A HYBRID PARSER MODEL FOR HINDI LANGUAGE

Sneha Asopa
Assistant Professor, Computer Science Department, Banasthali Vidyapith,
Newai, P.O. Banasthali Vidyapith, Newai, 304022, India
asopasneha@gmail.com

Neelam Sharma
Associate Professor, Computer Science Department, Banasthali Vidyapith,
Newai, P.O. Banasthali Vidyapith, Newai, 304022, India
sharmaneelam27@gmail.com

Abstract - Analyzing syntactic structure is the most complicated task for Indian Languages. In this paper, a probabilistic parser is proposed for Hindi language comprising the empirical and rationalist approaches. The task of tagging is accomplished with the help of TnT POS tagger. In this research work, along with the development and evaluation of probabilistic parser, evaluation of rule based and conditional random fields (CRF) based shallow parser is also done by using a test dataset of 100 tagged sentences of Hindi. The generation of probabilistic parser is formulated mainly by using rule based shallow parser, constructing grammar rules and assigning probabilities. The proposed probabilistic parser has shown the accuracy of 66%.

***Keywords*:** Hindi; parser; conditional random fields; rule based; probabilistic context free grammar.

1. Introduction

Language is flowing constantly and to make it definitive, grammar is formalized. Since 1950's researchers are working on empirical and statistical analysis of natural language. The main aim is to understand the structure of a language by generating an unbiased methodology and algorithm which is suitable for learning the syntax structure and lexicons from a given specific corpora. Groups of different phrases and words can help in understanding many aspects of a language. Based on the similarity in distributional behavior of words and phrases the task of categorization can be done. Chomsky [Brill (1997)] emphasized that the task of learning a language can be achieved by throwing insight on how child learns a language and what are the common features in all the languages. Chomsky formulated the universal grammar [Chomsky (1986)] which shifted the focus from empirical approaches to rationalist methods. With this approach researchers started developing the hand coded grammar rules [Chomsky (1957)] which were given as an input to the system. Two annotated corpora were widely used in 1980's namely Brown [Marcus et. al. (1993)] and Lancaster-Oslo–Bergen [Garside *et. al.* (1987)] which helped researchers in automatic formulation of lexical and syntactic information. This gave rise to the stochastic methods like Hidden Markov Model (HMM) [Rabiner (1989)]. HMM came into limelight which overruled rationalist approaches. This was the first time when stochastic approaches were used for assigning part of speech (POS) tags to each word with accuracy greater than 95 percent. In this era of stochastic approaches, along with HMM many models such as TnT models [Thorsten (2000)], Maximum Entropy Markov Model (MEMM) [Lafferty et. al. (2001)] and Conditional Random Fields (CRF) [Jurafsky and Martin (2009)] were introduced and developed for POS tagging. POS tagging is also considered as a sequential problem and is very crucial for analyzing the Natural Language.

Natural languages can be defined as the utterance of words in a syntactical structure. In other words, the finite sequence of morphemes and lexicons form a meaningful sentence by the use of grammar. Grammar restricts the formation of phrases, clauses and words by applying rules at every level. At morphological level, the formation of words of Hindi language is restricted by the help of affixes. The language consists of 22 prefixes from Sanskrit language, 10 prefixes of its own and 12 prefixes from Arabic and Persian language. The language also has compound words and five broad categories of suffixes. Figure 1 depicts the formation of words using affixes.

At phrase level, the formation of Hindi sentences cannot be bounded by the sequence of phrases. For example,

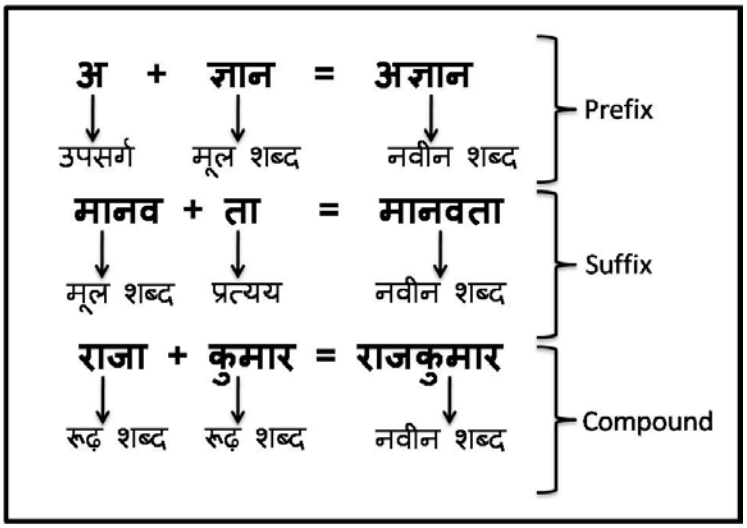

Fig. 1. Words formation in Hindi language

Hindi (HI): राम ने श्याम को खिलौना दिया

'Ram' 'erg' 'Shyam' 'to' 'toy' 'gave'

English (EN): Ram gave toy to Shyam

Hindi is a free Word order language, thus the above sentence remains grammatically correct in the following forms:

HI: राम ने दिया खिलौना श्याम को

HI: श्याम को राम ने खिलौना दिया

In Hindi, the nouns and pronouns relations with verb are shown by "Kaarakas". For example:

HI: कृष्ण मटकी फोड़ता है

'krishna' 'water pot' 'breaking' 'is'

EN: krishna is breaking water pot

In the above given sentence, कृष्ण (Noun) is a karta kaaraka, मटकी (Noun) is a karma karaka and फोड़ता है is a kriya(Verb). Karta and karma kaaraka both are related with verb. In the sentence, the questions like what is broken can be answered as the word मटकी (Noun) and who broke it can be answered as कृष्ण (Noun). Kaarakas [Bharati and Sangal (1990)] are divided into eight categories namely Karta (Nominative Case), Karma (Instrument Case), Karan (Ablative Case), Sampradan (Possessive Case), Apadaan (Objective Case), Sambandh (Dative Case), Adhikaran (Locative Case) and Sambodhan (Vocative Case). For example, in Figure 2, in sentence,

HI: सूरज निकलता है

'Sun' 'rising' 'is'

EN: Sun is rising

सूरज is the Karta having morphological feature 3rd form of person, Proper Noun, masculine, and singular and have the dependency on the verb निकलता है. In a sentence it is not necessary that all the kaaraka will be available and this makes Hindi tough to analyze. For the formulation of syntactic structure, phrase structure or dependency structure can be used.

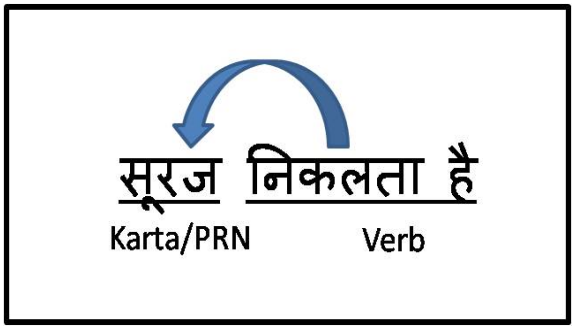

Fig. 2. Relations between words

In phrase structure part of speech tag defines the lexical category of a word. These lexical categories combine and form the head of the phrases and phrases are governed by certain rules. Various rules can be defined by looking at the Hindi sentences. For example, in a sentence,

HI: नटखट कृष्ण मटकी फोड़ता है

'Naughty' 'krishna' 'water pot' 'breaking' 'is'

EN: Naughty krishna is breaking water pot

Three phrases can be formed:

a) "नटखट कृष्ण" where, कृष्ण being a noun can be the head of the phrase and नटखट is the noun modifier with lexical category that belongs to adjective. The noun phrase can be extended by adding the adjectives and cardinals.

b) "मटकी" is another noun phrase

c) "फोड़ता है" is a verb phrase

Along with noun phrases, Hindi has primarily verb phrases, adjective phrases, adverb phrases, and pronoun phrases. For constructing a tree using phrase structure approach, grammar need to be formalized. Context free grammar cannot solve the problem of ambiguity so, instead of context free grammar, probabilistic context free grammar [Booth (1969)] can be used. Probabilistic context free grammar includes:

i) Context free grammar <N, S, Σ, P>

ii) Probabilistic parameter associated with every rule P(ω→ϒ) is as shown in "Eq. (1)" where, ω→ϒ belongs to Production rules <P> and ω is equal to a non terminal.

$$\Sigma P(\omega \rightarrow Y)=1. \quad (1)$$

For calculating the probability of a tree, we multiply all the P probabilities which are contained in rules of tree as shown in "Eq. (2)".

$$P(\omega \rightarrow Y)=P(NP \rightarrow DET\ NN) + P(NP \rightarrow NN)=1. \quad (2)$$

The probability of the parse tree as shown in figure 3 is calculated with "Eq. (3)" as

$$P(T) = P(S \rightarrow NP\ VP) * P(NP \rightarrow NN) * P(VP \rightarrow VM\ VAUX) * P(NN \rightarrow \text{राम}) * P(VM \rightarrow \text{जाता}) * P(VAUX \rightarrow \text{है}). \quad (3)$$

In this paper, we have proposed a hybrid approach for generating parse trees for Hindi language using empirical and rationalist methods.

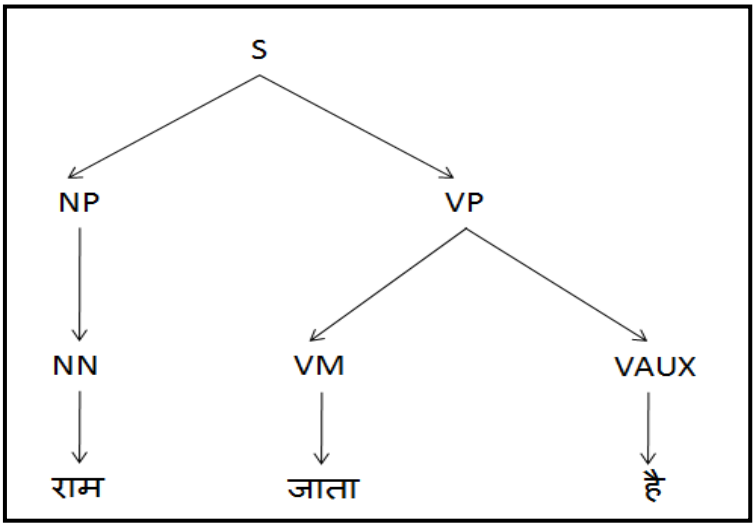

Fig. 3. Parse tree of a sentence "राम जाता है"

2. Related Works

Many researchers have done work for analyzing the syntactic structure of Indian Languages. [Agrawal (2007)] has developed probabilistic context free grammar model for Hindi and has also converted it to CNF form. [Dandapat (2007)] presented Conditional Random Field method for training the system. The authors tried to improve the machine learning without using any language tools such as morphological analyzers and dictionaries. The corpus was trained for B-L (Boundary Label) class for chunking. The method used to handle type-1 error improves the performance by 1.95% while method used to handle type-2 error degrades the performance by 11.6%. [Kumari and Rao (2012)] developed hybrid approach by combining output of both MALT and MST parsers. [Hambir and Srivastav (2015)] have developed a parser which uses CKY algorithm for parsing Hindi sentences. The parser is efficient in generating a matrix and parsing a sentence.

Inspired by the work done for part of speech tagging by [Reddy and Sharoff (2011)] [Bharati et.al. (2006)] and others [Agrawal (2007)], [Dandapat (2007)], [Kumari and Rao (2012)], [Hambir and Srivastav (2015)] on local word grouping and CRF based chunking [Sha and Pereira (2003)], [Nongmeikapam *et.al.* (2014)] shallow parsers of [Asopa *et.al.* (2016)] and [Asopa *et. al.* (2018)] were improved. In [Asopa *et. al.* (2016)] few more handcrafted rules were added and in [Asopa *et. al.* (2018)] the use of TnT Tagger [Reddy and Sharoff (2011)] was incorporated. The conceptual idea of selecting the best chunking technique and PCFG grammar motivated us to develop a parser based on PCFG approach for Hindi.

3. Methodology

In this paper, a parser based on PCFG grammar was developed for Hindi language along with the evaluation of PCFG parser and two Shallow Parsers. Figure 4 illustrates the proposed methodology.

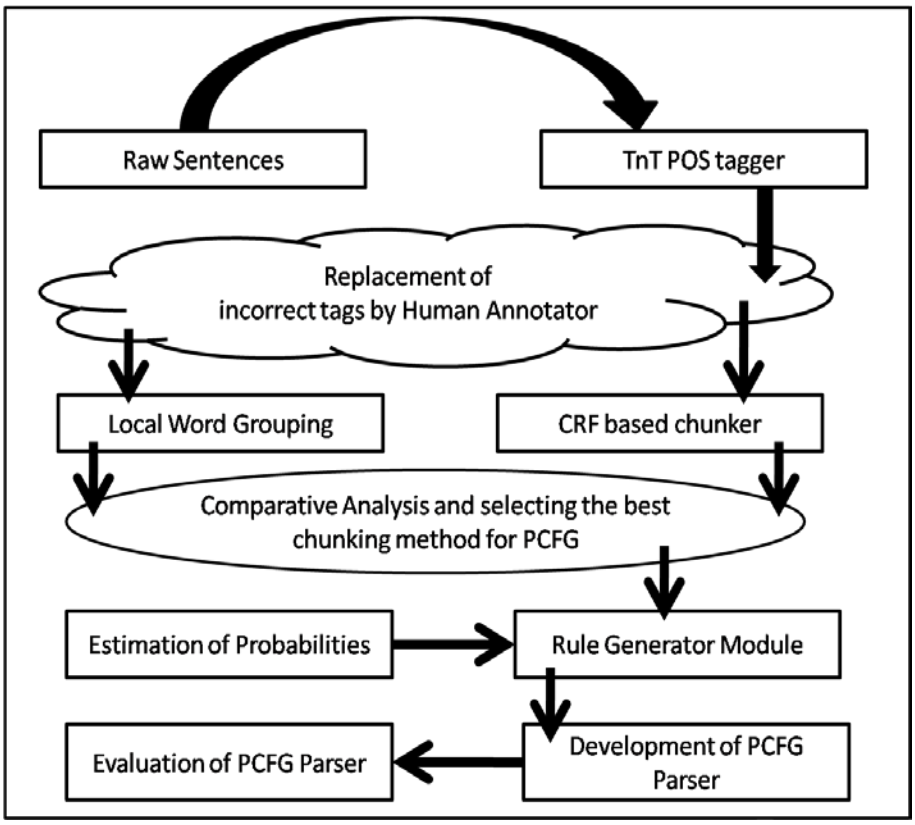

Fig. 4. Flowchart of proposed methodology

1000 raw sentences [Kunchukuttan et. al. (2018)] were given as an input to TnT POS tagger [Reddy and Sharoff (2011)]. If any incorrect tags were given by the tagger then they were replaced by the human annotated tags. From the tagger output, 100 tagged sentences were selected and were given as an input to a rule based shallow parsing algorithm which is based on rationalist approach. For using empiricist approach, the same set of 100 POS tagged sentences was given as an input to CRF based shallow parser model trained on rest 900 sentences. The outputs of both the approaches were evaluated by calculating the parameters F-Measure, Precision and Recall. When compared with CRF, Rule based shallow parser model generated better results. Figure 5 and 6 show the sample output of rule based shallow parser and CRF based shallow parser respectively.

As rule based shallow parser gave better results, further in methodology the rule based shallow parser generated output was given as an input to a rule generator algorithm. After generating the grammar rules, the probabilities were assigned to each rule manually.

Further, for developing a probabilistic parser, these rules were given as an input to the parsing algorithm through which the parse trees were formulated. After generation of parse trees, the evaluation was done by calculating the accuracy of the parse tree using "Eq. (4)".

$$Accuracy = (Correct\ Trees)/(Total\ Trees) \quad (4)$$

```
1  sentence:1 = आपने/PRP बचपन/NN में/PSP पंचतंत्र/NNP की/PSP
   कहानियाँ/NN जरुर/RB पढ़ी/VM होगी/VAUX |/PUN END/END
2
3  आपने/PRP --noun
4  बचपन/NN में/PSP --noun
5  पंचतंत्र/NNP की/PSP --noun
6  कहानियाँ/NN --noun
7  जरुर/RB --adverb
8  पढ़ी/VM होगी/VAUX --verb
9  |/PUN --punctuation
10
11 sentence:2 = मैं/PRP बचपन/NN में/PSP बहुत/JJ कहानियाँ/NN पढ़ता/VM
```

Fig. 5. Rule based shallow parser sample output

```
1  आपने  PRP  B-NP
2  बचपन  NN   B-NP
3  में    PSP  I-NP
4  पंचतंत्र  NNP  B-NP
5  की    PSP  I-NP
6  कहानियाँ  NN  B-NP
7  जरुर  RB   I-NP
8  पढ़ी  VM   B-VP
9  होगी  VAUX  I-VP
10 |     PUNC  O
```

Fig. 6. CRF based shallow parser sample output

4. Result and Discussion

The rule based and CRF based shallow parsers were compared by using the same set of 100 POS tagged sentences and was found that the rule based shallow parser was producing better results. While comparing between the Chunking Models, this was observed that in figure 6, जरुर (jarur) word is an adverb and is part of adverb phrase, but it is considered inside the noun phrase in CRF based shallow parser. However, in figure 5, the rule based shallow parser has given the correct output. The compared accuracies are given in table 1.

Table 1. Evaluation of Shallow Parsers

Shallow Parser	F-Measure (%)	Precision (%)	Recall (%)
CRF Based	98.04	98.04	98.04
Rule Based	99.59	99.54	99.65

The proposed probabilistic parser was evaluated on 100 sentences. The estimated accuracy was found to be as 66 %. The output of the parser for one sentence is depicted in Figure 7. The generated rules in the output are in CNF form with probabilities assigned to each rule.

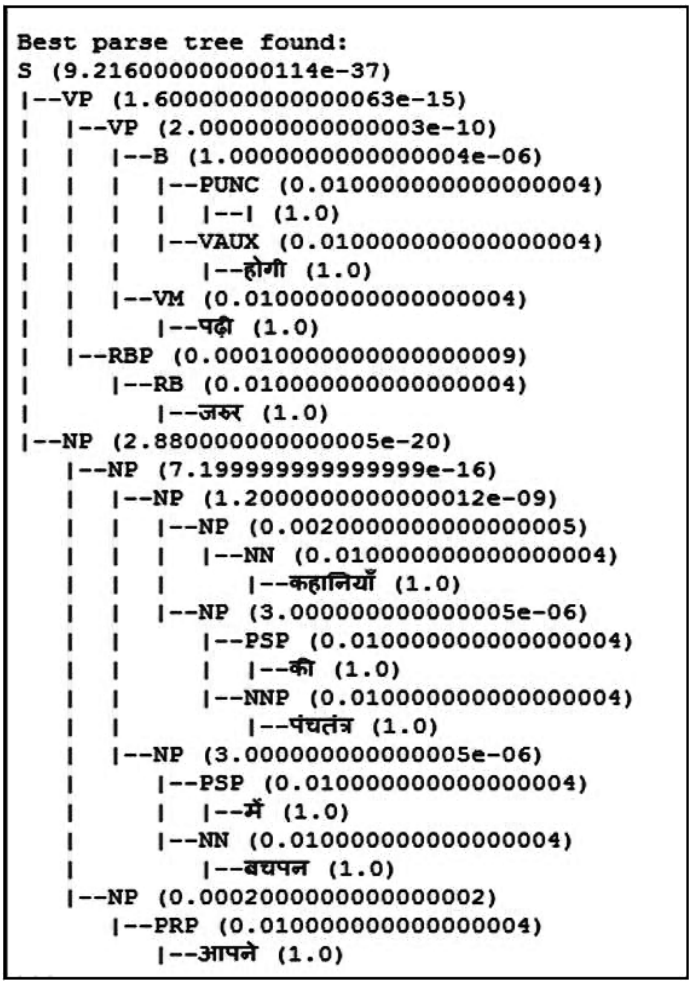

Fig. 7. Sample output of proposed PCFG parser

5. Conclusions

In this research work, empirical and rationalist approaches were compared and analyzed. The amalgamation of both statistical and rule based methods can endow with better accuracy for Hindi Language. Thus, in this research a hybrid model of PCFG parser is developed and evaluated producing the accuracy of 66%.

References

[1] Agrawal. H. (2007). POS Tagging and Chunking for Indian Languages, Shallow Parsing for South Asian Languages: 37 2007.
[2] Asopa, S; Asopa, P; Mathur, I; Joshi, N. (2016). Rule based chunker for Hindi, In Contemporary Computing and Informatics (IC3I), 2nd International Conference, IEEE, pp. 442-445.
[3] Asopa, S; Asopa, P; Mathur, I; Joshi, N. (2018). A Shallow Parsing Model for Hindi using Condition Random Fields, In Proceedings of International Conference on Innovative Computing and Communication.
[4] Bharati, A; Sangal, R; Sharma, D.M; Bai. L. (2006). Anncorra: Annotating corpora guidelines for pos and chunk annotation for indian languages, LTRC-TR31.
[5] Bharati, A; Sangal,R.(1990). A karaka based approach to parsing of Indian languages, In Proceedings of the 13th conference on Computational linguistics, Association for Computational Linguistics, Vol. 3, pp. 25-29.
[6] Booth, T.L. (1969). Probabilistic representation of formal languages, In IEEE Conference Record of the 1969 Tenth Annual Symposium on Switching and Automata Theory, 74–81.
[7] Brill, E.; Mooney, R. J. (1997). An Overview of Empirical Natural Language Processing, AI Magazine, 18(4), 13.
[8] Chomsky, N. (1957). Syntactic Structures, The Hague, The Netherlands: Mouton.
[9] Chomsky, N. (1986). Knowledge of Language: Its Nature, Origin and Use, New York: Praeger.
[10] Dandapat. S. (2007). Part of Speech Tagging and Chunking with Maximum Entropy Model, Shallow Parsing for South Asian Languages: 29.
[11] Garside, R.; Leech, G; Sampson, G.(1987) The Computational Analysis of English: A Corpus-Based Approach, London: Longman
[12] Hambir. N; Srivastav. A. (2015). Hindi Parser-based on CKY algorithm, Int. J. Computer Technology & Applications, 851-853.
[13] Jurafsky. D; Martin. J.H.(2009). Hidden Markov and Maximum Entropy Models, Chapter 6 in Speech and Language Processing, Second Edition, Prentice-Hall, Inc.

[14] Kumari, B.V.S; Rao. R.R. (2012). Hindi Dependency Parsing using a combined model of Malt and MST, In 24th International Conference on Computational Linguistics p. 171.
[15] Kunchukuttan, A; Mehta, P; Bhattacharyya, P. (2018). The IIT Bombay English-Hindi Parallel Corpus, Language Resources and Evaluation Conference.
[16] Lafferty, J; McCallum, A; Pereira. F.(2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data, Proc. 18th International Conference on Machine Learning.
[17] Marcus, M; Santorini, B; Marcinkiewicz, M. (1993) Building a Large Annotated Corpus of English: The Penn Tree Bank, Computational Linguistics 19(2): 313–330.
[18] Nongmeikapam, K; Chingangbam, C; Keisham. N; Varte, B; Bandopadhyay, S. (2014). CHUNKING IN MANIPURI USING CRF, In international journal on natural language computing (IJNLC) Vol.3,No.3.
[19] Rabiner, L.R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE 77(2): 257–286.
[20] Reddy. S; Sharoff. S. (2011). Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources, Cross Lingual Information Access: 11.
[21] Sha, F; Pereira, F.(2003). Shallow parsing with conditional random fields, In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Association for Computational Linguistics, Vol. 1, pp. 134-141.
[22] Thorsten, B. (2000). TnT: a statistical part-of speech tagger, In Proceedings of the sixth conference on Applied natural language processing, ANLC '00, Stroudsburg, PA, USA. Association for Computational Linguistics, pages 224–231.