A Comparative Study of Part-of-Speech Tagging Techniques of Nepali language.

Dr Prajadhip Sinha [1]
[1] *Assistant Professor, Dept of Computer Science*
Jotsoma, Kohima, Nagaland
(E-mail: prajadhip@rediffmail.com)

***Abstract*—** Part-of-speech tagging is the process of classifying or labeling the words in a text with their appropriate part of speech. Different POS tagging techniques in the literature have been developed and experimented mostly for English language. Some of the same work has been done for Nepali language. Comparative studies on POS tagging for Nepali language are relatively unexplored. There are many automatic POS taggers which have been developed worldwide using linguistic rules, stochastic models and hybrid models etc. Different types of taggers have their own advantages as well as disadvantages. In this paper we compare the performance of some POS tagging with different techniques of Nepali language and tried to see which technique maximizes the performance with our case.

Keywords*—*POS, stochastic, taggers, Nepali, hybrid model

I. INTRODUCTION

POS tagging is a technique to assign with its appropriate lexical categories. The process takes a word or a sentence as input, assigns a POS tag to the word or to each word in the sentence and produces the tagged text as output. Part of speech tagging is a preliminary and important component of computational linguistics or natural language processing. Human languages are generally known as natural languages; the science of studying natural languages falls under the area of linguistics and its implementation using computational means is regarded as computational linguistics.

Nepali language is originally belongs to the Indo-Aryan branch of indo-European family. This language takes its root from Sanskrit which is the classical language of India. Nepali language was known as Gurkha, Gurkhali or Khas Kura. In the 11th century AD Nepali Language developed from the Brahmi script. Nepali Language is written with the Devanagari alphabet. Nepali is spoken by more than 40 million people, mostly in Nepal, Bhutan, Myanmar, West Bengal and other parts of India. Linguistically, Nepali is most closely related to Sanskrit and Hindi. A large proportion of the technical vocabulary written in Nepali is influenced by Sanskrit.
Nepali is written in the Devanagari script and there are 12 vowels and 36 consonants in this language. The script is written from left to right. There is no provision of capital and small letters in the script. The Nepali alphabets are written in two separate groups, namely the vowels and the consonants

II. RELATED WORK

The Nepali National Corpus (NNC) from NELRALEC (Nepali Language Resources and Localization for Education and Communication) project, which contain 14 million Nepali words. It consists of speech corpus, spoken corpus, core sample (CS), general collection, and parallel data.
The Unitag1 has been developed or customized for Nepali language and was used for semi automatic tagging of Nepali National Corpus under the NERLAC project and tagset used is NERLAC project with 112 tags. Originally, Unitag was developed for Urdu language by Hardie etal. It consists of lexical analysis, a powerful morphological system and twin disambiguation modules, hand-written rules and the other using a probabilistic system based on a Hidden Markov model. After tagging, the corpus was manually reviewed and then correction was done. Since the tagset used was very large, it showed more error in tagging. Later the TnT tagger has been used as POS tagger with the 43 tags and training corpus of medium size as one of the pipelined modules for computational grammar analyzer.

III. THE PROPOSED APPROACH

There are different models for part of speech tagging. It can be classified as Supervised and Unsupervised. Both the supervised and unsupervised models can be classified as rule-based and stochastic model. In supervised POS tagging models a pre-tagged corpora is required, which is used for training to learn about the tagset, tag frequencies word, frequencies, rule sets etc. The accuracy or performance of these models generally increases when we increase size of the corpora.

Rule Based Approach: The basic principle of rule based approaches is that, the knowledge base consists of a set of linguistic generalizations, known most commonly as rules or constraints. Each rule contains the instructions for an operation to be performed, and the context describing where the rule should be applied. And these rules are responsible to provide the appropriate tags to the text. Typical rule based approaches use contextual information to assign tags to ambiguous words or unknown. These rules are often known as context frame rules. As an

example, "if any word is preceded by a determiner and followed by a noun, then it is tagged as an adjective".

Hidden Markov model: Markov models the state is directly visible to the observer, so the state transition probabilities are the only parameters. But in Hidden Markov model the state is hidden to the observer and output produced with the help of those states which is visible to the observer. Each state has a probability distribution over the possible output tokens or words. Therefore, the sequence of tokens generated by a Hidden Markov Model gives some information about the sequence of states.

Maximum Entropy Model: Maximum Entropy (ME) is a very flexible method of statistical modelling. In machine learning, a maximum-entropy Markov model (MEMM), or conditional Markov model (CMM), is a graphical model for sequence labelling that combines features of hidden Markov models (HMMs) and maximum entropy (MaxEnt) models. An MEMM is a discriminative model that extends a standard maximum entropy classifier by assuming that the unknown values to be learnt are connected in a Markov chain rather than being conditionally independent of each other. MEMMs find applications in natural language processing, specifically in part-of-speech tagging and information extraction.

Memory Based Learning: The Memory Based Learning (MBL) Model takes tagged data as input, and produces a lexicon and memory based POS tags as output.MBL consists of two components, one is a memory based learning component, and the other is a similarity based performance component. The learning component is called memory based as it memorizes examples while training. The performance component matches the similarity of the input with the output of the learning component to produce the actual output of the system. The different models described above have their own advantages and disadvantages, however, they all face one difficulty, which is to assign a tag to an unknown word which the tagger has not seen previously i.e. the word was not present in the training corpora.

Hybrid Approach: Our proposed method is based on hybrid approach; it combines the Rule-Based method presented with HMM probabilistic Techniques and makes new methods using strongest points from each method. It is makes use of essential feature from ML approaches and uses the rules to make it more efficient. Hybrid methods are ideally to be used to increase the accuracy of the system

Nepali tagset

For designing a Nepali tagset, apart from following the Eagles Guidelines and the Penn tree bank tagset, many other Indian tagging guidelines like IL-POST, ILMT and Sanskrit tagset were taken into consideration. After careful consideration a hierarchical tagset was favoured, the whole design of the tagset developed so far revolves around three distinct features into which the grammatical schema is distributed. The features are Category, Type and Attribute.

The tagset for Nepali currently includes 43 tags and covers almost all the grammatical categories in the Nepali language. By the reference of Penn Treebank [61] tagset, the tagset of the Nepali is designed and it also based on BIS (Bureau of Indian Standards) framework. The short description of tag set used here is given follow in table I:

POS Name	Tag	POS Name	Tag
Common Noun	NN	Coordinating	CC
Proper Noun	NNP	Subordinating Conjunction	CS
Personal Pronoun	PP	Interjection	UH
Possessive Pronoun	PP$	Cardinal Number	CD
Reflexive Pronoun	PPR	Ordinal Number	OD
Marked Demonstrative	DM	Plural Marker	HRU
Unmarked Demonstrative	DUM	Question Word	QW
Finite Verb	VBF	Classifier	CL
Auxiliary Verb	VBX	Particle	RP
Verb Infinitive	VBI	Determiner	DT
Prospective Participle	VBNE	Unknown Word	UNW
Aspectual Participle	VBKO	Foreign Word	FW
Other Participle Verb	VBO	sentence Final	YF
Normal/Unmarked	JJ	sentence Medieval	YM
Marked Adjective	JJM	Quotation	YQ
Degree Adjective	JJD	Brackets	YB
Manner Adverb	RBM	Header List	ALPH
Other Adverb	RBO	Symbol	SYM
Intensifier	INTF	Abbreviation	FB
Le-Postposition	PLE		
Lai-Postposition	PLAI		
Ko-Postposition	PKO		
Other Postpositions	POP		

Table I: Nepali Tagset

Experimental Result

Experiment was done in four phases according to the size of the lexicon. The four phases were based on lexicon sizes: 5000, 10000, 15000, and 20000. Lexicons were compiled based on different domains viz., government/politics, sports, tourism, etc.

Test data were randomly selected from different domains. Same test data were manually tagged in order to compare the accuracies of tagger. An application was built, which takes automatically tagged test data and manually tagged data as input. In order to see the percentage of error over test corpus, tag of a word in test corpus was compared against the tag of manually tagged corpus.

We have taken four different test sets with similar corpus sizes, and the tagging results obtained for each corpus are given below:

Experiment-1					
Lexicon Size (words)	**Exp**	**Total words**	**Rule Based Approach**	**Stochastic Approach**	**Hybrid Approach**
5000	1	1000	478	496	506
5000	2	1500	722	735	755
5000	3	2000	895	919	997
5000	4	2500	1190	1212	1269
5000	5	3000	1456	1479	1538
5000	6	3500	1743	1777	1829
5000	7	4000	2101	2140	2210
5000	8	4500	2345	2386	2512
5000	9	5000	2720	2745	2903

Table II: Experiment Set 1 based on lexicon size: 5000 words

Experiment-2					
Lexicon Size (words)	**Exp**	**Total words**	**Rule Based Approach**	**Stochastic Approach**	**Hybrid Approach**
10000	1	1000	602	624	682
10000	2	1500	910	942	1036
10000	3	2000	1215	1250	1420
10000	4	2500	1545	1562	1708
10000	5	3000	1830	1890	2078
10000	6	3500	2179	2257	2445
10000	7	4000	2489	2560	2734
10000	8	4500	2835	2965	3151
10000	9	5000	3199	3289	3523

Table III: Experiment Set 2 based on lexicon size: 10000 words

Experiment-3					
Lexicon Size (words)	**Exp**	**Total words**	**Rule Based Approach**	**Stochastic Approach**	**Hybrid Approach**
15000	1	1000	692	702	845
15000	2	1500	1050	1086	1292
15000	3	2000	1436	1444	1688
15000	4	2500	1779	1982	2170
15000	5	3000	2130	2245	2603
15000	6	3500	2487	2598	3075
15000	7	4000	2840	2971	3594
15000	8	4500	3210	3401	3921
15000	9	5000	3654	3754	4450

Table IV: Experiment Set 3 based on lexicon size:15000 words

Experiment-4					
Lexicon Size (words)	**Exp**	**Total words**	**Rule Based Approach**	**Stochastic Approach**	**Hybrid Approach**
20000	1	1000	791	832	922
20000	2	1500	1201	1250	1392
20000	3	2000	1625	1701	1837
20000	4	2500	2024	2109	2301
20000	5	3000	2439	2530	2803
20000	6	3500	2841	2971	3275
20000	7	4000	3242	3387	3790
20000	8	4500	3689	3829	4211
20000	9	5000	4112	4301	4706

Table V: Experiment Set 4 based on lexicon size: 20000 words

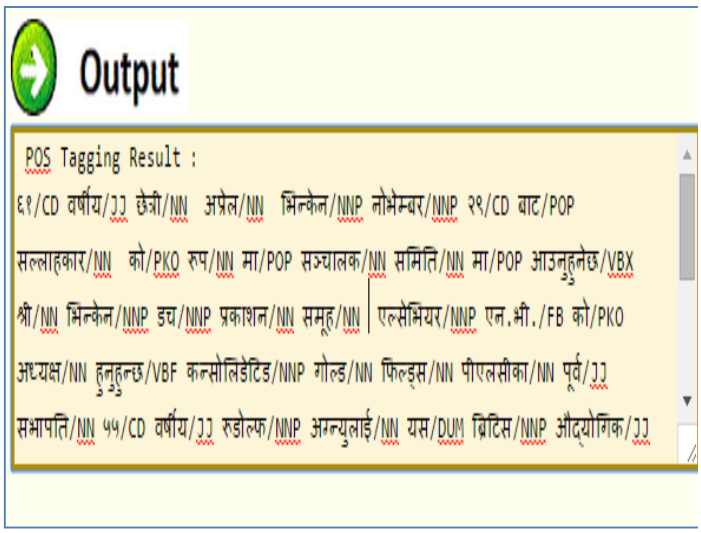
Input/Output of the Tagger

Input: ६१ वर्षीय छेत्री अप्रेल भिन्केन नोभेम्बर २९ बाट सल्लाहकार को रूप मा सञ्चालक समिति मा आउनुहुनेछ ।श्री भिन्केन डच प्रकाशन समूह एल्सेभियर एन.भी .को अध्यक्ष हुनुहुन्छ ।

कन्सोलिडेटिड गोल्ड फिल्ड्स पीएलसीका पूर्व सभापति ५५ वर्षीय रूडोल्फ अग्युलाई यस ब्रिटिस औद्योगिक समूहको सल्लाहकारको रूपमा मनोनयन गरिएको थियो ।एकताका केन्ट चुरोटको फिल्टर बनाउन प्रयोग भएको एक प्रकारको अस्बेस्टोस तीस वर्षभन्दा अगाडि यसको सम्पर्कमा आएका कामदारहरूको समूहमा क्यान्सरबाट मृत्यु हुनेको उच्च प्रतिशतको कारण बनेको छ , अनुसन्धाताहरूले जानकारी दिए ।

Output:

६१/CD वर्षीय/JJ छेत्री/NN अप्रेल/NN भिन्केन/NNP नोभेम्बर/NNP २९/CD बाट/POP सल्लाहकार/NN को/PKO रूप/NN मा/POP सञ्चालक/NN समिति/NN मा/POP आउनुहुनेछ/VBX श्री/NN भिन्केन/NNP डच/NNP प्रकाशन/NN समूह/NN एल्सेभियर/NNP एन.भी/.FB को/PKO अध्यक्ष/NN हुनुहुन्छ/VBF कन्सोलिडेटिड/NNP गोल्ड/NN फिल्ड्स/NN पीएलसीका/NN पूर्व/JJ सभापति/NN ५५/CD वर्षीय/JJ रूडोल्फ/NNP अग्युलाई/NN यस/DUM ब्रिटिस/NNP औद्योगिक/JJ समूहको/NN सल्लाहकारको/NN रूपमा/NN मनोनयन/NN गरिएको/VBKO थियो/VBX एकताका/RBO केन्ट/NNP चुरोटको/NN फिल्टर/NN बनाउन/VBI प्रयोग/NN भएको/VBKO एक/CD प्रकारको/NN अस्बेस्टोस/NNP तीस/CD वर्षभन्दा/NN अगाडि/RBO यसको/NN सम्पर्कमा/NN आएका/VBKO कामदारहरूको/NN समूहमा/NN क्यान्सरबाट/NN मृत्यु/NN हुनेको/VBKO उच्च/दर प्रतिशतको/NN कारण/NN बनेको/VBKO छ/VBF ,/YM अनुसन्धाताहरूले/NN जानकारी/NN दिए/VBF



Performance of POS Taggers:

We have tested many experiments using rule based approach and HMM with rule based approach on different corpus till we get the best accuracy. Then, we have seen that POS Tag using HMM with rule based approach get the better accuracy than using rule based approach only. Table II shows the performances of POS tagging according to the different approaches on different number of words in corpus. Figure 2 also, shows the comparison of these improvements in accuracy along with the increase in the size of annotated training data on different methods.

Number of words	Rule Based Approach	HMM Approach	Hybrid Approach
5000	47.11	47.87	52.66
7500	61.34	61.23	65.43
10000	61.54	61.23	69.33
12500	75.42	76.67	82.13
15000	77.87	79.90	88.11
17500	80.56	83.54	90.33
20000	84.89	87.09	93.50

Table VI: Performance of different number of words

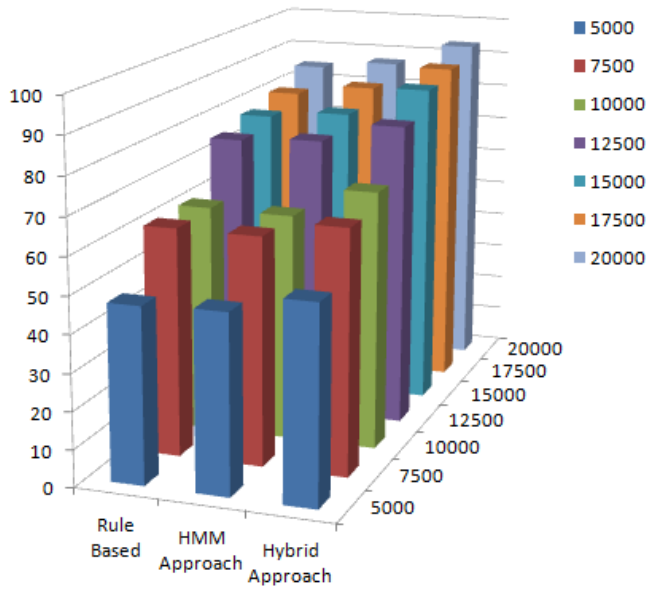


Fig I: Accuracy of various POS taggers on Nepali Text

VII CONCLUSION

We have compared the performance of Rule based, HMM and Hybrid method on Nepali language and found that hybrid taggers performed better for Nepali Language than other two methods. A hybrid solution for POS tagging in Nepali can be proposed that can be used in other advanced NLP applications, which might use a combination of the techniques mentioned

earlier to achieve a significant gain in performance and performs with very good accuracy as English or other languages in all domains

At present with the training corpus with a size of around 20000 words of a domain we get a performance of over 90%.If we can increase the training corpus size covering most of the domains then we might get a recognizable performance of 95%+ for Nepali too.

REFERENCES

[1] Arabic language - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Arabic_language

[2] F. Al Shamsi, and A. Guessoum. A Hidden Markov Model-Based POS Tagger for Arabic. JADT 2006.

[3] Abdelkareem M. Alashqar: A Comparative Study on Arabic POS Tagging Using Quran Corpus.

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

[6] Bal Krishna Bal Madan Puraskar Pustakalaya, Nepal Structure of Nepali Grammar.

[7] Mathew, D. A Course in Nepali, Ratna Pustak Bhandar, 1998

[8] en.wikipedia.org/wiki/Gurkha

[9] Bal Krishna Bal: Structure of Nepali Grammar.

[10]Prajadhip Sinha, Bhairab Sarma and Dr. Bipul Shyam Purkayastha." Kinship Terms in Nepali Language and its Morphology :International Journal of Computer Applications (0975 – 8887) Volume 58– No.9, November 2012.

[11] http://www.bhashasanchar.org/index.php

[12]Prajadhip Sinha, Bhairab Sarma and Dr. Bipul Shyam Purkayastha :"A Combined Approach to Part-of-Speech Tagging Using Features Extraction and Hidden Markov Model. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 2, February 2013. ISSN: 2278 – 1323.

[13]Prajadhip Sinha, Nuzotalu M Veyie and Bipul Syam Purkayastha "Enhancing the Performance of Part of Speech tagging of Nepali language through Hybrid approach" : International Journal of Emerging Technology & Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal), Volume 5, Issue 5, May, 2015.

[14] http://www.bhashasanchar.org/index.php

[15] Bal Krishna Bal: A Morphological Analyzer and a Stemmer for Nepali.

About Author:

Dr. Prajadhip Sinha is working as Assistant Professor in Kohima Science College and has 5 years of research experience in NLP based works. His research interests include Corpus Linguistics, Pats-of-Speech tagging, Morphological analyzer & generator, Speech data segmentation and annotation.