Rule Based Hindi Part of Speech Tagger

Navneet Garg[1,1], *Vishal Goyal*[2,1], *Suman Preet*[3,2]
(1) Department of Computer Science, Punjabi University, Patiala
(2) Department of Linguistics and Punjabi Lexicography, Punjabi University, Patiala.
navneetgarg123@rediffmail.com, vishal.pup@gmail.com,
virksumanpreet@yahoo.co.in

ABSTRACT

Part of Speech Tagger is an important tool that is used to develop language translator and information extraction. The problem of tagging in natural language processing is to find a way to tag every word in a sentence. In this paper, we present a Rule Based Part of Speech Tagger for Hindi. Our 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 stories. Our system achieved the accuracy of 87.55%.

KEYWORDS: POS, Tagging, Rules, Hindi.

Proceedings of COLING 2012: Demonstration Papers, pages 163–174,
COLING 2012, Mumbai, December 2012.

1. Introduction

Natural language processing is a field of computer science, artificial intelligence (also called machine learning) and linguistics concerned with the interactions between computers and human (natural) languages. Specifically, it is the process of a computer extracting meaningful information from natural language input and/or producing natural language output. Part of Speech tagger is an important application of natural language processing. Part of speech tagging is the process of assigning a part of speech like noun, verb, preposition, pronoun, adverb, adjective or other lexical class marker to each word in a sentence. 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.

2. Related Work

There have been many implementation of part of speech tagger using statistical approach, mainly for morphological rich languages like Hindi. Statistical techniques are easy to implement and require very less knowledge about the language.

Aniket Dalal et al., 2006 developed a system using Maximum Entropy Markov Model for Hindi. System required a feature function capturing the lexical and morphological feature of language and feature set was arrived after an in-depth analysis of an annotated corpus. The system was evaluated over a corpus of 15562 words with 27 different POS tags and system achieved the accuracy of 94.81%.

Smriti Singh et al., 2006 developed a part of speech tagger using decision tree base learning. This methodology uses locally annotated modestly sized corpora, exhaustive morphological analysis backed by high coverage lexicon. The heart of the system is detailed linguistic analysis of morph syntactic, handling of suffixes, accurate verb group identification and learning of disambiguation rules. The evaluation of the system was done with 4-fold cross validation of the corpora in the news domain and accuracy of the system is 93.45%.

Himanshu Aggarwal et al., 2006 developed a system using Conditional Random Fields for Hindi. A morph analyzer is used to provide information like root words and possible POS tags for training. The system was evaluated over a corpus of 21000 words with 27 different POS tags and system achieved the accuracy of 82.67%.

Manish Shrivastava et al., 2008 developed a system using Hidden Markov Model for Hindi. The System uses stemmer as a preprocessor to find the root of the words. The system was developed using 18 different pos tags and system achieved the accuracy of 93.12%.

Sanjeev Kumar Sharma et al., 2011 developed a system using Hidden Markov Model to improve the accuracy of Punjabi Part of Speech tagger. A module has been developed that takes output of the existing POS tagger as input and assign the correct tag to the words having more than one tag. The system was evaluated over a corpus of 26,479 words and system achieved the accuracy of 90.11%.

Pranjal Awasthi et al., 2006 developed a system using a combination of Hidden Markov Model and error driven learning. Tagging process consists of two stages, an initial statistical tagging using the TnT tagger, which is a second order Hidden Markov Model (HMM) and apply a set of transformation rules to correct the errors introduced by the TnT tagger. The system was developed using 26 different POS tags and accuracy of system is 79.66% using the TnT tagger and transformations in post processing improves the accuracy to 80.74%.

Shachi Mall et al., 2011 developed a system using a Rule based approach. The module reads the Hindi corpus and split the sentence into words according to the delimiter. The system finds the words in the database and assigns the appropriate tag to the words.

We can see that most of the taggers are developed using statistical techniques because these techniques are easy to implement and require very less knowledge about the language. In this paper, we presented a rule based approach to design part of speech tagger. Rule based approach required less amount of data and vast knowledge about the language. Rule based system is usually difficult to develop.

3. System Description

This system is developed using rule based approach and 30 different standard part of speech tags are [shown in Appendix A] used that are given by Department of Information Technology Ministry of Communications & Information Technology and some other tags that is time, date and number tag. Adverb tag is further classified into following categories i.e. adverb of manner (RB_AMN), adverb of location (RB_ALC), adverb of time (RB_TIME) and adverb of quantity (RB_Q). Collection of 18,249 words for different tags has been done. The system mainly works in two steps-firstly the input words are found in the database; if it is present then it is tagged. Secondly if it is not present then various rules are applied.

3.1 Algorithm

1. Input the text using file upload button or manually enter by user.
2. Tokenize the input text word by word.
3. Normalized the tokenized words. i.e. Separate out the punctuation marks and the symbols from the text.

4. Search the number tag by using Regular Expression.
For Example: - 2012, 1-2, 1.2, 12वें, २३, ६.७, ६-७ etc.

5. Search the date tag by using regular expression.
For Example: - 17/10/1985, 17-10-1985, etc.

6. Search the time tag by using regular expression.
For Example: - 12:10, 12:23:45 etc.

7. Search for the abbreviation using regular expression.
For Example: - ए.पी, आर.के. etc.

ē.pī, ār.kē.

8. Search in database for different input words and tag the word according to corresponding tag.

9. Then different rules are applied to tag the unknown words.

10. Display the tagged data to the user.

3.2 Following Rules are applied to identify different Tags

1. Noun Identification Rules

Rule 1: If word is adjective then there is high probability that next word will be noun.

For Example:-

वह एक सच्चा देशभक्त है।

vah ēk saccā dēshbhakt hai.

In above example सच्चा (saccā) is adjective and देशभक्त (dēshbhakt) is noun.

Rule 2: If word is relative pronoun then there is high probability that next word will be noun.

For Example:-

ये वो घर है जिसे राजा ने बनवाया था।

yē vō ghar hai jisē rājā nē banvāyā thā.

In above example वो (vō) and जिसे (jisē) is relative pronoun and घर (ghar) and राजा (rājā) is noun.

Rule 3: If word is reflexive pronoun then there is high probability that next word will be noun.

For Example:-

वह अपने घर चला गया।

vah apnē ghar calā gayā .

In above example अपने (apnē) is reflexive pronoun and घर (ghar) is noun.

Rule 4: If word is personal pronoun then there is high probability that next word will be noun.

For Example:-

यह हमारा घर है।

yah hamārā ghar hai .

In above example हमारा (hamārā) is personal pronoun and घर (ghar) is noun.

Rule 5: If current word is post position then there is high probability that previous word will be noun.

For Example:-

उसने पानी में पत्थर फेंका।

usnē pānī mēm patthar phēṅkā .

In above example पानी (pānī) is noun and में (mēṃ) is post position.

Rule 6: If current word is verb then there is probability that previous word will be noun.

For Example:-

वह भोजन खा रहा है।

vah bhōjan khā rahā hai.

In above example भोजन (bhōjan) is noun and खा (khā) is verb.

Rule 7: If word is noun then there is probability that next or previous word will be noun.

For Example:-

वह फ़ाइनल मुकाबले में हार गए ।

vah phaāinal mukāblē mēṃ hār gaē.

In above example फ़ाइनल (phaāinal) and मुकाबले (mukāblē) both are noun.

There are more rules are applied to find the noun tags.

2. Demonstrative Identification Rules

Rule 1: If word is pronoun in database and next word is also pronoun, then first word will be demonstrative.

For Example:-

वह कौन है।

vah kaun hai.

In above example वह (vah) and कौन (kaun) both are pronoun.

Rule 2: If current word is pronoun in database and next word is noun, then curremt word will be demonstrative.

For Example: -

वह मुंबई नहीं जाएंगे।

vah mumbī nahīṃ jāēṅgē.

In above example वह (vah) is pronoun and मुंबई (mumbī) is noun.

3. Proper Noun Identification Rules:-

Rule 1: If current word is not tagged and next word is tagged as proper noun, then there is high probability that current word will be proper noun.

For Example: - आर.के. गोयल, राम गोयल

ār.kē. gōyal, rām gōyal

In above example आर.के.(ār.kē.), गोयल (gōyal) and राम (rām) are proper noun.

Rule 2: If current word is name and next word is surname then we tagged them as single proper name.

For Example: - सुरेश कुमार <N_NNP>

surēsh kumār

In above example सुरेश (surēsh) is name and कुमार (kumār) is surname.

4. Adjective Identification Rules:-

Rule 1: If word ends with तर (tar), तम (tam), िक (ik) postfix then word is tagged as adjective.

For Example: - लघुतर, विशालतम, प्रामाणिक

laghutar, vishāltam, prāmāṇik

5. Verb Identification Rules:-

Rule 1: If current word is not tagged and next word tagged as a auxiliary verb, then there is high probability that current word will be main verb.

For Example:-

वह खाना खा रहा है।

vah khānā khā rahā hai.

In above example खा (khā) is main verb and रहा (rahā) is auxiliary verb.

The system can be understood by following example:-

Input Hindi Sentence

श्रीनगर में एक 200 साल पुरानी दरगाह में आग लगने के बाद प्रदर्शनकारियों ने पुलिस पर पथराव किया है और इलाके में तनाव है।

shrīngar mēṃ ēk 200 sāl purānī dargāh mēṃ āg lagnē kē bād pradrshankāriyōṃ nē pulis par pathrāv kiyā hai aur ilākē mēṃ tanāv hai.

Output

shrīngar <N_NNP> mēṃ <PSP> ēk <QT_QTC>200<NUMBER> sāl <N_NN> purānī <JJ> dargāh <N_NN> mēṃ <PSP> āg <N_NN> lagnē <V_VM> kē <PSP> bād <PSP> pradrshankāriyōṃ <N_NN> nē <PSP> pulis <N_NN> par <PSP> pathrāv <N_NN> kiyā <V_VM> hai <V_VAUX> aur <CC_CCD> ilākē <N_NN> mēṃ <PSP> tanāv <N_NN> hai <V_VAUX> .<RD_PUNC>

4. Evaluation and Result

Evaluation is done to enhance the performance of system on different domains of news. These domains include news, essay, and short stories. The system was evaluated on 26,149 words. The overall accuracy achieved by system is 87.55%. We have constructed three test data sets for testing. These test data sets are collected from different websites [15][16] of Hindi. Following table shows the different test cases for testing.

Test No.	Domain	No. of words
Test Case 1	News	17233
Test Case 2	Essay	5039
Test Case 3	Short Stories	3877

Table 5.1 Test Cases

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 = (β^2 + 1) PR / β^2 R + P

β is the weighting between precision and recall and typically $\beta = 1$.

	Recall	Precision	F-Meaure
Set-1	92.84%	89.94%	91.37%
Set-2	87.32%	81.36%	84.23%
Set-3	88.99%	85.11%	87.06%

Table 5.2 Accuracy of System on different Test Cases

Conclusion and Future Work

In this paper Part of Speech tagger using rule based technique has been discussed. Tokenized words are search in the database and if not found then appropriate rules are applied. Sometimes when we apply rules then system may tag the words with wrong POS tags.

If a sentence consists of 12 words out of which 8 words are unknown, then system fails to tag them. The reason behind it is hard to decide which rules should be handled first because word tagging resolution is based on neighbour's words.

By increasing the size of database accuracy of part of speech tagger can be increased. Hybrid based system can be developed to increase the accuracy of system. There is problem in handling the words that can act as both common noun and proper noun. So it becomes difficult for the system to tag the word correctly. When such a situation occur system tag the word as a common noun, there is high probability that word will be a common noun but in few cases it can act as proper noun. This limitation can be handled by using Hindi Named Entity Recognition system in future.

References

[1] Aniket Dalal, Kumar Nagaraj, Uma Sawant and Sandeep Shelke. (2006). *Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach*, In Proceeding of the NLPAI Machine Learning Competition, 2006.

[2] Manish Shrivastava and Pushpak Bhattacharyya. (2008). *Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information without Extensive Linguistic Knowledge*, International Conference on NLP (ICON08), Pune, India, and December, 2008.

[3] Agarwal Himashu, Amni Anirudh. (2006). *Part of Speech Tagging and Chunking with Conditional Random Fields,* In the proceedings of NLPAI Contest, 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] Nidhi Mishra and Amit Mishra. (2011). *Part of Speech Tagging for Hindi Corpus*, In the proceedings of 2011 International Conference on Communication systems and Network Technologies, pp.554-558.

[6] Eric Brill. (1992). *A Simple Rule Based Part of Speech Tagger*, In Proceeding of the Third Conference on Applied Computational Linguistics (ACL), Trento, Italy, 1992, pp.112–116.

[7] Sanjeev Kumar Sharma and Gurpreet Singh Lehal. (2011). *Using Hidden Markov Model to Improve the Accuracy of Punjabi POS Tagger*, Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on June 2011, pp. 697-701.

[8] Shachi Mall and Umesh Chandra Jaiswal. (2011). *Hindi Part of Speech Tagging and Translation*, In the proceedings of Int. J. Tech. 2011, Vol. 1: Issue 1, pp. 29-32.

[9] Pranjal Awasthi, Delip Rao and Balaraman Ravindran. (2006). *Part Of Speech Tagging and Chunking with HMM and CRF*, In the proceedings of NLPAI Contest, 2006.

[10] Dinesh Kumar and Gurpreet Singh Josan. (2010). *Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey*, International Journal of Computer Applications (0975 – 8887) Volume 6– No.5, September 2010, pp.1-9.

[11] Sankaran Baskaran. *Hindi POS Tagging and Chunking*, Microsoft Research India, Banglore.

[12] Antony P J and Dr. Soman K P. (2011). *Part of Speech Tagging for Indian Languages: A Literature Survey*, International Journal of Computer Applications (0975 – 8887) Volume 34– No.8, pp.22-29.

[13] Mandeep Singh, Lehal Gurpreet, and Sharma Shiv. (2008). *A Part-of-Speech Tagset for Grammar Checking of Punjabi*, published in The Linguistic Journal, Vol 4, Issue 1, pp 6-22.

[14] http://en.wikipedia.org/wiki/Part-of-speech_tagging

[15] http://www.bbc.co.uk/hindi/

[16] http://www.bhaskar.com/

[17] http://en.wikipedia.org/wiki/Natural_language_processing

[18] http://en.wikipedia.org/wiki/Support_vector_ machine

[19] http://en.wikipedia.org/wiki/Hidden_Markov _model

[20] http://en.wikipedia.org/wiki/Maximum_entropy

Appendix A

Standard POS Tags

Sr. No	Category		Label	Annotation Convention	Examples
	Top Level	Subtype			
1.	Noun		N	N	ladakaa, raajaa, kitaaba
1.1		Common	NN	N_NN	kitaaba, kalama, cashmaa
1.2		Proper	NNP	N_NNP	Mohan, ravi, rashmi
1.3		Nloc	NST	N_NST	Uupara, niice, aage, piiche
2	Pronoun		PR	PR	Yaha, vaha, jo
2.1		Personal	PRP	PR__PRP	Vaha, main, tuma, ve
2.2		Reflexive	PRF	PR_PRF	Apanaa, swayam, khuda
2.3		Relative	PRL	PR_PRL	Jo, jis, jab, jahaaM,
2.4		Reciprocal	PRC	PR_PRC	Paraspara, aapasa

2.5	Pronoun	Wh-word	PRQ	PR_PRQ	Kauna, kab, kahaaM
2.6		Indefinite	PRI	PR_PRI	Koii, kis
3	Demonstrative		DM	DM	Vaha, jo, yaha,
3.1		Deictic	DMD	DM_DMD	Vaha, yaha
3.2		Relative	DMR	DM_DMR	jo, jis
3.3		Wh-word	DMQ	DM_DMQ	kis, kaun
3.4		Indefinite	DMI	DM_DMI	KoI, kis
4	Verb		V	V	giraa, gayaa, sonaa, haMstaa, hai, rahaa
4.1		Main	VM	V_VM	giraa, gayaa, sonaa, haMstaa,
4.2		Auxiliary	VAUX	V_VAUX	hai, rahaa, huaa,
5	Adjective		JJ	JJ	sundara, acchaa, baRaa
6	Adverb		RB	RB	jaldii, teza
7	Postposition		PSP	PSP	ne, ko, se, mein

8	Conjunction		CC	CC	aur, agar, tathaa, kyonki
8.1		Co-ordinator	CCD	CC_CCD	aur, balki, parantu
8.2		Subordinator	CCS	CC_CCS	Agar, kyonki, to, ki
9	Particles		RP	RP	to, bhii, hii
9.1		Default	RPD	RP_RPD	to,bhii, hii
9.2		Interjection	INJ	RP_INJ	are, he, o
9.3		Intensifier	INTF	RP_INTF	bahuta, behada
9.4		Negation	NEG	RP_NEG	nahiin, mata, binaa
10	Quantifiers		QT	QT	thoRaa, bahuta, kucha, eka, pahalaa
10.1		General	QTF	QT_QTF	thoRaa, bahuta, kucha
10.2		Cardinals	QTC	QT_QTC	eka, do, tiina,
10.3		Ordinals	QTO	QT_QTO	pahalaa, duusaraa
11	Residuals		RD	RD	
11.1		Foreign word	RDF	RD_RDF	

	Residuals				
11.2		Symbol	SYM	RD_SYM	$, &, *, (,)
11.3		Punctuation	PUNC	RD_PUNC	., : ; ? | !
11.4		Unknown	UNK	RD_UNK	