HMM BASED POS TAGGER FOR HINDI

Nisheeth Joshi[1], Hemant Darbari[2] and Iti Mathur[3]

[1,3]Department of Computer Science, Banasthali University, India
[1]nisheeth.joshi@rediffmail.com
[3]mathur_iti@rediffmail.com
[2]Center for Development of Advanced Computing, Pune, Maharashtra, India
[2]darbari@cdac.in

ABSTRACT

Part of Speech tagging in Indian Languages is still an open problem. We still lack a clear approach in implementing a POS tagger for Indian Languages. In this paper we describe our efforts to build a Hidden Markov Model based Part of Speech Tagger. We have used IL POS tag set for the development of this tagger. We have achieved the accuracy of 92%.

KEYWORDS

Hidden Markov Model, POS Tagging, Hindi, IL POS Tag set

1. INTRODUCTION

Part of Speech (POS) Tagging is the first step in the development of any NLP Application. It is a task which assigns POS labels to words supplied in the text. This is the reason why researchers consider this as a sequence labeling task where words are considered as sequences which needs to be labeled. Each word's tag is identified within a context using the previous word/tag combination. POS tagging is used in various applications like parsing where word and their tags are transformed into chunks which can be combined to generate the complete parse of a text.
Taggers are used in Machine Translation (MT) while developing a transfer based MT Engine. Here, we require the text in the source language to be POS tagged and then parsed which can then be transferred to the target side using transfer grammar. Taggers can also be used in Name Entity Recognition (NER) where a word tagged as a noun (either proper or common noun) is further classified as a name of a person, organization, location, time, date etc.

Tagging of text is a complex task as many times we get words which have different tag categories as they are used in different context. This phenomenon is termed as lexical ambiguity. For example, let us consider text in Table 1. The same word 'सोने' is given a different label in the two sentences. In the first case it is termed as a common noun as it is referring to an object (Gold Ornament). In the second case it is termed as a verb as it is referring to an experience (feelings) of the speaker. This problem can be resolved by looking at the word/tag combinations of the surrounding words with respect to the ambiguous word (the word which has multiple tags).

Over the years, a lot of research has been done on POS tagging. Broadly, all the efforts can be categorized in three directions. They are: rule based approach where a human annotator is required to develop rules for tagging words or statistical approach where we use mathematical

Jan Zizka (Eds) : CCSIT, SIPP, AISC, PDCTA - 2013
pp. 341–349, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3639

formulations and tag words or hybrid approach which is partially rule based and partially statistical. In the context of European languages POS taggers are generally developed using machine learning approach, but in the Indian context, we still do not have a clear good approach. In this paper we discuss the development of a POS tagger for Hindi using Hidden Markov Model (HMM).

सोने	के	आभूषण	महंगे	हो	गए	है	
NN	PSP	NN	JJ	VM	VAUX	VAUX	
उसका	दिल	**सोने**	का	है			
PRP	NN	VM	PSP	VM			

Table 1. Example of Lexical Ambiguity

The paper is organized with literature survey in Section 2 which is succeeded by section 3 which describes the HMM approach for POS tagging Hindi text and section 4 which shows evaluation results followed by section 5 which concludes the paper.

2. 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.[1] 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. [2] 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. [3] used a pure maximum entropy based machine learning approach for labeling Hindi words with various POS tag categories. This tagger reported to have 88.4% accuracy Shrivastava and Bhattacharya [4] 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 [5] and Avinesh and Gali [6] 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%.

Considerable Effort of developing a POS Tagger in other Indian Languages have also been put in for Malayalam, an HMM based tagger was proposed by Manju et. al. [7], since they did not had an annotated corpus, they used a morphological analyzer to generate the corpus which was then used for training the HMM algorithm. Another tagger for Malayalam was developed by Anthony et. al. [7] who used Support Vector Machines (SVM). They used a SVMTool for tagging which was developed by Giménez and Màrquez [8]. For developing this tagger Anthony et. al. first proposed a tagset which they claim is suitable for Malayalam and then created an annotated corpus using this tagset. Their tagger reported 94% accuracy with their tagset.

For Bengali, Dandapat et. al.[9] studied the possibility of developing a tagger using HMM and Maximum Entropy (ME) models. They too used a morphological analyzer for compensating the shortage of annotated corpus. With these two modes they implemented a supervised tagger and a semi-supervised tagger and reported an accuracy of around 88% for the two approaches. Ekbal and Bandyopadhyay[10] annotated news corpus and developed an SVM based tagger. They reported an accuracy of 86.84% for their tagger. Ekbal et. al. [11] also developed a Conditional

Random Fields(CRF) based tagger. For training the tagger they used the information of prefix and suffix of Bengali words along with normal word/tags and reported an accuracy of 90.3%. For Tamil, Selvam and Natarajan[12] proposed a POS tagger which used a rule based morphological analyzer to annotate the corpora which was used to train the tagger. They used the Tamil version of the Bible for annotation of POS tagged corpus and reported an accuracy of 85.56%. Dhanalakshmi et. al.[13] proposed an SVM based tagger using linear programming and developed their own POS tagset for Tamil which has 32 tags. They used this tagset to annotate their corpus and then trained their model and reported an accuracy of 95.63%. Dhanalakshmi et. al.[14] also proposed another tagger where they used machine learning techniques to extract linguistic information which was then used to train the tagger based on SVM approach. They used their own 32 tags tagset for annotating the corpus and reported an accuracy of 95.64%. For Marathi, Singh et al [15] proposed a POS tagger using trigram method. They used a pos tagset proposed by Bharti et al [16] which had 24 tags. They showed an accuracy of 91.63%.

3. HIDDEN MARKOV MODEL BASED POS TAGGER FOR HINDI

For developing a HMM based tagger we were first required to annotate a corpus based on a tagset. We used IL POS tagset[16] proposed by Bharti et. al. Table 2 shows brief description of the tags used. A detailed explanation can be sought from their paper.

We used 15,200 sentences (3,58,288 words) from tourism domain to train our system. Since this is a sizable corpus, we did not put in our efforts in developing morphological analyzers. Instead, we used 5 annotators for creation of the POS tagged corpora who completed the task of annotating this corpus in four months. The working of the tagger is as follows:

3.1 Working of Tagger

A POS tagger based on HMM assigns the best tag to a word by calculating the forward and backward probabilities of tags along with the sequence provided as an input. The following equation explains this phenomenon.

$$P(t_i|w_i) = P(t_i|t_{i-1}).P(t_{i+1}|t_i).P(w_i|t_i) \quad (1)$$

Here, $P(t_i|t_{i-1})$ is the probability of a current tag given the previous tag and $P(t_{i+1}|t_i)$ is the probability of the future tag given the current tag. This captures the transition between the tags. These probabilities are computed using equation 2.

$$P(t_i|t_{i-1}) = \frac{freq\ (t_{i-1}, t_i)}{freq(t_{i-1})} \quad (2)$$

Each tag transition probability is computed by calculating the frequency count of two tags seen together in the corpus divided by the frequency count of the previous tag seen independently in the corpus. This is done because we know that it is more likely for some tags to precede the other tags. For example, an adjective (JJ) will be followed by a common noun (NN) and not by a postposition (PSP) or a pronoun (PRP). Figure 1 shows this example.

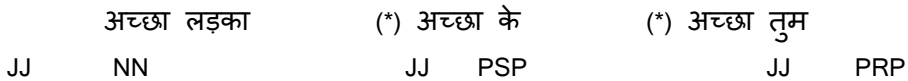

Figure 1. Tag transition probabilities

By looking at this figure, we know that probability of P(JJ|NN) will fetch a high score then P(JJ|PSP) and P(JJ|PRP). Since the last two are wrong, we might not get even a single count for them.

S.No.	Tag	Description (Tag Used for)	Example
1.	NN	Common Nouns	लड़का, लड़के, किताब, पुस्तक
2.	NST	Nourn Denotating Spatial and Temporal Expressions	ऊपर, पहले, बहार, आगे
3.	NNP	Proper Nourns (name of person)	मोहन, राम, सुरेश
4.	PRP	Pronoun	वह, वो, उसे, तुम
5.	DEM	Demonstrative	वह, वो, उस
6.	VM	Verb Main (Finite or Non-Finite)	खाता, सोता, रोता, खाते, सोते, रोते
7.	VAUX	Verb Auxilary (Any verb, present besides main verb shall be marked as auxillary verb)	है, हुए, कर
8.	JJ	Adjective (Modifier of Noun)	सांस्कृतिक, पुरानी, दुपहिया
9.	RB	Adverb (Modifier of Verb)	जल्दी, धीरे, धीमे
10.	PSP	Postposition	में, को, ने
11.	RP	Particles	भी, तो, ही
12.	QF	Quantifiers	बहुत, थोडा, कम
13.	QC	Cardinals	एक, दो, तीन
14.	CC	Conjuncts (Coordinating and Subordinating)	और, की
15.	WQ	Question Words	क्यों, क्या, कहा
16.	QO	Ordinals	पहला, दूसरा, तीसरा
17.	INTF	Intensifier	बहुत, थोडा, कम
18.	INJ	Interjection	अरे, हाय
19.	NEG	Negative	नहीं, ना
20.	SYM	Symbol	? , ; : !
21.	XC	Compounds	केंद्र/XC सरकार/NN रंग/XC बिरंगे/JJ
22.	RDP	Reduplications	धीरे/RB धीरे/RDP गली/NN गली/RDP

S.No.	Tag	Description (Tag Used for)	Example
23.	ECH	Echo Words	चाय-व्हाय, प्यार-व्यार
24.	UNK	Forigen Words	English, বাংলা, ગુજરાતી

Table 2. Description of IL POS Tagset

We also computed the word likelihood probabilities using $P(w_i|t_i)$ i.e. the probability of the word given a current tag. This probability is computed using equation 3.

$$P(w_i|t_i) = \frac{freq\ (t_i, w_i)}{freq(t_i)} \quad (3)$$

Here, the probability of word provided a tag is computed by calculating the frequency count of the tag in question and the word occurring together in the corpus divided by the frequency count of the occurrence of the tag alone in the corpus.

We used two special tags <S> to denote the starting of the sentence and </S> denoting the ending of the sentence which was added to all the sentences of the training corpus. Using the above two equations we created a tag-tag database which computed all the tag transition probabilities of tag combinations available in the corpus and a word-tag database which computed all word likelihood probabilities available in the corpus.

Suppose if we have a word which is an open class word i.e. a noun or verb or adjective or adverb. Then it is a possibility that it might be assigned to multiple tags and we may face the ambiguity issue. For example, in table 1 we have an ambiguous word which is assigned to a noun and a verb. A human expert can very easily distinguish the two contexts and thus assign a different POS tag to the words. Using HMM this phenomenon can be captured intuitively as we are considering the context of tags (before and after) with respect to the current tag. This context description is a powerful feature of HMM which can decides the tag for a word by looking at the tag of the previous word and the tag of the future word. Figure 2 shows this phenomenon which is a generative model where there is a hidden underlying generator of observable events (tag-tag probabilities) and this hidden generator can be modelled as a set of states. Our goal is to find the underlying state sequence from the observed events.

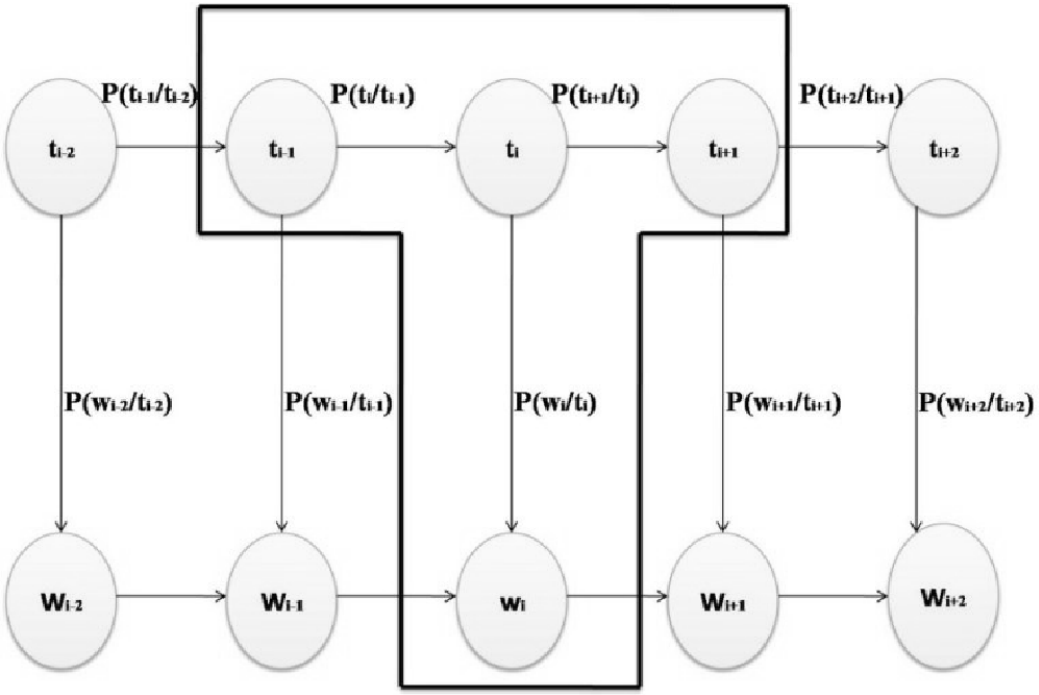

Figure 2. Context Dependency of Hidden Markov Model

Let us consider this computation using the example "उसका दिल सोने का है". Here, "सोने" is an ambiguous word which can either have an NN or a VM tag. Figure 3 and 4 show the context dependency of the possible tags.

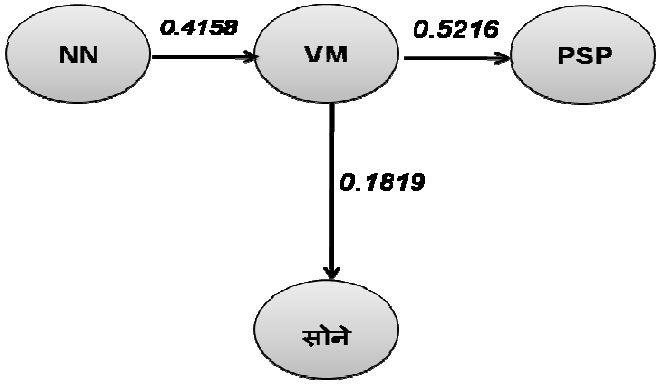

Figure 2. Context Dependency of "उसका दिल सोने का है" with VM assigned to "सोने"

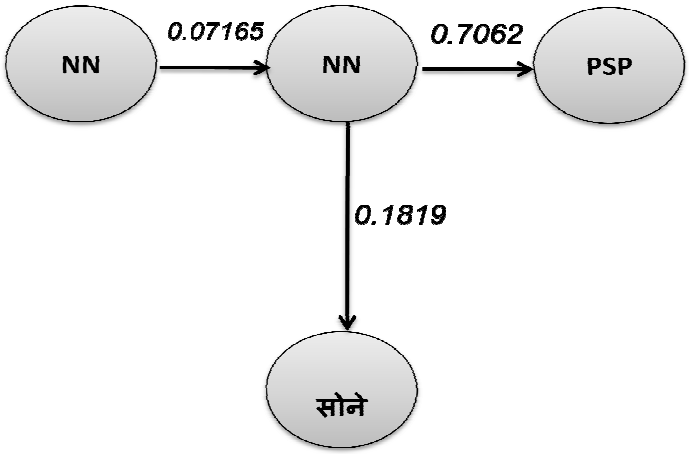

Figure3. Context Dependency of "उसका दिल सोने का है" with NN assigned to "सोने"

Since all the other word-tag combinations are same for the ambiguous words, their computation will also be the same. It is the ambiguous word-tag context which will make the difference to the final score. The one which is higher will get selected. So, the computation of P(VM|NN) × P(PSP|VM) × P(सोने|VM) is computed to be 0.03945 and the computation of P(NN|NN) × P(PSP|NN) × P(सोने|NN) is computed to be 0.0092. As the computation of VM is higher than that of NN, VM gets selected for "सोने". Once all the tags are identified for the input words. They are displayed to the user.

4. EVALUATION

For testing the performance of our system, we developed a test corpus of 500 sentences (11,720 words). We calculated precision, recall and f-measure as they are considered to be standard performance indicators of a system. These were calculated using the following equations.

$$Precision\ (P) = \frac{No.of\ correct\ POS\ tags\ assigned\ by\ the\ system}{No.of\ POS\ tags\ assigned\ by\ the\ system} \quad (4)$$

$$Recall\ (R) = \frac{No.of\ correct\ POS\ tags\ assigned\ by\ the\ system}{No.of\ POS\ tags\ in\ the\ text} \quad (5)$$

$$F - Measure = \frac{2\ \times P\ \times R}{P + R} \quad (6)$$

Recall is also considered to be the accuracy score of the system as it assigns correct tags against the total no. of tags in the corpus. Test scores of our system are as follows:

No. of Correct POS tags assigned by the system = 10798
No. of POS tags assigned by the system = 11720
No. of POS tags in the text = 11720
Since the score of precision and recall are same the f-measure would also be same. Thus the accuracy of the system is 92.13%.

5. CONCLUSION

We used HMM based statistical technique to train our POS tagger for Hindi. We disambiguated correct word-tag combinations using the contextual information available in the text. We attained the accuracy of 92.13% on test data. Future enhancements of this work would be to improve the accuracy of the tagger. This can be achieved by improving the tagset and adding more tags so that the tagger can make less ambiguous classification of the text.

We can also use this tagger to generate some possible 8-10 tags for a word which can be transformed into elementary tress and can generate super tags with derivation trees (α and β trees), this will help us in training a super tagger (calculating the probabilities of α and β trees) which can be used to implement a fully functional Tree Adjoining Grammar (TAG) based parser for Hindi.

REFERENCES

[1] Bharati, A., Chaitanya V., Sangal R., (1995) "Natural Language Processing – A Paninian Perspective". Prentice-Hall India, New Delhi (1995).
[2] Singh, S., Gupta, K., Shrivastava, M., Bhattacharya, P., (2006) "Morphological Richness Offsets Resource Demand- Experiences in Constructing a POS Tagger for Hindi". In: COLING/ACL, pp. 779-786.
[3] Dalal, A., Nagaraj, K., Sawant, U., Shelke, S., (2006) "Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach". In: NLPAI Machine Learning Competition.
[4] Shrivastava, M., Bhattacharyya, P., (2008) "Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information Without Extensive Linguistic Knowledge". In: International Conference on NLP (ICON08), Macmillan Press, New Delhi.
[5] Agarwal, H., Amni, A., (2006) "Part of Speech Tagging and Chunking with Conditional Random Fields". In: NLPAI Machine Learning Competition.
[6] Avinesh, PVS, Karthik, G., (2006) "Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning". In: NLPAI Machine Learning Competition.
[7] Manju K., Soumya S., Sumam, M. I., (2009) "Development of a POS Tagger for Malayalam - An Experience". In: International Conference on Advances in Recent Technologies in Communication and Computing, pp.709-713.
[8] Jeśus Giménez and Llúis Màrquez., (2006) "SVMTool. Technical manual v1.3".
[9] Dandapat, S., Sarkar, S., Basu, A., (2007) "Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario". In: Association for Computational Linguistic, pp 221-224.
[10] Ekbal, A., Bandyopadhyay, S., (2007) "Lexicon Development and POS tagging using a Tagged Bengali News Corpus". In: FLAIRS-2007, Florida, pp 261-263.
[11] Ekbal, A., Haque, R., Bandyopadhyay, S., (2007) "Bengali Part of Speech Tagging using Conditional Random Field". In: 7th International Symposium of Natural Language Processing(SNLP-2007), Thailand Pattaya, 13-15 December 2007, pp.131-136.
[12] Selvam, M., Natarajan, A.M., (2009) "Improvement of Rule Based Morphological Analysis and POS Tagging in Tamil Language via Projection and Induction Techniques". International Journal of Computers, 3(4).
[13] Dhanalakshmi, V., Kumar, A., Shivapratap, G, Soman, K.P., Rajendran, S, (2009) "Tamil POS Tagging using Linear Programming". International Journal of Recent Trends in Engineering, 1(2).
[14] Dhanalakshmi V, Anand kumar M, Rajendran S, Soman K P., (2009) "POS Tagger and Chunker for Tamil Language". Proceedings of Tamil Internet Conference 2009.
[15] Singh, J., Joshi, N., Mathur I., (2013) "Part of Speech Tagging of Marathi Text Using Trigram Method", International Journal of Advanced Information Technology, pp 35-41, Vol 3. No. 2.
[16] Bharati, A., Sharma, D.M., Bai, L., Sangal, R., (2006) "AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages", http://ltrc.iiit.ac.in/tr031/posguidelines.pdf

AUTHORS

Mr. Nisheeth Joshi is a researcher working in the area of Machine Translation. He has been primarily working in design and development of evaluation metrics in Indian Languages. Besides this he is also very actively involved in the development of MT engines for English to Indian Languages. He is one of the experts empanelled with TDIL Programme, Department of Information Technology, Govt. of India, a premier organization which foresees Language Technology Funding and Research in India. He has several publications in various journals and conferences and also serves on the Programme Committees and Editorial Boards of several conferences and journals

Dr. Hemant Darbari, Executive Director of the Centre for Development of Advanced Computing (CDAC) specialises in Artificial Intelligence System. He has opened new avenues through his extensive research on major R&D projects in Natural Language Processing (NLP), Machine assisted Translation (MT), Information Extraction and Information Retrieval (IE/IR), Speech Technology, Mobile computing, Decision Support System and Simulations. He is a recipient of the prestigious "Computerworld Smithsonian Award Medal" from the Smithsonian Institution, USA for his outstanding work on MANTRA-Machine Assisted Translation Tool which is also a part of "The 1999 Innovation Collection" at National Museum of American History, Washington DC, USA.

Mrs. Iti Mathur is an Assistant Professor at Banasthali University. Her primary area of research is Computational Semantics and Ontological Engineering. Besides this she is also involved in the development of MT engines for English to Indian Languages. She is one of the experts empanelled with TDIL Programme, Department of Electronics and Information Technology (DeitY), Govt. of India, a premier organization which foresees Language Technology Funding and Research in India. She has several publications in various journals and conferences and also serves on the Programme Committees and Editorial Boards of several conferences and journals.