ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 2, February 2013

A Combined Approach to Part-of-Speech Tagging Using Features Extraction and Hidden Markov Model

Bhairab Sarma
Lecturer, Faculty of Science & Technology, The Icfai University Tripura, Kamalghat, Tripura-799210

Prajadhip Sinha
Asst. Professor, Department of Computer Science, Kohima Science College, Kohima, Nagaland

Dr. Bipul Shyam Purkayastha
Professor, Computer Science Department, Assam University, Silchar 788011

Abstract

Words are characterized by its features. In an inflectional language, category of a word can be express by its tense, aspect and modality (TAM). Extracting features from an inflected word, one can categorised it with proper morphology. Hence features extraction could be a technique of part-of-speech (POS) tagging for morphologically inflected languages. Again, many words could have same features with distinguish meaning in context. However contextual meaning could be recovered using Hidden Markov Model (HMM). In this paper we try to find out a common solution for part-of-speech tagging of English text using both approaches. Here we attempt to tag words with two perspectives: one is feature analysis where the morphological characteristics of the word are analyse and second is HMM to measure the maximum probability of tag based on contextual meaning with previous tag.

KEYWORDS: HMM, POS, contextual meaning, features extraction, TAM

1. Introduction

Part-of Speech Tagging is an activity to assign a proper tag (symbol) to each word from a written text or corpora. The primary objective of Part-of-Speech tagging (POST) is to word-category disambiguation. It is the process of marking up the words in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context —i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph[1]. The goal of POST is to identify words as nouns, verbs, adjectives, adverbs in a sentence according to the context as they appear to represent a meaning. Problem in here is for the words having multiple meaning[2] . There are many approaches to retrieve the sense of a word when they are playing a particular role in the context. [2] discussed an approach to express multiword expression where a tree technology is used for context retrieval. Hidden Markov Models is a common technique being used in POST since 1980. Scott and Harper (2008) used HMM to disambiguate parts of speech, during working to tag the Lancaster-Oslo-Bergen Corpus of British English in[3]. This involves counting cases such as from the Brown Corpus as in[4], and making a table of probabilities of certain sequences. For example, if there is an article 'the' in the sentence, perhaps the next word is a noun 40% of the time, an adjective 40%, and a number 20%. Retrieving these information's, a program can take a decision based on the maximum probability of the category. Rantaparikhi (1996) in [5] proposed his maximum entropy model where he claimed about 95% accuracy in part-of-speech tagging of structural language. With this entropy model, maximum probability of a tag could be determined without using HMM where comparatively more mathematical computations were required for higher accuracy level. Mark Jhonson (2007) in [6] pointed some reasons in favour of EM model comparative to HMM. Knowledge about the contextual meaning could be represented for the succeeding words in text. A second order HMM learn the probability up to two previous words in the sentence where as in higher order learn for triples or more. Higher in order of HMM used in POST, better is the accuracy result of tagging. Dynamic programming method is one alternative, developed by Steven DeRose and Ken Church (1987) with similar technique used in Viterbi algorithm. Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states – called the Viterbi path

– that results in a sequence of observed events, especially in the context of Markov information sources, and more generally, hidden Markov models. The forward algorithm is a closely related algorithm for computing the probability of a sequence of observed events. These algorithms belong to the realm of information theory.

Viterbi algorithm makes a number of assumptions.

- Two states of event sequences correspond to time: observed events and hidden events.
- There is a one-to-one mapping of these two sequences. An instance of an observed event needs to correspond to exactly one instance of a hidden event.
- A transition which computing the most likely hidden sequence up to a certain point t must depend only on the observed event at point t, and the most likely sequence at point $t - 1$.

These assumptions are all satisfied in a first-order hidden Markov model. But DeRose as in [3] used a table of pairs, while Church used a table of triples and an ingenious method of estimating the values for triples that were rare or nonexistent in the Brown Corpus. Both methods claimed accuracy over 95%.

2. Our approach

Taggers are broadly classified as: Rule-based, Stochastic and Hybrid (Eric Brill, 1992). A stochastic tagger falls under unsupervised category[7]. Unsupervised tagging techniques use an untagged corpus for their training data and produce the tagset by induction. Here patterns in word are observed, and derive part-of-speech categories themselves. For example, in most of the cases, “the”, “a”, and “an” occur in similar contexts, while “eat”, “run” occurs in different ones. With sufficient iteration, similar classes of words emerge that are remarkably similar to those human linguists would expect; and the differences themselves sometimes suggest valuable new insights. In[8], Adam and Hal proposed a lexicon independent unsupervised part-of-speech tagger using HMM approach. It reveals that lexicon is not an essential component to build a tagger.

Our approach is a stochastic in nature perhaps falls in unsupervised category but not depend on HMM only. Brill (1992) in [7] demonstrate that the stochastic method is not only viable method for part of speech tagging. Here we combine two approaches; features extraction with morphological analysis and HMM. The first approach is used to tag closed word which are minimum in numbers and the second approach is used to smooth the probability of tagging for confused tags. For this purpose, we tag a text manually and extract the features of each unique word from the text depending on the assigned tags. We consider the relevance of previous tags as one of the feature of the word which reflects the application of Markov property. In this study, we collect the sample data from the first chapter of the book “Natural Language Processing: a Paninian Perspective”. In [9], the authors Bharati et al. gives a brief introduction and overview of natural language processing including prospect and some key area of NLP development. They also describe some application area of NLP in this chapter. We tag manually 1409 root word based on morphological features. Depending on the tag frequency, we classify the entire tag into two classes as open and closed class. Closed class tags are taken special consideration in tagging. We used Penn Treebank tag set that consist of 45 tags. Out of all 45 tags, we exclude 10 punctuation marker tags and consider 35 tags for study. We follow the tagging guideline provided by Santorini of University of Pennsylvania, Philadelphia[1].

3. Observation

In this study, we first collect the word list by a special module called ‘tokenizer’ that tokenize individual token from the text file. Here, we tokenize the text with multiple successions. First white space is used to segregate the independent word, and then take care of about the ending character of each word. If the ending character is a special character (i.e. ., ;, :, ? etc), the word is again segregate in two tokens. If two consecutive words begin with capital character, we group them in a one token and assign a probable tag as ‘NNP’. The frequency of each word is counted by a separate module called ‘word-frequency-count’. This module is used to count the number of appearance of a particular word in the text. The system has been developed with PHP and MySql as front and back end respectively. We collect total 1409 word and

[1] Part-of-S peech Tagging Guidelines For The Penn Treebank Project (3rd Revision) MS-CIS-90-47 LINC LAB 178; Presented by Beatrice Santorini, Department of Computer and Information Science School of Engineering and Applied Science University of Pennsylvania Philadelphia, PA 19 104 July 1990 available at: http://repository.upenn.edu/cis reports/570

tagged manually according to Penn Treebank tagset and the frequency of tags are given in table 1.

Table 1: Tag Frequency

CC	62	JJS	3	VBG	28
CD	39	LS	4	PRP	7
DT	115	MD	7	PRP$	34
EX	8	NN	312	RB	21
FW	0	NNS	111	RBR	7
IN	160	NNP	17	RBS	2
JJ	206	NNPS	2	RP	0
JJR	4	PDT	6	SYM	0
TO	31	VB	92	VBD	59
VBN	28	VBP	3	VBZ	14
WDT	14	WP	1	WRB	7

Our approach does not follow HMM directly for context retrieval. In HMM, the meaning of a word W_i is depend on the meaning of its previous word sequence. However, in this approach it is difficult to predict the tag of the first word of the sentence. According to[10], in unsupervised method, no previous information is available in the system for tagging. Our approach is slightly different from this approach. We consider HMM finding tag transition probability for those words which have same features and confused in tagging them as per their appearance. Here we take the advantages form both HMM and features extraction method so that maximum possibility could be determined. For example, capitalized feature identified as noun, however each line is started with a capital letter. So, if the word is not a beginning word, and if it is not an abbreviation, the probability of being a noun is maximum. In next iteration, we take the transition matrix and find the transition probability for the tag. If the beginning word is a determiner for the next i.e. the previous tag is DT, then the possibility of the word to be noun is maximum. Hence the word would be tagged as a noun because:

1. The word is not a beginning word and it start with capital letter
2. The word is not an abbreviation because next character is not capital letter
3. The previous tag is a DT.

Next we identify the unique words that falls in same category. For example, in the third row of table 1, DT is given as 115 i.e. in the selected text, 115 times articles appeared. However in table two we show the number of unique word that falls in a particular category. For example in row two of table 2, DT is given three only. Meaning is that in our specimen text, although 115 of DT category word exist, only three separate word are exist in this category, and these are ‘a’, ‘an’, and ‘the’. We found total 334 unique words in this table.

From these collections of words, we analyze each group of word morphologically and derive some features that hold in each word. These features are used to predict a probable tag for an unknown word. All features are enlisted in rows against each open category tags to form a features matrix. We show the matrix in table 5. Finally, we find the maximum probability of a tag by considering all features that hold in the specimen word and assign the tag to the word. For example, in case of tag NNS, the following features are observed:

1. If the word is ended with ‘s’ or ‘es’
2. The root word is ‘noun’ category
3. The previous tag is JJ
4. The word is start with a capital letter.

Table 2: List of Number of word in unique category

Tag category	**No. of Words**	**Tag category**	**No. of words**	**Tag category**	**No. of Word**	**Tag category**	**No. of Words**
CC	8	CD	40	EX	1	IN	6
DT	3	JJ	29	JJR	4	JJS	4
LS	8	MD	6	NN	39	NNS	27
NNP	17	NNPS	6	PDT	6	PRP	5
RBR	7	RB	17	PRP$	12	RBS	4
VB	24	VBD	32	VBG	22	VBN	17
VBP	2	VBZ	12	WDT	3	WRB	5
WP	2	WP$		SYM			

For a particular tag, if the number of word collection is less than 10% of total tags/word, the word is considered as rare word and that tags are not analysis for features extraction. Table 3 shows the sample size of unique word per tags. For example, in EX tags (in Table 2), we found only one word i.e. existential ‘there’. Hence no features are valid for EX tag except the word ‘there’. Sometimes ‘there’ could be used as the beginning word but it is not classified as noun.

Table 3: Unique word per tag

Sample size	Unique word	Number of Tag	Average no. of W/T
1409	334	31	10.77

The following tags do not come under study due to small number (rare word) of word exists in the sample. In this case we simply tagged with specific word. For example, as stated previously, under EX tag we have only one word. So, the only feature is ‘word’ and consider as unique features, not applicable to other tags.

1. If the word is ‘there’ then the word tagged as ‘EX’.

Few tags falls under this closed group is given below:

CC: Coordinating conjunction. Word group is “or”, “and”, “also”, “but”, “either”, “neither”, “nor”. ,

DT: Determiner. If the word is either “a”, “an”, “the”, “every”, “any”, “some”, “each”, “ether”, “neither”, “this”, “that”, “these”, “those”.

EX: Existential there –“there”

IN: Interjection group of word consist “by”, “at”, “for”, “in”, “of”, “on”, “under”, “over”.

MD: Modal verb, a type of auxiliary verb. Modal verbs do not take the inflection *-s* or *-es* in the third person singular, unlike other verbs. Includes “shall”, “should”, “will”, “would”, “can”, “could”, ”may”, “might”, “mote”, “must”.

WDT: Words include “which”, “that”.

WP: Word include “what”, “why”, “who”, “whom”.

WRB: In this group the word includes “how”, “however”, “whereas”, “whether”, “where”.

3.1 The Features Matrix

We consider the following tags for analysis. These tags are frequently used in text and shows ambiguity in tagging. Words found here are very confusing in their sense and considered as open class. Following table 4: shows some confused tags with their frequencies.

Table 4: Confusion tags with frequency

NNP	16	VB	24	CD	40	JJ	23	RB	17
VBD	32	VBG	22	NN	30	PRP$	12	VBZ	12
NNS	19	VBN	7						

From this list, we observed some peculiar features that exist in these words and find the relevancy of their category by a probabilistic factor (Fc). We measure this factor in terms of percentage using the formula as:

$$Fc = \frac{\text{number of word that hold that features}}{|\text{number of unique word in that category}|} x100\%$$

For instance in ‘NN’ groups, there are 40 unique words and out of 40, fifteen words are ended with “ive”, then the rank of that features to be grouped under ‘NN’ is 15/40*100=37.5%. For each rows, we compute Fc value for all tags and take the transition values from transition matrix. Finally, we compute the average of both and find the maximum value of that average. During tagging we take a word from the text, check it whether it falls under rare group and the if not, we analyse the features matrix and for each tag and sum with the rank given by previous tag probability from transition matrix (HMM) and assign that particular tag which give the maximum average. For example: if given a word ended with “ogy” (features-last three characters), we move to the second row of the feature matrix and read each value for each category and look in transition matrix for that category and compute the average. The features matrix is given in table 5.

3.2 The Transition Matrix

The transition matrix is a square matrix computed from confusion tags of the corpus. Here we consider fifteen confusion tags and compute the transition probabilities as given below:

Table 5: The Features-rank Matrix

Features/Tags		NN	NNS	JJ	NNP	VB	VBG	VBN	VBZ	VBD	RB	PRP$
Begins with cap		23.08	11.11	17.24	100.00	4.17	4.55	0.00	0.00	0.00	0.00	8.33
Beginning word of a line		20.51	11.11	0.00	17.65	8.33	9.09	0.00	0.00	0.00	0.00	0.00
All cap		5.13	0.00	0.00	47.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Suffixes	**s**	2.56	74.07	0.00	0.00	0.00	4.55	5.88	50.00	3.13	11.76	16.67
	es	0.00	11.11	0.00	0.00	0.00	0.00	0.00	33.33	0.00	0.00	0.00
	al	2.56	0.00	13.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ed	0.00	0.00	10.34	0.00	4.17	0.00	0.00	8.33	50.00	0.00	0.00
	ly	0.00	0.00	6.90	0.00	0.00	0.00	0.00	0.00	0.00	35.29	0.00
	tic	0.00	0.00	6.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ess	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ses	0.00	7.41	6.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ion	10.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ive	2.56	0.00	3.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ogy	2.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ism	2.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ing	0.00	0.00	0.00	0.00	4.17	63.64	52.94	0.00	0.00	0.00	0.00
Prefixes	**im**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	de	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	un	0.00	0.00	0.00	0.00	0.00	0.00	5.88	8.33	6.25	0.00	0.00
	in	5.13	10.53	6.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Transition probability is computed as:

$$p(t_2 | t_1) = \frac{C(t_1 \& t_2)}{C(t_1)}$$

For example determining the probability of noun following by a determiner is

$$p(noun | \det) = \frac{c(\det \& noun)}{c(\det)}$$

That is number of determiner and noun occurred together divided by the number of times that a determiner occurred. In table 6 we show the transition probability matrix of 15 confusion tags-

4. Findings

We test our model in various text corpora by inputting simple text file. Words which are not fall in our consideration, we don't tag them in confusion and we simply kept as untagged word. The accuracy matrix of our experiment is given in table 7. For instance we input a text consists of 627 words and it tag accordingly and found the following result. Out of 627 words, 40 are unable to tag and remain as untagged, which include those words that do not hold any features or may be foreign word. Some punctuation marks are also included in this category. Large numbers of words that are assigning wrongly because of less number of observation samples were taken for features consideration. For tags 'CC', 'IN', 'DT', 'RBR', 'TO', 'EX' etc., we achieve almost 100% accuracy due to their direct word consideration and we exclude these from our analysis. Considering only the confusion words, we achieve almost 70% accuracy. A very good figure of accuracy level is found in tags 'JJ', 'VBN', 'VBG' and 'VBZ' due to their unique features. The overall accuracy will increase up to 80% including rare word category.

Table 6: Transition Probability Matrix

	NN	NNP	NNS	CC	CD	DT	IN	JJ	PRP$	RB	VB	VBG	VBN	VBD	VBZ
NN	16.35	1.6	8.33	6.41	4.49	3.21	21.15	9.29	1.6	0.64	8.01	3.21	0.96	1.6	2.24
NNP	11.76	11.76	0	5.88	35.29	0	23.53	0	0	0	5.88	0	0	5.88	0
NNS	11.71	0.9	6.31	9.01	0	1.8	21.62	15.32	1.8	0.9	9.91	4.5	1.8	4.5	0
CC	20.97	3.23	4.84	0	0	12.9	4.84	17.74	4.84	1.61	4.84	1.61	1.61	8.06	3.23
CD	7.69	0	12.82	7.69	7.69	5.13	5.13	12.82	5.13	5.13	2.56	0	0	2.56	2.56
DT	58.26	0	4.35	0	1.74	0	0	22.61	0	6.96	0	0.87	1.74	0.87	0
IN	62.26	0.94	25.47	6.6	2.83	16.04	21.7	29.25	2.83	0.94	2.83	2.83	4.72	4.72	0
JJ	32.04	0.49	13.11	3.4	1.46	8.25	11.17	15.05	1.46	0.49	1.46	1.46	2.43	2.43	0
PRP$	26.47	0	5.88	5.88	0	0	5.88	8.82	0	2.94	20.59	0	0	5.88	8.82
RB	4.76	0	19.05	4.76	14.29	19.05	0	4.76	4.76	9.52	0	0	0	14.29	0
VB	7.61	1.09	2.17	4.35	1.09	8.7	9.78	13.04	5.43	1.09	2.17	4.35	11.96	17.39	0
VBG	14.29	3.57	39.29	0	0	3.57	0	17.86	0	0	7.14	0	0	3.57	0
VBN	10.71	0	10.71	0	0	3.57	10.71	25	7.14	3.57	3.57	0	3.57	14.29	0
VBD	8.47	0	0	10.17	1.69	10.17	25.42	23.73	0	0	0	3.39	0	1.69	0
VBZ	7.14	0	0	0	7.14	28.57	7.14	7.14	0	0	14.29	0	0	7.14	0

Some confused words are tagged as rare category for which the number of incorrect tags being increased specifically in ‘VB’, ‘NN’ tag, because these words directly match with the rare category.

Table 7: Accuracy level of confused tags

Tag	Number of word	Correctly tag	Wrongly tag	Accuracy %
NN	127	87	40	68.5
NNP	116	90	26	77.59
JJ	54	44	10	81.48
PRP$	46	34	12	73.91
VB	98	66	32	67.35
VBN	47	37	10	78.72
VBG	32	26	6	81.25
VBD	35	29	6	82.86
VBZ	24	21	3	87.5
Untagged	48			
Total	627	434	145	69.21

5. Conclusion

In this paper, we try to find a weight of a tag in two perspectives: one by features analysis and the other by using HMM. Feature extraction method could be enhanced using constraint grammar satisfaction method. Constraint Grammar (CG) as proposed by Fred Karlsson (in 1990) is a methodological paradigm for Natural language processing. In this approach linguist-written, context dependent rules are compiled into a grammar that assigns grammatical tags ("readings") to words or other tokens in running text. In future we suppose to imply this approach to achieve maximum weightage with higher order HMM. This method could be enhance to multiple perspective model which could include rule based, constraint satisfaction, HMM and features extraction method.

References

[1] Daniel Jurafsky & James H. Martin, WORD CLASSES AND PART OF SPEECH TAGGING, Speech and Language Processing: *An introduction to natural language processing, computational linguistics, and speech recognition.* 2005

[2] Spence Green, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning, Multiword expression identification with tree substitution grammars: a parsing tour de force with French**, in** *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ACL 2011

[3] Scott M. Thede and Mary P. Harper, A Second-Order Hidden Markov Model for Part-of-Speech Tagging, *In Proceedings of the 37th Annual Meeting of the ACL*

[4] Julia Hockenmaier , Building a (statistical) POS tagger: HMM POS-tagging with Hidden Markov Models at 3324 Siebel Center

[5] Adwait Ratnaparkhi, A Maximum Entropy Model for Part-Of-Speech Tagging, *In Proceedings of the Empirical Methods in Natural Language Processing* (1996), pp. 133-142.

[6] Mark Johnson, Why doesn't EM find good HMM POS-taggers? Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 296–305, Prague, June 2007. c 2007 ACL

[7] Eric Brill. 1992. A simple rule-based part of speech tagger. *In Proceedings of the third conference on Applied natural language processing* (ANLC '92).

[8] Adam R. Teichert , Hal Daume III, Unsupervised Part of Speech Tagging without a lexicon, School of Computing, university of Utah, 2009

[9] Akshar Bharati, Vineet Chaitanya Rajeev Sagal, Natural Language Processing, A Paninian Perspective, Prentice-Hall of India, New Delhi

[10] Doug Cutting and Julian Kupiec Jan Pedersen and Penelope Sibun, A Practical Part-of-Speech Tagger, *In Proceeding of the third conference on applied Natural language Procesing*, Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA (1992)

Kinship Terms in Nepali Language and its Morphology

Prajadhip Sinha
Assistant professor
Department Of Computer Sc.
Kohima Science College

Bairab Sarma
Research Scholar
Assam University
Silchar-Assam

Bipul Shyam Purkayastha
Professor
Assam University
Silchar-Assam

ABSTRACT

Kinship relations are blood relations or those relations which are created naturally. Kinship is a method of acknowledging relations and it is a social bond initiated by blood or genetic as well as marriage. Kinship is important in all cultures and in all human interaction. Kinship is important to anthropology because anthropology is the study of human behaviour and human behaviour is variable. According to M.Lamp[12] to understand fully the nature of kinship systems it is necessary to understand what kind of linguistic elements these are, and what kind of linguistic relationships. The kinship terminology of any language is a natural meeting point for the disciplines of anthropology and linguistics [6]. Like other languages of the world, Nepali kinship terms are also common for Nepali Language. Nepali[1] or Nepalese (नेपाली) is a language in the Indo-Aryan branch of the Indo-European language family. It takes its root from Sanskrit, the classical language of India. This paper explores the area of kinship terms in Nepali language, and its outlines the standard kinship relations, associated set of terms in the language. The formations of such terms are also elaborated with grammatical analysis. Kinship terms form a considerable part of the WordNet in any language because the kinship terms interacts each other with different relational characteristics of WordNet. Nepali is a morphologically rich language and Kinship terms form an important aspect in morphology of Nepali Language. There are many number of morphological studies in Nepali but most of them are descriptive in nature. Morphology of Nepali has not yet been fully analyzed from computational perspective.

Keywords

Kinship,Terminology,WordNet, Morphology, Computational

1. INTRODUCTION

Man is social by nature. He establishes many types of relations with a number of persons. The most important of these relationships is known as kinship and it is a method of acknowledging relationship. It is a social bond initiated by genetic or blood ties as well as marriage. Kinship ties are of fundamental importance in every society all over the world. Family is the point of departure for studying kinship. There are basically two types of kinship within a family and they are Affinal kinship and consanguineous kinship. Affinal kinship is based on marriage and most primary affinal relationship is one between a husband and a wife which in its extended form includes parents and siblings of both sides and their spouses and children. Consanguineal kinship based on descent is called consanguineal kinship, commonly known as blood relation. The relation between a child and his parents, between children of the same set of parents, between uncles and nephews/nieces are examples of consanguineous kinship. Like other languages the kinship terms are also common in the Nepali language. Kinship terms form an important aspect in morphology of Language. Nepali is the national language of Nepal. It is also a medium of a uniform, nationwide, educational system, public administration and mass communication. The most recent official census conducted by the government of Nepal in 2001 reports that there are around 20 million Nepali speakers in Nepal, out of which, it is spoken as the mother tongue by 11 million people, and as a lingua franca by others. Nepali or Nepalese (नेपाली) is a language[2] which takes its root from Sanskrit, the classical language of India. Nepali was previously known as Khas Kura and the language of the Khasa kingdom, which ruled over the foothills of current Nepal during the 13th and 14th centuries. The history of the usage of Nepali in writing dates as back as the 12th century AD. Nepali is written with the Devanagari alphabet, which developed from the Brahmi script in the 11th century AD. Linguistically, Nepali is most closely related to Hindi. A large proportion of the technical vocabulary is shared by Hindi and Nepali. Even the script is more or less the same for both languages and differing with each other in only a few minor details. It is the official language of Nepal and is also spoken in Bhutan, parts of India and parts of Myanmar (Burma) and in India; it is one of the country's 22 official languages[3] of India.

Nepali Kinship Terms

We describe the Nepali relations through both consanguinity and affinity. In this case, the consanguinity is the relation by blood or the connection of persons descended from the same stock or common ancestors. Furthermore, the consanguineal relations consist of two types of relation, viz. core and peripheral. On the other hand, the affinal relation is the relation made by marriage but not by blood. The affinal relations also consist of two types, viz. core and peripheral. For convenience, we have followed the common ethnological abbreviations[5] which are : [P= parents, M= mother, F= father, B= brother; Z= sister; S= son; D= daughter; H=husband; W= wife; E= spouse; G= siblings; C= child; e= elder; y= younger]

Consanguineal relations: The kinship relations are made by different ways. One among them is consanguineal relation which is made by blood. It is the relation among the same stock or common ancestors. The consanguineal relations can be classified into two types of relation, namely core and peripheral.

Representing kinship terms of core consanguineal relations: The relations made by the ego directly are the core consanguineal relations. The core consanguineal relations, as

[1] http://en.wikipedia.org/wiki/Nepali_language

[2] http://en.wikipedia.org/wiki/Nepali_language

[3] At present there are 22 languages listed under the ES **(Eighth Schedule)**. Three languages, namely Manipuri, Konkani, and Nepali, were included in ES in 1992 through the Seventy-First Amendment

its name suggests are really the core relations and consist of ego's parents, siblings and offspring. For these relations, there are the following kinship terms[2] .

Kinship relation	Kinship terms	Generation from ego
F	Baba	G+1
M	Amma	G+1
eB	Daju	G=0
yB	Bhai	G=0
eZ	Didi	G=0
yZ	Buhari	G=0
C	Nani	G-1
S	Chhora	G-1
D	Choori	G-1

TABLE 1-- Representing kinship terms of core consanguineal relations

The core consanguineal relations are father, mother, elder/younger brother, elder/younger sister, child, son and daughter. So there are nine terms for core consanguineal relations in Nepali. There are not sex and age neutral terms in Nepali which are found in English like parents, brother, sister, etc. In Nepali, the terms, viz. daju (eB), Bhai (yB) make the age distinction and the terms, viz. Daju (eB), Didi (eZ) make the sex distinction.

Representing kinship terms of peripheral consanguineal relations

The peripheral consanguineal relations are not the direct relation of the ego but the ego's relations through the core consanguineal relations are called the peripheral consanguineal relations. It is also called the relations through relations. The peripheral consanguineal relations can be interpreted by a number of relations which are as follows:

Through parents' up generation:

Kinship relation	Kinship terms	Generation from ego
FFF, FMF, MFF, MMF	Jiju baje	G+3
FFM, FMM, MFM, MMM	jiju boju	G+3
FF, MF	baje	G+2
FM, MM	boju	G+2

TABLE 2-- Representing kinship terms through Parents' up generation

The peripheral consanguineal relations made by parents' up generation are father's/mother's father's father, father's/mother's mother's father, father's/mother's father's mother, father's/mother's mother's mother, father's/mother's father, father's/ mother's mother.

Through father:

Kinship relation	Kinship terms	Generation from ego
FeB	thula babu	G+1
FeBW	thuli ama	G+1
FyB	kaka/kancha babu	G+1
FyBW	kaki/kanchhiama	G+1
FeZ/ FyZ	Phupu	G+1
FZH	Phupaju	G+1

TABLE 3-- Representing kinship terms through father

The peripheral consanguineal relations through father are father's elder/younger brother, father's elder/younger sister. These terms make a clear age distinction which is dissimilar to English terms uncle for FB and aunt for FZ.

Through mother:

Kinship relation	Kinship terms	Generation from ego
MeB	Mama	G+1
MyB	Mama	G+1
MeZ	Thool-ama	G+1
MyZ	Chema	G+1

TABLE 4-- Representing kinship terms through mother

In Nepali, there is similar term mama for mother's elder/younger brother (MeB, MyB), which is an age-neutral

Through father's/mother's siblings:

Kinship relation	Kinship terms	Generation from ego
FBeS, FZeS,MBeS, MZeS	Daju	G=0
FByS, FZyS MByS, MZyS	Bhai	G=0
FBeD, FZeD, MBeD, MZeD	Didi	G=0
FByD, FZyD, MByD, MZyD	bohini	G=0

TABLE 5-- Representing kinship terms through father's/mother's siblings

The peripheral consanguineal relations through father's siblings are father's brother's/sister's elder/younger son and father's brother's/sister's elder/younger daughter. The relations made by father's siblings are upholding both age and sex distinction. Elder and younger are distinctly noticed by the like ‘daju’ and ‘bhai’ which is dissimilar to the terms found in English like cousin for both sex.

Through male ego:

Kinship relation	Kinship terms	Generation from ego
BS	Bhatija	G-1
BD	Bhatiji	G-1
ZS	Bhanja	G-1
ZD	Bhanji	G-1

TABLE 6-- Representing kinship terms through male ego

Through female ego:

Kinship relation	Kinship terms	Generation from ego
BS	Bhoda	G-1
BD	Bhodi	G-1
ZS	Chhora	G-1
ZD	Choori	G-1

TABLE 7-- Representing kinship terms through female ego

The peripheral consanguineal relations made by female ego are the same to the relation made by male ego. But the terms are different. The relations, namely brother's/sister's son and brother's/sister's daughter by male ego and female ego are termed differently because of sex distinction.

Through child:

Kinship relation	Kinship terms	Generation from ego
SS, DS	Nati	G-2
SD, DD	Natini	G-2

TABLE 8-- Representing kinship terms through child

The peripheral consanguineal relations made by child are overlapping. Son's son and daughter's son are called by the same term nati. Similarly, son's daughter and daughter's daughter are also called by the term natini.

Affinal relations

The relations made by marriage but not by blood are called the affinal relations. In this way, there are two major ways to establish the relations although there are other ways for establishing the relation. Similar to the consanguineal relations, the affinal relations also have two further types. They are core an peripheral which are described below.

Representing kinship terms of core affinal relations

The concept of the core affinal relation is similar to the core consanguineal relations but they are considerabley different to each other. The affinal relations (made by marriage) through the core consanguineal relations are called the core affinal relations. The core affinal relations are made by the ego's father, mother, siblings and offspring. They are gradually treated in the following ways.

Through father:

Kinship relation	Kinship terms	Generation from ego
FeBW	Thuloama	G+1
FyBW	Kaki	G+1
FeZH	Phupaju	G+1
FyZH	Phupaju	G+1

TABLE 9-- Representing kinship terms through father

The affinal relations made by father are father's elder/younger brother's wife and father's elder/younger sister's husband.

Through mother:

Kinship relation	Kinship terms	Generation from ego
MeBW	Maiju	G+1
MyBW	Maiju	G+1
MeZH	Thoolababa	G+1
MyZH	Thoolababa	G+1

TABLE 10-- Representing kinship terms through mother

The affinal relations through mother are mother's elder/younger brother's wife and mother's elder/younger sister's husband.

Through the ego's siblings:

Kinship relation	Kinship terms	Generation from ego
GeBW	Buhari	G=0
GyBW	Bhauju	G=0
GeZH	Bhinaju	G=0
GyZH	Juwai	G=0

TABLE 11-- Representing kinship terms through ego's siblings

The affinal relations through ego's siblings are siblings' elder/younger brother's wife and siblings' elder/younger sister's husband. For these relations, there are four different

Terms in nepali.

Through the ego's child

Kinship relation	**Kinship terms**	**Generation from ego**
SW, SSW, DSW	buhari	G=0-1-2
DH,	Juwai	G=0-1-2
DDH, SDH	nati Juwai	G=0-1-2

TABLE 12-- Representing kinship terms through ego's child

The terms buhari is common term for SW, SSW, and DSW.

Representing kinship terms of peripheral affinal relations

The affinal relations through peripheral consanguineal relations are known as the peripheral affinal relations. The peripheral affinal relations are also established through different relations which are treated below.

Through parents' siblings:

Kinship relation	Kinship terms	Generation from ego
PZDeH, PBDeH	Juwai	G=0
PZDyH, PBDyH	Juwai	G=0
PZSeW, PBSeW	Bouji	G=0
PZSyW, PBSyW	Buhari	G=0

TABLE 13-- Representing kinship terms through parents' siblings

The affinal relations through parent's sibling are parent's brother's/sister's elder daughter's husband, parent's brother's/sister's younger daughter's husband, parent's brother's/sister's elder son's wife and parent's brother's/sister's younger son's wife.

Through parents' child:

Kinship relation	**Kinship terms**	**Generation from ego**
SWF, DHF	Samdhi	G-0
SWM, DHM	Samdhini	G-0

TABLE 14-- Representing kinship terms through parents'child

The peripheral affinal relations made by parents' child are son's wife's/daughter's husband's father, son's wife's/daughter's husband's mother. These are cross relations and reciprocal terms.

Through ego's siblings

Kinship relation	**Kinship terms**	**Generation from ego**
BSW	bhatija buhari	G-1
BDH	bhatiji Juwai	G-1
ZSW	bhanja buhari	G-1
ZDH	bhanji Juwai	G-1

TABLE 15-- Representing kinship terms through ego's siblings

The peripheral affinal relations made by ego's siblings are brother's/sister's son's wife and brother's/sister's daughter's husband.

Through ego's wife:

Kinship relation	**Kinship terms**	**Generation from ego**
W	Svasni	G=0
WeB	Jethan	G=0
WeBW	Jethandidi	G=0
WyB	Salo	G=0
WyBW	Bhaini	G=0
WyZ	jethi sasu	G=0
WeZH	Daju	G=0
WyZ	Sali	G=0
WyZH	Sadubhai	G=0

TABLE 16-- Representing kinship terms through ego's wife

These peripheral affinal relations, as mentioned in the table are made by ego's wife perspective. The ego is male and the relations are established through his wife. The relations through wife are wife, wife's elder brother and his wife, wife's younger brother and his wife, wife's elder sister and her husband, wife's younger sister and her husband.

Through ego's husband:

Kinship relation	**Kinship terms**	**Generation from ego**
H	Logne	G=0
HeB	Jethaju	G=0
HeBW	Jethani	G=0
HyB	Devar	G=0
HyBW	Devarani	G=0
HeZ	Amaju	G=0

HeZH	Buwa	G=0
HyZ	Nanda	G=0
HyZH	Nandebhai	G=0

TABLE 17-- Representing kinship terms through ego's husband

The relations made by ego's husband are more similar to the relations made by ego's wife. But there are different terms for some relations like wife's younger brother (WyB) and husband younger brother (HyB).

Through brother's wife and sister's husband:

Kinship relation	Kinship terms	Generation from ego
BWeB, SHeB	Buwa	G-0
BWyB, SHyB	Saiba	G-0
ZWeB, ZHeB	buwa	G-0
ZWyB, ZHyB	Saiba	G-0

TABLE 18-- Representing kinship terms through brother's wife and sister's husband

The peripheral affinal relations made by brother's wife and sister's husband are brother's wife's/sister's husband's elder/younger brother, sister's wife's/sister's husband's elder/younger brother.

Through his or her spouse:

Kinship relation	Kinship terms	Generation from ego
EFF, EMF	Bajsosura	G+2
EFM, EMM	Bojusasu	G+2
EF,	Sosura	G+1
EFeB	thula babusosura	G+1
EFeBW	thuliamasasu	G+1
EFyB	kanchhasosura	G+1
EFyBW	kanchhisasu	G+1
EMyZ/ EMeZ	Sanimasasu	G+1
EMe/yB	mamasosura	G+1
EMe/yBW	Maijusasu	G+1
EFe/yZ	Phupusasu	G+1
EFe/yZH	phupajusosura	G+1
EMZe/yH,	sanababusora	G+1
EM	Sasu	G+1

TABLE 19-- Representing kinship terms through his or her spouse

The morphology of Nepali kinship terms

Morphology (linguistics)[4] means, the study of the structure and content of word forms. Nowadays morphology is considered an autonomous component on the same footing as syntax and semantics. The term morphology[5] is Greek and is a makeup of morph- meaning 'shape, form', and -ology which means 'the study of something'. The term is used not only in linguistics but also in biology as the scientific study of forms and structure of animals and plants, and in geology as the study of formation and evolution of rocks and land forms. We are going to stick to morphology in linguistics, as the scientific study of forms and structure of words in a language. The knowledge of morphology is necessary in order to know the way the human brain works and processes language. It will help to produce new alternatives to learn languages and it will permit its application to artificial intelligence.

The words are considered to be fundamental building blocks of language. A word (i.e. word form), in real sense, can either be in simple, complex, compound or reduplicated. A simple word consists of a root or stem together with suffixes or prefixes. A compound term can be broken up into two or more independent terms. Nepali has a number of kinship terms in which female gender is indicated by a final suffix i.e /-i/ or /-ni/, such as chorā 'son' versus chorī 'daughter', and is also used to indicate biological sex in non-human animates, such as kukur 'hound' versus kukurnī 'bitch'[7].

Nouns in Nepali in which the gender is indicated either by certain change or by some sorts of marker are said to have morphological gender. A small set of Nepali kinship terms(nouns)such as काका kaka 'paternal uncle' takes -◌ी-iː and changes to feminine as काकी kakiː 'aunt' Another set of terms such as नाति nati 'grandson' changes to its feminine form as नातिनी natiniː by taking suffix -नी-niː

Masculine	Gloss	Feminine	Gloss	Gender marker
काका\ kaka	Uncle	काको \kaki	Aunt	/-i
नाति \ nati	Grandson	नातिनी\ natini	grand daughter	/-ni
Devar	Brother in law	Devarani	Sister in law	/-ni
छोरा \Chorā	Son	छोरी\Chori	Daughter	/-i
भाई\Bhai	Brother	बहिनि\Bhoini	Sister	/ni

Table 20 -- Representing Morphological gender

In the above Table20 o-ending and non-o-ending terms in Nepali that inflect for feminine gender by various ways are demonstrated [4]). Another popular suffixes in nepali[6] are

[4] http://en.wikipedia.org/wiki/Morphology

[5] Aronoff_sample chapter_What is morphology.pdf

[6] Nepal Bhasa - Wikipedia, the free encyclopedia.mht

chaa and ju. "Chaa" is added to signify "junior" or "lesser". But when added to a name, it is used derogatorily. For example, kya'ah-chaa means nephew where "chaa" is being added to kya'ah(son). When added to name like Amit for "Amitt-chaa", it is being used derogatorily. The suffix "ju" is added to show respect. For example, "Baa-ju" means "father-in-law" where "ju" is added to "Baa(father)". Unlike "chaa", "ju" is not added to a first/last name directly. Instead, an honorific term like “Bhaaju” is added for males and "Mayju" for females. Example, "Birat bhaaju" for a male name (Birat) and "Suja Mayju" for a female name (Suja).

Prefix – "Tap'ah" is added to denote "remote" or "distant" relative ('distance' in relationship irrespective of spatial extent). A distant (younger) brother (kija) becomes "tap'ah-kija". "Tuh" is added to denote "higher". Father (baa)'s senior brother is referred to as "Tuh-baa".

In Nepali kinship terminology, as in the lexicon in general, biological gender can be emphasized or disambiguated with of the two gendered suffixes “Sosura" (EF) and "sasu" (EM) in Nepali. It may be observed the addition of “sosura” and “sasu” is not only used for distinguishing the sex of the kinsmen but also found attached to the kin terms. The ‘sosura’ suffix generally added with male and ‘sasu’ with female gender .For example Jethasosura(HeB) husband elder brother and jethisasu(HeBW) and both derived from the kinship term jetha(FeB). Some examples with their corresponding gender marker are given below:

Noun(kinship term)	Suffix/gender marker
Budhasosura	sosura<male>
Budhisasu	sasu<female>
kanchhasosura	sosura <male>
Kanchhisasu	sasu<female>
Jethasosura	sosura<male>
Jethisasu	sasu<female>
Mamasosura	sosura <male>
Maijusasu	sasu<female>
Thulobabusosura	sosura <male>
Thuliamasasu	sasu<female>
Sanobabusosura	sosura <male>
Saniamasasu	sasu<female>
phupajusosura	sosura <male>
Phupusasu	sasu<female>

Table 21-- Representing gender marker

Doubling, reduplicative, near reduplicative or mirror forms is a common and well-attested feature of kinship terminologies[7] in many of the world’s languages, including Nepali. In reduplication, the reduplicate is most often repeated only once. However, in some languages, reduplication can occur more than once. The kinship term ‘kaka’ (father’s younger brother) is a reduplication form of ‘ka.’ And other examples are ‘mama’ (mother’s brother),’ didi’(‘elder sister) etc.

A word can either be in simple, complex, or compound. A simple word consists of a root or stem together with suffixes or prefixes. A compound term can be broken up into two or more independent terms. Same rule is applicable for Nepali Kinship terms. Some Nepali kinship term consists of one or more independent kinship terms. ‘Mama’ and ‘sosura’ are two independent kinship term in Nepali language. ‘Mamasosura’ also another kinship term in Nepali which is the combination of previous two kinship term Mama(MB) and sosura(EF). Some examples with their corresponding word form (simple & Compound) are given below:

Jethasosura(Compound)--→ Jetha(simple) +Sosura(simple)

Kanchhasosura(Compd)→Kanchha(simple)+Sosura(simple)

Kanchhisasu (Compound)-→ Kanchhi(simple)+Sasu(simple)

In Nepali, there are two numbers and they are singular and Plural. For the plural of noun, generally it adds the suffix-haru to the end of a word.

Nepali Singular	Nepali plural
Bhai(yB)	Bhaiharu
Daju(eB)	Dajuharu
Didi(eZ)	Didiharu
Chhora(S)	Chhoraharu

Table 22- Representing Number

Nepali kinship terms are usually closely linked to specific pronouns which in turn require specific verbal endings [6]. In other words, when addressing one’s own father or someone of that age, one should use the kinship term Phupu and the pronoun tapai.

Noun (Kinship term)	Pronoun:	Verbal ending:
Phupu/ Ama, dai, didî	Topai/Hajur	-nuhuncha (indicative present) -nuhos (imperative)
Bhai, Bahina	Timi	-chau (indicative present) -au ; -a (imperative)
Babu, Nani	Timita	-chau (indicative present) -au ; -a (imperative) -chas (indicative present) -Ø (imperative

Table 22—Relation between Noun, pronoun & verb of kinship term

On the other hand, in a conversation with one’s younger sister, one would use the term bahinî and most likely the personal pronoun timi. The table 22 above is a provisional attempt at categorizing these relationships.

2. Conclusion:

Kinship terms generally replace an individual’s given name, both as a term of address and for reference. Moreover the metaphorical usage of kinship terms to non-kin is widely observed among the Nepali language. Formation of kinship terms in Nepali language and its morphology have been discussed in the paper. It is seen that the prefixes and suffixes play a major role in formation of the kinship terms in Nepali Language. Moreover for formation of kinship terms it follows some pattern or paradigm in Nepali language. This report can serve as a strong base document for Nepali WordNet[1] development in Nepali language.

3. References

[1] Alok chakraborty, Bipul Shyam Purkayastha, Arindam Roy (2010), Experiences in building the Nepali Wordnet, Proceedings of the 5th Global WordNet Conference, Mumbai.India.

[2] Anju Giri (2001): English and nepali kinship terms Journal NELTA, Vol 6,No 1.

[3] Arindam Roy,Sunita Sarkar,Bipul Syam Purkayastha,(2012): A Proposed Nepali Synset entry and extraction tool, Proceedings of the 6th Global WordNet Conference, Matsue, Japan

[4] Balaram prasain (2012): A computational analysis of Nepali morphology: A model for natural language processing.

[5] Ichchha Purna Rai (2010) : Bantawa kinship Terminology.

[6] Mark Turin (2001): Call me uncle: an outsider’s experience of nepali kinship.278 CNAS journal, vol 28,no 2.

[7] Mark Turin(2004) : ‘Thangmi kinship terminology in comparative perspective’, pp. 101-139 in Anju Saxena, ed., Himalayan Languages: Past and Present. Berlin Mouton.

[8] Prajwal Rupakheti, Laxmi Prasad Khatiwada Bal Krishna Bal (2006): Report on Nepali Computational Grammar.

[9] Shikhar kr. Sarma, Biswajit Brahma, Mane Bala Ramchiary(2010): Formation of kinship terms in Bodo language.

[10] Satarupa Dattamajumdar, Kinship Terminology in Lepcha (2010):The Buckingham Journal of Language and Linguistics Volume 3 pp 179-185.

[11] Shikhar Kr. Sarma, Utpal Saikia and Mayashree Mahanta:(2010): Kinship terms in Assmese Language.

[12] Sydney M.LAMP(1965) New Series, Vol. 67, No. 5, Part 2: Formal Semantic Analysis (Oct., 1965), pp. 37-64 Kinship terminology and linguistic structure.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 5, May 2015)

Enhancing the Performance of Part of Speech tagging of Nepali language through Hybrid approach.

Prajadhip Sinha[1], Nuzotalu M Veyie[2], Bipul Syam Purkayastha[3]
[1,2]Asst. Professor, Department of Computer Science, Kohima Science College, Kohima, Nagaland
[3]Professor, Department of Computer Science, Assam University, Silchar, Assam, India.

***Abstract*—Part-of-speech tagging is the process of marking up the words in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context —i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. Part-of-Speech (POS) tagging is the process of assigning the appropriate part of speech or lexical category to each word in a natural language sentence. Part-of-speech tagging is an important part of Natural Language Processing (NLP) and is useful for most NLP applications. It is often the first stage of natural language processing following which further processing like chunking, parsing, etc. are done. There are a number of approaches to implement part of speech tagger [1], i.e. Rule Based approach, Statistical approach and Hybrid approach. Rule-based tagger uses 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 a combination of Rule based approach and Statistical approach. In this paper, we have proposed a Hybrid approach using Hidden Markov Model (statistical approach) integrated with Rule-Based method towards POS tagging and achieved the accuracy of 93.15%.**

***Keywords*— Corpus, POS, NLP, chunking, parsing, Rule based, Statistical, Hybrid, ambiguity, HMM.**

I. INTRODUCTION

Part of Speech (POS) Tagging is the essential basis of Natural Language Processing (NLP). It is the process in which each word is assigned to a corresponding POS tag that describes how this word is used in a sentence. The development of an automatic POS tagger requires either a comprehensive set of linguistically motivated rules or a large annotated corpus[1]. But such rules and corpora have been developed for a few languages like English and some other languages. POS taggers for Indian languages are not readily available due to lack of such rules and large annotated corpora and moderate accuracy can only be achieved in rule based techniques.

To overcome these problems, we propose Nepali Part-Of-Speech Tagging method based on hybrid approach which combines the rule-based approach with a statistical approach that relies on the Nepali sentence structure improving the POS Tagging.

Nepali is an Indo-Aryan language spoken in Nepal, Bhutan, and some parts of India and Myanmar [2]. It is the national language of Nepal and also one of 23 Official languages of India incorporated in 8th annex of the Indian Constitution.

II. LITERATURE SURVEY

In this section, we would be focusing on the work done in the Indian context instead of discussing POS tagging approaches and efforts of implementing a POS tagger in general. POS tagging efforts in Indian context dates back to 1990s with Bharti et. al.[3] proposing a POS tagger for Hindi with morphological analyzer where a morphological analyzer would first provide a root word with its morphological features and a general POS category with can then be further classified using this generic pos category and morphological features. This approach was slightly modified by Singh et. al. [4] where they used the results of morphological analysis for training using a decision tree based classifier. Their tagger gave an accuracy of 93.45%. Dalal et. al. [5] used a pure maximum entropy based machine learning approach for labelling Hindi words with various POS tag categories. This tagger reported to have 88.4% accuracy Shrivastava and Bhattacharya [6] proposed an approach where instead of developing a full morphological analyzer, they used a stemmer to generate suffixes which was then used to generate POS tags. Their tagger reported 93.12% accuracy. Agarwal and Amni [7] and Avinesh and Gali [8] used Conditional Random Fields (CRF) with morphological analyzer to train their tagger. Agarwal and Amni's tagger reported an accuracy of 82.67% and Avinesh and Gali's tagger reported an accuracy of 78.66%.

NELRALAC[9] tagset is the first work in developing Nepali tagset which consist of 112 tags. This tagset has been compiled with reference to widely published grammars of Nepali. This tagset was used to tag (Nepali National Corpus) NNC manually and semi manually. As showed that error rates of annotation could be much higher with a large tagset, the reason primarily being the chances of assigning incorrect tags to the words out of confusion while manually annotating the training data itself.

A. Rule based Tagger: Rule based part of speech tagging is the approach that uses handwritten rules for tagging. Rule based tagger depends on dictionary or lexicon to get possible tags for each word to be tagged. Hand-written rules are used to identify the correct tag when a word has more than one possible tag. These rules are often known as context frame rules. Disambiguation is done by analysing the linguistic features of the word, its preceding word, its following word and other aspects. For example, if the preceding word is an adjective then the word in question must be adjective or noun. This information is coded in the form of rules.

B. Statistical taggers: These tagging algorithms apply probabilistic methods. They are usually trained by using a tagged corpus. They learn the POS tags of the words and their probabilities from the corpus and also they learn the distributional probability of the word. When these taggers encounter an unknown word they use distributional information of the word to suggest a tag for it. Statistical taggers are of high accuracy but their performance is difficult to improve. Also they need a tagged training corpus which would be unavailable in some languages.

Hidden Markov Model is a well-known stochastic based approach and probability is the basic principle behind HMM.

The intuition behind all stochastic taggers is simple generalization of the “pick the most-likely tag for this word”. For a given sentence or a word sequence, HMM tagger chooses the tag sequence that maximizes:

$$P\ (\text{word} \mid \text{tag}) * P\ (\text{tag} \mid \text{previous n tags}).$$

HMM tagger generally chooses a tag sequence for a given sentence rather than for a single word. This approach assumes that we are trying to compute the most probable tag sequence of tags T = (t1,t2,…,tn) for a given sequence of words in the sentence W = (w1,w2,…,wn)

III. System Description

This system is developed using hybrid based approach and the proposed POS tagging is implemented by undergoing several distinct steps. Given below is the flow diagram and its related explanation:

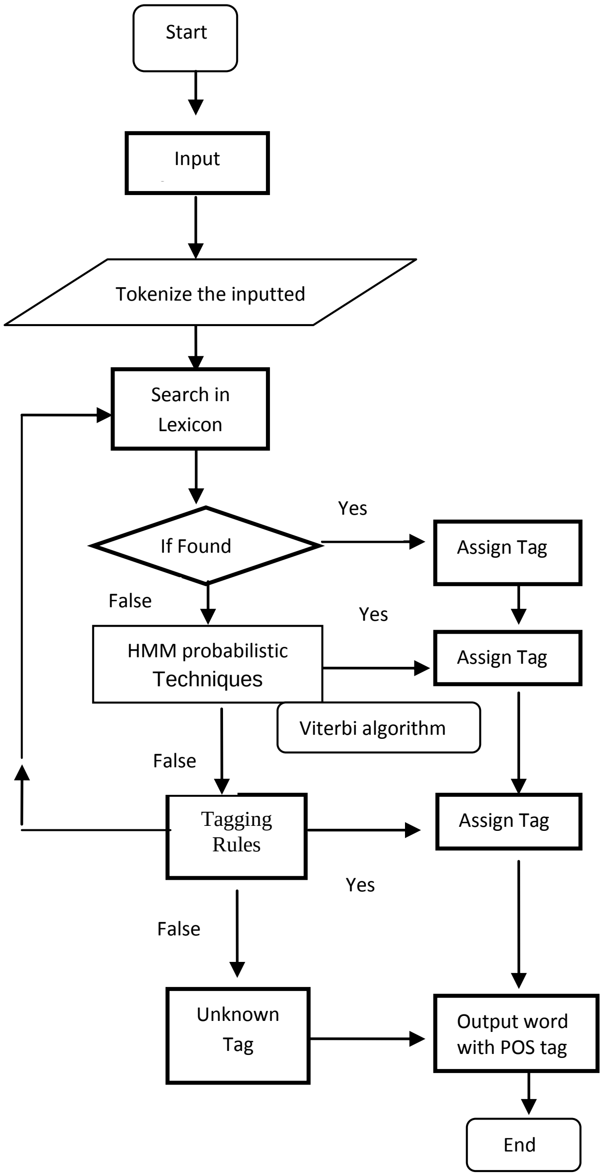

Fig 1: Proposed System Design of Nepali POS Tagger

1. *Tokenizer:* Tokenization is the first step in part of speech tagging of any natural language. It segregates words, punctuation marks and symbols of an input text, and subsequently assigns them into tokens by creating whitespaces between them.
2. *Lexicon:* In this step, a lexicon of word list in Nepali language is defined. This lexicon includes all categories of POS tagging viz., prepositions, adverbs, conjunctions, interrogative particles, etc. Initially a tagged lexicon is developed manually by collecting limited words from Nepali newspapers, books and dictionaries. All the words in the input text have to pass through this phase. If a word is found in the lexicon, then the entered word itself will be assigned with an appropriate tag. Else, it passes to the next step (HMM probabilistic Techniques).
3. *Hidden Markov Model (HMM) probabilistic Techniques:* A HMM[11] is Statistical Model which can be used to generate tag sequences. Basic idea of HMM is to determine the most likely tag sequences[10]. For this purpose we have to calculate Transition probability. Transition probability shows the probability of travelling between two tags i.e. forward tag and backward tags. The Transition probability is generally estimated based on previous tags and future tags with the sequence provided as an input. The following equation (1) explains this idea-

$$P\ (ti/wi) = P\ (ti/ti-1).\ P\ (ti+1/ti).\ P\ (wi/ti)............\ (1)$$

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
It is calculated as-

$$P\ (wi/ti) = freq\ (ti, wi)/\ freq\ (ti)\ldots....................\ (2)$$

This is done because we know that it is more likely for some tags to precede the other tags. In HMM we consider the context of tags with respect to the current tag. It 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 which can decide the tag for a word by looking at the tag of the previous word and the tag of the future word.

4. *Viterbi algorithm*: **The** main idea of the Viterbi algorithm is that instead of iterating over all possible state sequences to choose the best state sequence, we iterate over all possible candidates of each state to get the best one for that individual state.
The concatenation of the best individual states produces the best state sequence. Applied to tagging, this algorithm searches for the best tag for each word in order to find the best tag sequence
5. *Rule based tagging:* Almost all the words are recognized by the previous two phases. However, some disambiguated words require to be further analyzed, which can be resolved with the help of Rule based tagging.

IV. Algorithm Used For This Tagging Is As Follows

Step 1: Input the Nepali text.

Step 2: Tokenize the input text.

Step 3: Search for the tokens in lexicon. If the word is in the form of affixation, derivation and compounding then feed the word to segmenter for splitting and checks the word with the lexicon for a match.

Step 4: If the word is found in the lexicon then the entered word itself will be the output as appropriate tag.

Step 5: If match is not found or multiple tags exists for a single word then tagger tagged the word by using HMM probabilistic Techniques. HMM probabilistic Techniques determined the frequency of each word and two words sequence in the corpus.

Step 6: Returned the tagged output text.

Step 7: Extract those terms which has been misclassified or unanalysed during Step 5.

Step 8: Make the new entry for the unknown new word to the lexicon.

Step 9: Apply POS rules for those words and returned the tagged output text

Step 10: Repeat step 5 and 9 till the end of the input text.

Step 11: Display the tag output and save tag structure into HTML file.

V. RESULT

In order to measure the performance of the system, we use tag set consisting of 43 grammatical tags[8] and create corpus with different number of words collecting from Nepali Newspaper and other sources.

Initially, a very limited lexicon was present in POS tagger and its accuracy was low.

However, when more text is tagged and manual corrections are done for those words that are new words to lexicon, the lexicon will grow. After some time, the accuracy level will also be increased. All these are summarized in the Table 1,2,3.

Table1: Tagging accuracy for 5000 Tokens

Experiment set-1 5000(Tokens)			
Exp	Total words	Correctly tagged	Accuracy
1	1000	506	50.6
2	1500	755	50.333
3	2000	997	49.85
4	2500	1269	50.76
5	3000	1538	51.26
6	3500	1829	52.25
7	4000	2210	55.25
8	4500	2512	55.82
9	5000	2903	58.06

Table2: Tagging accuracy for 10000 Tokens

Experiment set-2 10000(Tokens)			
Exp	Total words	Correctly tagged	Accuracy
1	1000	682	68.2
2	1500	1036	69.066
3	2000	1420	71
4	2500	1708	68.32
5	3000	2078	69.266
6	3500	2445	69.85
7	4000	2734	68.35
8	4500	3151	70.022
9	5000	3523	70.46

Table3: Tagging accuracy for 15000 Tokens

Experiment set-3 15000(Tokens)			
Exp	Total words	Correctly tagged	Accuracy
1	1000	922	92.2
2	1500	1392	92.8
3	2000	1837	91.85
4	2500	2301	92.04
5	3000	2803	93.43
6	3500	3275	93.57
7	4000	3790	94.75
8	4500	4211	93.57
9	5000	4706	94.12

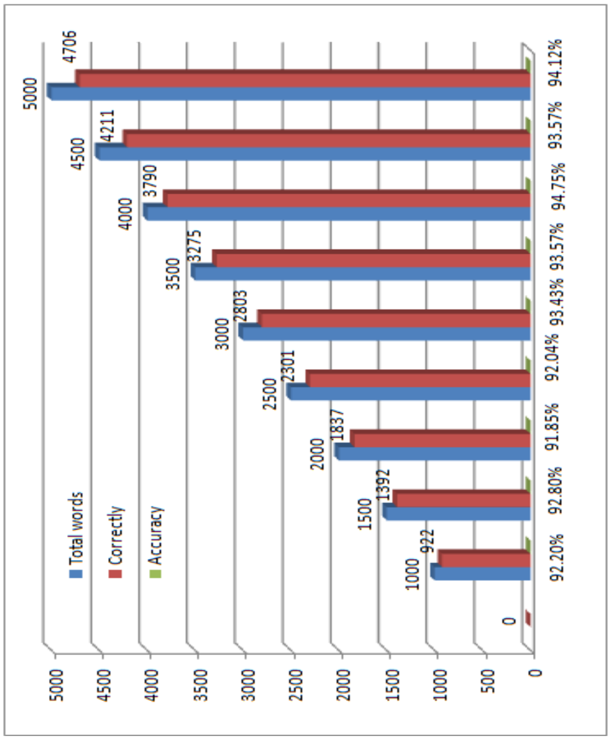

Fig 2.: Part-of-speech tagging accuracy percentage for 15000 tokens

A. Performance of pos tagging for different approaches

No of words	Rule Based	HMM	Our Approach
5,000	44.43%	48.23%	53.06%
7,500	51.00%	54.60%	61.67%
10,000	56.0%	60.19%	69.32%
12,500	64.23%	67.67%	77.12%
15,000	71.89%	80.56 %	93.24%

Fig 3. Tagging accuracies on different word

VI. EVALUATION

For testing the performance of our system, we developed a test corpus of 550 sentences (15,720 words). We calculated precision, recall and performance indicators of a system. Precision or accuracy of the POS tagging is calculated using the following equation:

$$\text{Precision} = \frac{No\ of\ correct\ POS\ tags\ assigned\ by\ the\ system}{No\ of\ POS\ tagsassmigned\ by\ the\ syste}$$

The present lexicon size is approximately 15720. As such the computation of accuracy of the system is based on the results derived in experiment set 3; the values ranging from 92.2 to 94.2. Thus the POS tagging of the Nepali language through hybrid approach is yielding an accuracy of 93.15%.

INPUT TEXT: छ वर्ष सम्म टि .मार्शल हान जुनियर ले भद्र र शान्त जर्ज बुस शैली मा निगम अभिग्रहण गरे । अब को प्रश्न : उहाँ कठोर रूप मा काम गर्ने टेडी रुजेवेल्ट ले जस्तो काम गर्न सक्नुहुन्छ ? जर्जिया-प्यासिफिक कर्पोरेसनका अध्यक्ष तथा मुख्य कार्यकारी अधिकृत ६२ वर्षीय श्री हानले ग्रेट नर्दन नेकुसा कर्पोरेसन को लागि वन्य-उत्पादन विषय को ३.१९ विलियन डलरको अप्रार्थित बोलपत्र हेरिरहनु भएको छ ।

OUTPUT: छ/CD वर्ष/NN सम्म/POP टि/.FB मार्शल/NNP हान/NNP जुनियर/NNP ले/PLE भद्र/JJ र/CC शान्त/JJ जर्ज/NNP बुस/NNP शैली/NN मा/POP निगम/NN अभिग्रहण/NN गरे/VBO अब/RBO को/PKO प्रश्न/NN :/YM उहाँ/PP कठोर/JJ रूप/NN मा/POP काम/NN गर्ने/VBNE टेडी/NNP रुजेवेल्ट/NNP ले/PLE जस्तो/JJM काम/NN गर्न/VBI सक्नुहुन्/VBF ?/YF जर्जिया/NNP प्यासिफिक/NNP कर्पोरेसनका/NN अध्यक्ष/NN तथा/CC मुख्य/JJ कार्यकारी/JJ अधिकृत/NN ६२/CD वर्षीय/JJ श्री/NN हानले/NN ग्रेट/NNP नर्दन/NNP नेकुसा/NNP कर्पोरेसन/NN को/PKO लागि/POP वन्य/NN उत्पादन/NN विषय/NN को/PKO ३/१९.CD विलियन/CD डलरको/NN अप्रार्थित/JJ बोलपत्र/NN हेरिरहनु/VBI भएको/VBKO छ/CD ,/YM नेकुसाले/NN प्रस्तावलाई/NN सार्वजनिक/JJ उदासिन/JJ व्यवहार/NN गरेको/VBKO छ/VBF। /YM

Graphical User Interface Tool: The researcher has highlighted his work in Nepali POS tagging on the website create by him www.researchnlp.com.A Graphical User Interface tool named "POSTIN" has been developed by using PHP.

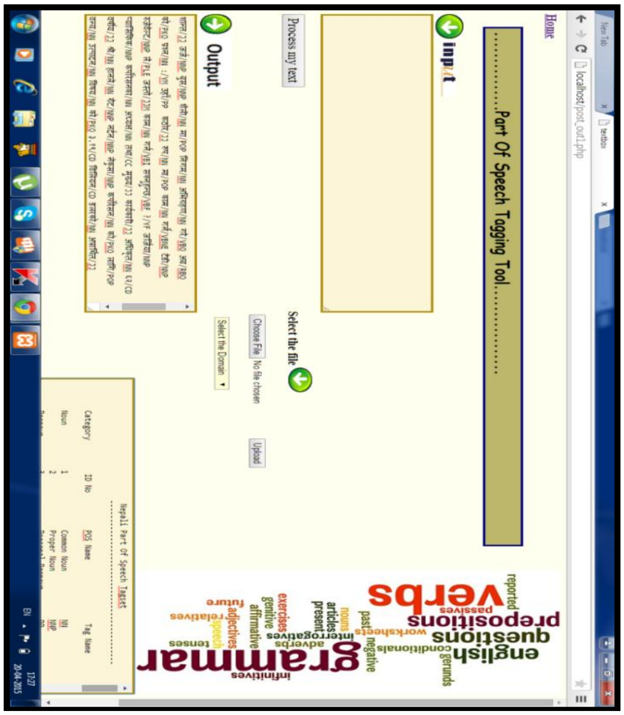

Fig4: POS Tagging Tool

VII. Conclusion And Future Work

In this paper, we have presented an approach for part-of-speech (POS) Tagger of Nepali language using statistical approach. The developed tagger employed an approach that combines the Rule-based method with Hidden Markov Models (HMMs). This tagger significantly outperforms Natural Processing System which assigns to a word the most likely tag assigned to that word in the training corpus To evaluate the accuracy of the proposed POS Tagger, a series of experiments are conducted using Nepali corpus containing 15430 words. The experiments were performed with conducted further tests on more interesting dataset to evaluate the real performance of this approach. Accuracy about 93.15.6% represents a very good result of our method compared to Rule-Based. We note that the accuracy slightly increased with the increasing of the number of words in the training corpus. In the future, we plan to improve the tagging accuracy of unknown words by using other training corpus, and applying our POS tagger in extraction of Multi-Word Terms.

References

[1] SandipanDandapat(2009): Part-of-Speech Tagging for Bengali.

[2] en.wikipedia.org/wiki/Nepali_language.

[3] Nisheeth Joshi , Hemant Darbari and Iti Mathur: HMM based pos tagger for hindi

[4] Fahim Muhammad Hasan:Comparison of different pos tagging techniques for some south asian languages.

[5] en.wikipedia.org/wiki/Hidden_Markov_model.

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

[7] Yoga Raj Joshi May, 2010: A Chunk Alignment Model for Statistical Machine Translation on English-Nepali Parallel Corpus.

[8] Bal Krishna Bal,Madan PuraskarPustakalaya, Nepal POS Tagset Design, Tagging

[9] Eric Brill.(1992): A simple rule-based part of speech tagger. In Proceedings Third Conference on Applied.

[10] Jyoti Singh, Nisheeth Joshi , Iti Mathur Development of Marathi Part of Speech Tagger Using Statistical Approach.

[11] Status and Challenges Tagset and Tagging Nepali Corpus

[12] Lowrance R. Rabiner: A tutorial on Hidden Morkov Models and Selected Applications in Speech Recognition.

[13] Kanak Mohnot, Neha Bansal, Shashi Pal Singh, Ajai Kumar : Hybrid approach for Part of Speech Tagger for Hindi language

AUTHOR'S PROFILE

Mr. Prajadhip Sinha is working as Assistant Professor in Kohima Science College. His research interests include Natural Language Processing (NLP), E-learning, Corpus Based Learning and Computer Applications etc.

Miss Nuzotalu M Veyie is working as Assistant Professor in Kohima Science College. She is specialized in Natural Language Processing (NLP), Machine assisted Translation (MT), Computing and Mobile Computing and C language.

Prof. Bipul Syam Purkayastha, is working as Head of the Department, Computer Science in Assam University, Silchar, India. He is handling various projects in the area of Natural Language Processing, Information Extraction and Retrieval. He has published various national & international papers.