



Reduced Tagset To Improve Accuracy Of HMM Based Parts Of Speech Tagger In Telugu Language

Dr. V. Suresh

Assistant Professor, Department of IT, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam-531162, Andhra Pradesh, India.

Corresponding Author: Dr. V. Suresh

Abstract - The accuracy of a tagger is the biggest challenge today. A lot of tagger have been proposed by different Researchers for the different languages (Telugu, Tamil, Hindi, Bengali etc.) using different techniques like HMM (Hidden Markov Model), SVM (Support Vector Machine), ME (Maximum Entropy) etc. POS tagging is the fundamental task in almost all the NLP applications like Grammar Checking, Speech processing, Machine translation etc. that assign the correct tag to the word for a number of available tags. A Telugu POS tagger based on HMM model is one of them. This tagger uses Hidden Markov Model., a statistical technique to accurately tag the words in Telugu language using 620 tags. This large tag set results in data sparseness problem. To cope up with this problem, in this research paper an experiment with reduced POS tag set (34 tags) proposed by Technical Development of Indian Languages (TDIL) has been used to improve the tagging accuracy of HMM based POS tagger.

Keywords-NLP, Parts-of-speech tagger, Telugu, HMM technique, TDIL proposed Telugu tag set.

Received 20 Jun., 2020; Accepted 05 Feb., 2020 © The author(s) 2020.

Published with open access at www.questjournals.org

I. INTRODUCTION

Recently NLP has been applied to Android security also and has got promising results. App reviews or app code and the manifest files are used for this purpose to analyze security. The framework called WHYPER detects risky applications automatically. The CHABADA system verifies description of textual applications. These systems help the user to take decisions according to the security threat. This NLP is used a lot than human language processing and we can apply to any text in written format.

POS tagging is the process of assigning the correct tag to the word from a number of available tags. Here the tag means grammatical information of the word. It is well known that a computer will understand and process the language if the meaning of each and every word of that language is known. In most of the Natural Language Processing applications like grammar checking, sentence identification, phrase chunking etc. the computer required only grammatical information of the input text. This grammatical information is given in the form tags called part of speech tags. Here the parts of speech are different word classes in which a word lies like noun, adjective, verb etc. A word can occur in more than one word class in different context. Same word can act as a noun in one sentence and the same can act as a verb in other sentence.

So in order to assign the exact grammatical information to a word one must know the context in which that word has occurred. For a computer system it is very difficult to understand the context of the sentence. Therefore different techniques are used to assign the part of speech tag to a word.

II. RELATED WORK

Different techniques used for POS tagging of Indian Languages:

There are basically three techniques used for part of speech tagging. 1) Rule based method 2) Statistical based method and 3) Neural network based method. Besides these three a hybrid method is also used. This hybrid method is the combination of two or three of above mention techniques. In rule based technique different hand written rules are used for disambiguation of tags. These rules are developed manually. Therefore thorough knowledge of language is required to develop the rules. This rule based technique has been used by Sreeganesh (2006) for Telugu language; Statistical method used by R. J. Rama Sree et al. (2008) is another technique commonly used for Telugu part of speech tagging. Most commonly used statistical methods are

support vector machine (SVM) used by Ekbal and S. Bandyopadhyay(2008) for Bengali language; V.Dhanalakshmi et al. (2008)for Tamil language, M Anandkumar, Vijaya M.S, Loganathan R, Soman K.P, Rajendran S (2008) ;Sindhiya Binulal et al. for POS tagging of Telugu language. Antony P.J et al. for Malayalam language. Hidden markov model based technique used by Manish Shrivastava&Pushpak Bhattacharyya for POS tagger for Hindi language; Manju K et al. for Malayalam language; Navanath Saharia et.al for Assamese; Sanjeevkumar Sharma et al. (2011) for Punjabi Language; Ekbal, S. Mondal et al. for Bengali language. Maximum entropy based technique was used by Aniket Dalal et al. for Hindi language; Ekbal et al. (2008) for Bengali language. Another rule based POS tagger was developed for Punjabi language by Mandeep Singh Gill, Gurpreet Singh Lehal (2008). Conditional Random Field based technique has been used by Ravindran et al. and Himanshu et al. for POS tagging and chunking of Hindi language; other Indian languages on which this CRF technique has been applied are Bengali and Manipuri. Neural network based technique has been used by Ankur Parikh for Hindi Language. Chirag Patel and Karthik Gali used a combination of rule based method and CRF for Gujarati Pos tagger. In hybrid based approach used a combination of rule based and HMM based technique has been used by Arulmozhi P et al. for development of Tamil POS tagger.

Existing POS tagger of Telugu Language

Telugu language is a member of the Indo-Aryan family of languages. It is also known as Indic languages. Other members of this family are Tamil, Hindi, Bengali, Gujarati, and Marathi, Punjabi etc. Two POS tagging system with two different techniques have been developed for Telugu language. First by Sreeganesh (2006) and second was R. J. Rama Sree et al. (2008). The first system was developed as a sub part of grammar checker project. These rules were implemented by using regular expression. The main reason for using this rule based technique was that the rules can be edited i.e. new rules can be added or deleted. A tag set of more than 620 tags was also developed. Second POS tagging system has been developed by using statistical method. Hidden Markov model was used to disambiguate the tags. Viterbi algorithm was used for implementation of Hidden Markov model. The tag set used in the second system was same as was proposed by R. J. Rama Sree et al. They also tried a hybrid approach that is combination of rule based system and statistical approach in which the output of rule based system was fed to the statistical based system. This gives further improvement on the accuracy of the POS tagger.

Tag set

A tag set is a set of all the tags used to represent the grammatical information of the language. The number of tags used for a language depends upon the length of the tag which further depends upon the amount of information that we want to represent using a tag. e.g. if just basic word class is to be represented with each word then the length of the tag will be 2, 3 or 4. One extra character will be required for extra grammatical information that is to be represented with tag. E.g. to represent only word class we can use NN tag for noun. But if we want to represent gender information also then an extra character will be added to this tag. This extra character may be M for masculine gender, F for feminine gender and B for both types. Therefore proposed tag for a masculine noun will be NNM, for feminine it will be NNF and for both categories it may be NNB. This extra information not only increases the length of the tag but also increase the no of tags. As in above case if the information of only word class is to be represented then only one tag was sufficient and as the information increases the number of tags also increase and becomes 3 in above case. From above discussion it is concluded that the information has a direct effect on the number of tags.

Telugu tag set

Existing Telugu POS Tagset

As discussed in above section two POS tagger has been developed for Telugu language. In both of these POS taggers same tag set has been used. This tag set was developed by keeping in mind that this POS tagger has to be used for grammar checking software of Telugu language. This tag set was fine grained and more than 620 tags were used.

New proposed tag set by TDIL:

Depending on some general principle of tag set design strategy, a number of POS tag sets have been developed by different organizations. For POS annotation of texts in Telugu, we have used tag set proposed by TDIL (Technical Development of Indian Languages). There were 36 tags proposed by TDIL for Telugu language.

III. INTRODUCTION TO HMM

Hidden Markov Model was proposed by L. E. Baum. It is a statistical model used to solve classification type of problems. This model is used to assigns the joint probability to paired observation and label sequence. In order

to maximize the joint likelihood of training sets parameters are trained. In NLP applications this type of training is done by using accurate annotated corpus. The main advantage of this model is that it is easy to understand and implemented. The accuracy of this model is directly proportional to size of training data.

Basic Definitions and Notation

According to (Rabiner, 1989), the HMM can be defines by using the following five elements:

1. N , it is the number of distinct states in the model. For part-of-speech tagging, N is the total number of tags that can be used by the system. In existing system these are more than 360 tags and in our propose system these are 36 tags only. Each possible tag for the system corresponds to one state of the HMM.

2. M , the number of distinct output symbols in the alphabet of the HMM. For part-of-speech tagging, M is the number of words in the lexicon of the system. As the exact number of words of a language can't be counted so the distinct words present in the training corpus is taken as M .

3. $A = \{a_{ij}\}$, the state transition probability distribution. Where a_{ij} is the probability that the system will move from state i to state j in one transition. For part-of-speech tagging, these states are represented by the tags, so a_{ij} is the probability that the model will move from tag t_i to t_j . This probability can be estimated using training corpus.

4. $B = \{b_j(k)\}$, the emission probability. The probability $b_j(k)$ is the probability that the k -th output symbol will be emitted when the model is in state j . For part-of-speech tagging, this is the probability that the word W_k will be assign tag

t_j (i.e., $P(W_k/t_j)$). This probability can be again estimated from a training corpus.

5. $\Pi = \{\pi_i\}$, the initial state distribution. It is the probability that the model will start in state i . For part-of-speech tagging, this is the probability that the sentence will begin with tag t_i . of tags that maximizes $P(V/U)$. The Viterbi algorithm is a common method for calculating the most likely tag sequence when using an HMM. The proposed model is a type of first order HMM, also referred to as bigram POS tagging. For POS-tagging problem presented Hidden Markov Model is composed of two probabilities: lexical (emission) probability and contextual (transition) probability (Samuelsson, 1996).

Using Baye's law above equation can be rewritten as:

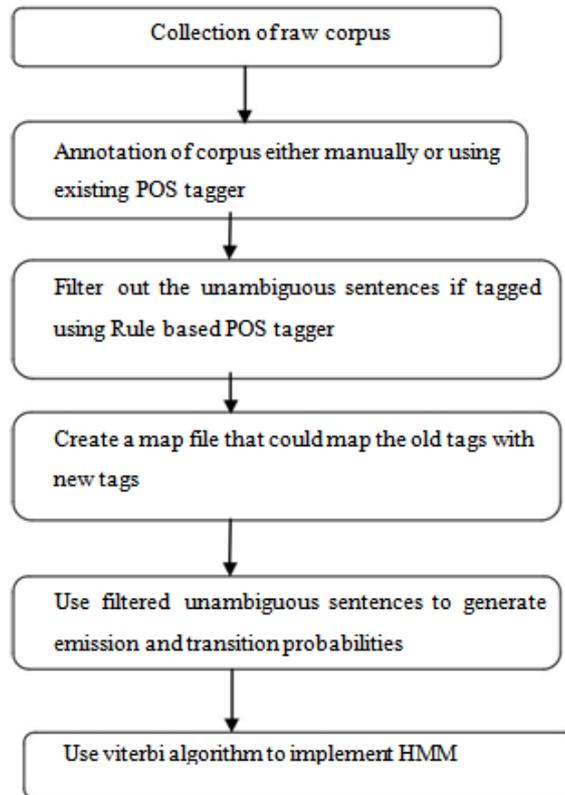
$$P(t_1, \dots, t_n | W_1, \dots, W_n) = P(t_1, \dots, t_n) \times \frac{P(W_1, \dots, W_n | t_1, \dots, t_n)}{P(W_1, \dots, W_n)}$$

$$(t_1, \dots, t_n) = \underset{t_1, \dots, t_n}{\text{argmax}} P(t_1, \dots, t_n) \times \frac{P(W_1, \dots, W_n | t_1, \dots, t_n)}{P(W_1, \dots, W_n)}$$

$$(w_0, \dots, w_n) = \underset{w_0, \dots, w_n}{\text{argmax}} [P(t_i | t_{i-1}) (\text{TRANSITION PROBABILITY}) * P(W_i | T_i) (\text{EMMISSION PROBABILITY})]$$

IV. METHODOLOGY

Following flow diagram shows the steps used to implement new tag set in HMM.



Step 1:- collection of raw corpus.

We collected a large accurate corpus of nearly 200 pages containing nearly 8000 sentences and approximately 42,000 words. This corpus was collected from internet. Following web sites were used for the collection of corpus

- [http:// www.eenadu.net](http://www.eenadu.net)
- [http:// www.andhrajyothy.com](http://www.andhrajyothy.com)
- [http:// www.sakshi.com](http://www.sakshi.com)
- [http:// www.epaper.vaartha.com](http://www.epaper.vaartha.com)

Step 2:- Annotation of corpus.

For the annotation we used an existing lexicon based morphological analyzer. This morphological analyzer contains more than one million Telugu words with their part of speech tag

Step 3:- Mapping with new tags.

Since the existing morph used in step 2 contains tags that have been developed using a tagset of 620 tags and in new system we have to use the tags proposed by TDIL, so mapping of tags was done to reduce the 620 tags to 34 tags. The tags of annotated corpus developed in above step were converted in to standard common tags set for Indian Languages and as per IIT tagset guidelines. This mapping was partially done manually and partially by computer by using a map file. The map file was manually developed with the help of linguistic.

Mapped File

Sr No	Old tag starts with	New tag
1	NN	N_NN/N_NST/N_NNP
2	PNP	PR_PRP
3	PNR	PR_PRF

4		PND	PR_PRC
5		PNI	PR_PRI
6		PNE	PR_PRL
7		PNN	PR_PRQ
8		AJI	JJ
9		CD	QT_C
10		OD	QT_O
11		PPU	PSP
12		AVI	RB
13		PPI	PSP
14		AVU	RB
15		VBP	V
16		CJ	CC
17		PTU	RP
18		PTV	RP
19		AJU	JJ
20		VBO	V
21		VBMA	V_VM
22		BVAX	V_VAUX
23		Unknown	RD_UNK
24	Comma , Dot, Colon, Semicolon, Opening Single Quote, Closing Single Quote, Opening Double Quote, Closing Double Quote	Question Sentence, Sentence, ,Exclamation	RD_PUNC
25	Opening Bracket, Opening Brace , Closing Bracket, Closing Brace, Opening Parenthesis, Closing Parenthesis, Less Than, Greater Than		RD_SYM
26		VBMAXPSXXTNE	V_VM_VNG
27		AJU	RP_INTF
28		PTUN	RP_NEG

29	AJIMSD	QT_QTF
30	No specific match but generally unknown words	RD_RDF
31	No specific tag but words with hyphen in between them	RD_ECH
32	VBMA XXXX XINIAN	V_VM_VINF
33	VBMA XXXX XINDIAN	V_VM_VNF

Step 4:-Development of emission and transition probability file. Now in order to find out transition and emission probabilities we developed an application in visual studio (c#.net). The probability files were kept in txt format. These file were used for implementing HMM using viterby algorithm.

Sample transition file

Tag1/Tag2 pair	Probability
N_NN/V	0.190476
V/V_VM	0.005376
V_VM/RP	0.040698
RP/PR_PRP	0.016484
PR_PRP/N_NN	0.01066
N_NN/PSP	0.470769
PSP/N_NN	0.028161
N_NN/V_VM	0.135289

Step 5:- Viterbi algorithm was used to implement HMM.

V. EXPERIMENTAL EVALUATION

The accuracy of Natural Language product is generally measured in terms precision and recall. Precision is the percentage of correctly disambiguate tags. And recall is If A is the number of correctly disambiguate tags and B is the number of tags that were not disambiguate by our system then

$$\text{Recall} = A / (A+B)$$

Similarly if A is the no of correctly disambiguated tags and C is the number of incorrectly disambiguated tags then Precision= A / (C+A)

For evaluation of the proposed POS tagger, a corpus having texts from different online resources i.e. Telugu websites were used. The outcome was manually evaluated through a linguistic expert to mark the correct and incorrect disambiguate tags. The result obtained has been given in Table 1. The precision and recall values are given in table 2.

Table 1 Experimental Result

corpus	Total number of words	No of unknown words (not tagged by the system)	No of known words	Existing HMM based system	Proposed system
				No of correctly disambiguated tags	No of correctly disambiguated tags
Essay	5995	325	5670	5233	5610
News	4007	344	3663	3369	3595
Short stories	8047	69	9333	8602	9297
Novel	2508	211	2297	2188	2267
Book Chapter	2134	23	2111	2013	2089

Table 2 Precision and Recall of HMM based and Proposed system

Corpus type	Existing HMM based system					Proposed System				
	A	B	C	Precision	Recall	A	B	C	Precision	Recall
Essay	5233	775	0	100%	85.2%	5610	0	60	98.9%	100%
News	3369	472	0	100%	86%	3595	0	68	98.1%	100%
Short stories	8602	1162	0	100%	86.5%	9297	0	36	99.6%	100%
Novel	2188	285	0	100%	87%	2267	0	30	98.6%	100%
Book Chapter	2013	292	0	100%	85.5%	2089	0	22	98.9%	100%

VI. CONCLUSION

NLP plays a vital role in Data Governance. NLP collects enormous data from the users. It handles ownership of data, the privacy of a user and providing security. It has more control over whatever the user uses his data even. Google, Amazon, Microsoft, and others are the big tech corporations for this NLP. So the government needs to regulate on collecting data and disseminated data through Natural Language Processing. Tagset had reduced for efforts to improve the accuracy of HMM based Telugu POS tagger. The tagset has been reduced from more than 620 tags to 34 tags. We observed a significant improvement in the accuracy of tagging. Our proposed tagger shows an accuracy of 92-95% whereas the existing HMM based POS tagger was reported to give an accuracy of 85-87%. This significant improvement is due to reduction in the tagset from more than 630 tags to 36 tags. The main problem with large tag set results in data sparseness. The reduction in the tags results in reduction in data sparseness and hence the accuracy improved.

REFERENCES

- [1]. R. J. Rama Sree, G Umamaheshwar Rao, and K. V Madhu Murthy, "Assessment and Development of POS Tagset for Telugu", In Proceedings of the Sixth Workshop on Asian Language Resources, 3rd International Joint Conference on Natural Language Processing (IJCNLP-08), IIT Hyderabad, Hyderabad, India, 2008, pp 85–88.
- [2]. Srinivasu Badugu "Morphology Based POS Tagging on Telugu" IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 1, No 1, January 2014 ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784. www.IJCSI.org ,pp 181 – 167.
- [3]. Sharma S.K, Lehal G.S (2011) "Using HMM to Improve accuracy of Punjabi POS tagger" 2011 IEEE International Conference on computer science and Automation Engineering. Shanghai (China)
- [4]. Ahmed, Raju S.B, Chandrasekhar Pammi V. S., Prasad M.K (2002), "Application of multilayer perceptron network for tagging parts-of- speech", Proceedings of the Language Engineering Conference, IEEE Workshop on Artificial Intelligence.
- [5]. Ankur Parikh (2009), "Part-Of-Speech Tagging using Neural network", Proceedings of ICON-2009: 7th International Conference on Natural Language Processing.
- [6]. Antony P.J, Mohan S. P., Soman K.P (2010), "SVM Based Part of Speech Tagger for Malayalam", Proceedings of 2010 International Conference on Recent Trends in Information, Telecommunication and Computing, IEEE.
- [7]. Anupam Basu, Ray, Ranjan Pradipta, Harish V. and Sarkar Sudeshna (2003), "Part of speech tagging and local word grouping techniques for natural language parsing in Hindi", Proceedings of the International Conference on Natural Language Processing (ICON 2003).
- [8]. Arulmozhi P, L Sobha (2006) "A Hybrid POS Tagger for a Relatively Free Word Order Language", Proceedings of MSPIL-2006, Indian Institute of Technology, Bombay. Avinesh PVS and Gali Karthik (2007), "Part-of-speech tagging and chunking using conditional random fields and transformation based learning", Proceedings of the IJCAI and the Workshop On Shallow Parsing for South Asian Languages (SPSAL), pp. 21–24.
- [9]. Chirag Patel and Gali Karthik (2008), "Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields", Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, Hyderabad, India, pp. 117– 122.
- [10]. Ekbal, S. Mondal and S. Bandyopadhyay (2007). POS Tagging using HMM and Rule-based Chunking. In Proceedings of the Workshop on Shallow Parsing in South Asian Languages, International Joint Conference on Artificial Intelligence (IJCAI 2007), 6-12 January 2007, Hyderabad, India, PP. 25-28
- [11]. Ekbal, R. Haque and S. Bandyopadhyay (2007), "Bengali Part of Speech Tagging using Conditional Random Field", Proceedings of the 7th International Symposium on Natural Language Processing (SNLP-07), Thailand, pp.131-136.
- [12]. Ekbal and S. Bandyopadhyay (2008), "Part of Speech Tagging in Bengali using Support Vector Machine", Proceedings of the International Conference on Information Technology (ICIT 2008), pp.106-111, IEEE.
- [13]. Ganesan M (2007), "Morph and POS Tagger for Tamil" (Software) Annamalai University, Annamalai Nagar. G. Sindhiya Binulal, Goud P. A, K.P.Soman (2009), "A SVM based approach to Telugu Parts Of Speech Tagging using SVMTool", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.
- [14]. Himanshu Agrawal, Mani Anirudh (2006), "Part Of Speech Tagging and Chunking Using Conditional Random Fields" Proceedings of the NLP AI ML contest workshop, National Workshop on Artificial Intelligence.
- [15]. Manjuk, S. Soumya, Idicula S.M. (2009), "Development of A Pos Tagger for Malayalam-An Experience". Proceedings of 2009 International Conference on Advances in Recent Technologies in Communication and Computing, IEEE
- [16]. Manish Shrivastava, Bhattacharyya Pushpak (2008), "Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information Without Extensive Linguistic Knowledge", Proceedings of ICON-2008: 6th International Conference on Natural Language Processing.
- [17]. Navanath Saharia, Das Dhruvajyoti, Sharma Utpal, Kalita Jugal (2009), "Part of Speech Tagger for Assamese Text", Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Suntec, Singapore, pp. 33–36.
- [18]. Ekbal, M. Hasanuzzaman and S. Bandyopadhyay (2009), "Voted Approach for Part of Speech Tagging in Bengali", Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC-09), December 3-5, Hong Kong, pp. 120-129.
- [19]. http://en.wikipedia.org/wiki/Telugu_language
- [20]. http://en.wikipedia.org/wiki/Telugu_grammar

Dr. V. Suresh "Reduced Tagset To Improve Accuracy Of HMM Based Parts Of Speech Tagger In Telugu Language" Quest Journals Journal Of Software Engineering And Simulation, Vol. 05, No. 02, 2019, Pp. 10-17