ISSN 2349-6193
Impact Factor: 2.805

IJESMR

International Journal OF Engineering Sciences &Management Research

IMPROVES THE ACCURACY TO COMBINE SEVERAL POS TAGGERS FOR TEXTS IN TELUGU LANGUAGE

Vayyasi Suresh*
*HOD and Assistant Professor, Department of Electronics and Computer Engineering(ECM),
Vignan's Institute of Information Technology, Duvvada,Visakhapatnam, Andhra Pradesh, India

Keywords: *NLP , Telugu morphological analyzer , POS tagger, Sentence Tokenizer, word sense disambiguator*

ABSTRACT

POS Taggers are developed by modeling the morpho-syntactic structure of natural language text.POS Tagging is the process of assigning a correct POS tag (noun, verb, adjective, adverb) to each word of the sentence. The three Telugu POS taggers are improve the accuracy of existing Telugu POS taggers by using an voting algorithm. viz., (1) Rule-based POS tagger (2) Brill Tagger (3) Maximum Entropy.POS taggers are developed with an accuracy of 97.014%, 93.248%, and 85.914 respectively. An annotated corpus of 14000 words is used to train the last two taggers. To improve the accuracy of these taggers, an error analysis is made to find out the errors made by these three taggers and methods are then examined. As a first step, a voting algorithm is proposed to get better results to build an ensemble Telugu POS tagger. This tagged output could be used for word sense disambiguation (WSD) is retrieving Telugu documents and a variety of NLP (Natural Language Processing) applications..

INTRODUCTION

In considering the role or function of the word in the sentence, POS tagging is the process of assigning a tag like noun, verb, pronoun, preposition, adverb, adjective or other lexical class marker to each word in a given sentence. POS tagging is a difficult process due to the following reasons.

(a) **Morpho-syntactic ambiguity** : For example, the word "pen" can be taken as *verb* or *noun* in English.

(b) **Existence of unknown words in the language**: For each language, new words always get added and it becomes impractical to keep track of all borrowed words of the language.

However, We tried to analyze small Telugu corpus using a Telugu morphological analyzer (MA),POS tagging is very much required to reduce the syntactic ambiguity. we observed that 33% of the words are identified by Telugu Morphological analyzer that has coverage of 97%. More than 42% of the words are ambiguous and 25% of the words are unknown. The non-identification of words by MA is due to (i) the presence of proper nouns (ii) conjoining of two or more number of words and (iii) existence of foreign words. POS tagger with high accuracy is very much useful, in order to identity the correct analysis in the given context.

RELATED WORK

All related work in the area of POS tagging can be broadly classified into four categories viz.,

(i) **Rule-based**: Rule-based taggers generally consist of two phases. The first phase is concerned with getting all possible tags of each word of the sentence and the second phase is concerned with identification of the correct tag by using some hand written rules.

(ii) **Stochastic based:** Stochastic based taggers which in turn can be classified as
- Hidden Markov Models-HMM taggers.
- Maximum Entropy taggers - MXPOST, Maccent system, Swedish POS tagging, Chinese
- Memory Based
- Connectionist
- Decision Tree etc., depending on how language modeling was done to assign POS tags to the words in a given sentence.

(iii) **Transformation based Learning** and

IJESMR
International Journal OF Engineering Sciences &Management Research

(iv) **Ensemble approaches -** Statistical n-gram taggers assign a part-of-speech label to each word in a text on the basis of probability estimates that are automatically derived from a large, already tagged training corpus.

Some researchers examined the grammatical constructions which cause such taggers to falter most frequently. As one would expect, certain of these errors are due to linguistic dependencies that extend beyond the limited scope of statistical taggers which lead to the idea of combining classifiers in the area of machine learning for enhancing accuracy of POS tagging. These works showed the process of combining the existing freely available taggers by using linguistically motivated rules so that tagging accuracy of the combination exceeds that of the best of the individual taggers.The bulk of literature on POS is for English. As far as Indian Languages are concerned, non-availability of lexical resources is a bottle neck for POS tagging. However some works are being done in the area of POS taggers in IITs and Language Technology institutes.

POS TAGGERS FOR TELUGU

The words are formed by joining morphemes together is an agglutinative

Telugu language in which. In this paper, we describe three POS taggers developed in different ways viz., (1) Rule-based approach, (2) using Transformation based learning (TBL) approach of Erich Brill (3) using Maximum Entropy Model, a machine learning technique. An annotated corpus of 14000 words is constructed to train the taggers for the last two methods.

For all the three taggers, the input is a Telugu sentence transliterated in wx-notation as shown in Appendix-1 and output is the same sentence where each word is

tagged with its right tag.For example,

Input : govu manaku cAlA paviwramEna jaMwuvu **.**

(గోవు మనకు చాలా పవిత్రమైన జంతువు)

(cow to us very holy animal.)

(cow is a very holy animal to us)

Output : govu/nn1 manaku/pr4 cAlA/if paviwramEna/jj jaMwuvu/nn1 ./sym

(గోవు/nn1 మనకు/pr4 చాలా/if పవిత్రమైన/jj జంతువు/nn1 ./sym)

Rule-based POS tagger for Telugu : The overview of Telugu Rule-based tagger is shown in Figure-1. It consists of a series of modules as described below. **Sentence Tokenizer** which is responsible for segregating the input text into a series of sentences and each sentence into words such that each sentence and word are given a identification number. **Telugu Morphological Analyzer** which gives all possible analyses of each word of the given input sentence. At present care is taken in such a way that all words are recognized by the Morphological Analyzer. This is done by pre-editing the Telugu Texts. How ever, some words may not be identified by MA

 ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences &Management Research

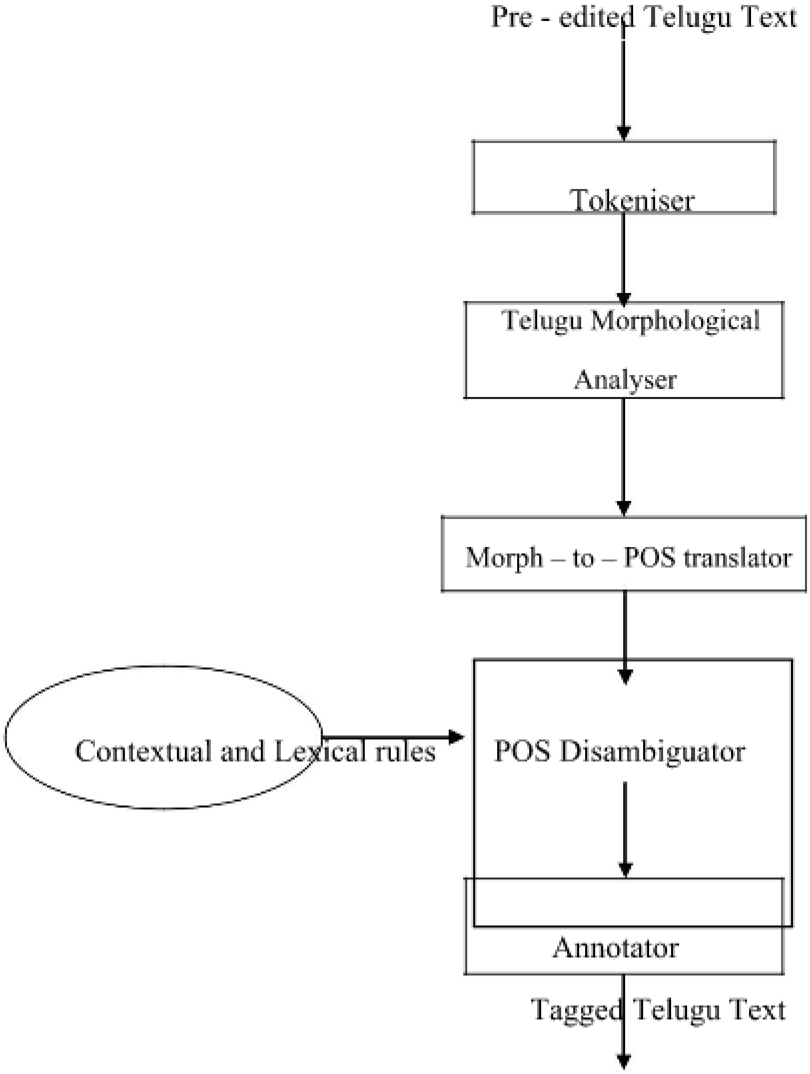

Fig. 1. Overview of Telugu Rule–based POS tagger

due to the presence of – (i) foreign words and (ii) compound words. In the case of foreign words, if the words are used in the Telugu language frequently, then these words are added into the dictionaries (For example, bus, gas etc). Otherwise they are translated into Telugu. Compound words are segregated.

- **Morph to POS Translator** which converts all the morphological analyses into their corresponding POS tags in the tag set using some pattern rules. The number of POS tags for each word is equal to the number of analyses.

- **POS Disambiguator** which reduces the above POS ambiguity for each word. Presence of more than one POS tag for a word indicates the ambiguity at word level. This ambiguity is reduced by the application of ungram and bigram rules which are written taking context into consideration

- **Annotator** which produces the tagged text.

The baseline performance of this tagger is found to be 98%, provided the Telugu texts are pre-edited. However, the task of pre-editing is little bit a difficult task. The lower performance of the tagger for some texts can be attributed to the scope of the ambiguity. Some times the ambiguity is beyond the scope of bigrams. If the domain is large, it is very difficult to write rules. The reasons for this, are as follows – (i) we need to have to write more number of rules. (ii) The complexity of the problem increases as the size of the domain increases. It

IJESMR
International Journal OF Engineering Sciences &Management Research
is found that it is difficult to develop a general purpose rule-based POS tagger for Telugu as its syntactic distribution varies from speaker to speaker.

Implementation of Brill's Tagger for Telugu
There are three main phases in implementing Brill Tagger for any language.

They are
(i) Training phase – in which it first extracts rules from the training corpus using statistical techniques.
(ii) Verification phase – in which these rules are verified by taking an annotated text with its tags removed as the input and generates the tagged text; this tagged text is compared with its original tagged text and learns where it has gone wrong;
(iii) Testing phase – in which new unseen texts can be tagged. The accuracy of the tagger when applied to different European languages is above 95%. The results of applying Brill's Transformation Rule-Based Learning (TBL) for Telugu are studied and it is shown that the present system does not obtain a very high accuracy but results are still promising with base line accuracy of 90%.

Implementation of Maximum Entropy Tagger for Telugu
Training a Maximum Entropy model is relatively easy. There is a Maximum

Entropy Modeling toolkit freely available on the net. This toolkit consists of both Python and C++ modules to implement Maximum Entropy Modeling. More over, there is a separate language and tag set independent toolkit in Python (maxent) as a case study for building a POS tagger. This is straightly used to build POS tagger for Telugu. The maxent tagger was tested for Telugu and found that average performance was 83.47 which is also comparatively less when compared to European languages.

Telugu Training Corpus
An annotated corpus of 14000 words is created for this purpose. The following table gives the information of the Telugu training corpus.

Statistics of the Training Corpus

CHARACTERISTICS	NUMBER
Number of sentences	15012
Number of words	13146
Number of unambiguous words	6452
Unknown words to Telugu Morph	976
Number of ambiguous words with 3 tags	3657
Number of ambiguous words with 4 tags	959
Number of ambiguous words with 5 tags	231
Number of ambiguous words with 6 tags	53

THE ACCURACY OF IMPROVING POS TAGGING

The accuracy of POS tagging is increased by a simple voting algorithm which gives one vote to each tagger output. The accuracy of the tagged Telugu texts is increased not by optimizing the performance of the individual taggers but it was done by improving beyond the accurate single tagger. The overall error rate reduces by 5% for

IJESMR
International Journal OF Engineering Sciences &Management Research
machine learning tagger and 0.82% for Rule-base Telugu Tagger. But it was observed that errors made by the three taggers are independent. Hence simple voting may seems to be better for limited text inputs. However it takes a great amount of time and effort to test and evaluate these outputs of the taggers as gold standard data and tag set definitions are not yet standardized for Indian Languages. Also it is required to test these taggers and voting algorithm on a large testing samples.

However it is required to explore whether accuracy can be improved by giving different weights for voting to different taggers depending on their performance accuracy. Another way is to train each classifier on the tagged texts so generated by voting algorithm using good tagger which is known as stacked classifier.

RESULTS

The following table shows the one of the output sentence from the sample output of comparative study of the three taggers. Refer appendix-1 for transliteration.

Word number	Word	Morph Output	Rule-based Tagging	Brill Tagging	Maximum Entropy Tagging	Output of voting algorithm
0	oVka	jj	Jj	jj	Jj	jj
1	vyApAri	nn1,nni	Nni	nn1	nn1	nn1
2	oVkasAri	nn1,nni	nn1	nn1	nn1	nn1
3	oVka	jj	Jj	jj	Jj	jj
4	mahanIyudu	jj	Jj	nn1	nn1	nn1
5	cese	vnf,vrb	Vrb	vnf	Vnf	vrb
6	prasaMgAlanu	nn2	nn2	nn2	nn1	nn2
7	vinadAniki	nn4	nn4	nn4	nn4	nn4
8	poyAdu	vf	Vf	vf	Vf	vf
9	.	sym	Sym	sym	Sym	sym

CONCLUSION

Even when simple methods are used for combining several taggers improves the accuracy of tagged texts which help in turn to generate good applications of NLP. The overall error rate reduces by 5% for machine learning tagger and 0.82% for Rule-base Telugu Tagger. This leads to a fewer errors and reduces human effort to evolve a new tagger. A main task in the process of Information Retrieval, the Telugu annotated text so generated is useful mainly in word sense disambiguation.

IJESMR
International Journal OF Engineering Sciences &Management Research

REFERENCES

1. *Brant T. 2000. TnT a statisticalpart-of-speech tagger.Proceedings of the 6th Applied NLP Conference, pp 224-231.*
2. *Brill E. 1992. A Simple Rule-Based Part of Speech Tagger. Proceedings of the Third Conference on Applied Natural Language Processing, Torento,Italy.*
3. *Brent A Olde, James Hoeffner, Patrick Chipman, Arthur C Graesser, the Trutoring Research Group.A Connectionist Model for Part of Speech Tagging.*
4. *Chanod J P, Tapanainen P. 1995. Tagging French: comparing a statistical and a constraint-based method. Procs. 7th Conference of the European Chapter of the Association for Computational Linguistics, pp. 149-157, ACL.*
5. *Andrew Freeman. Brill's POS tagger and a Morphology parser forArabic. Department of Near Eastern Studies, Ann Arbor, Michigan, 48109,USA.*
6. *Beat Megyesi. Data Driven Methods for POS tagging and Chunking of Swedish.*
7. *Berthelsen H, Magyesi B. 2000. Ensemble of Classifiers for Noise*
8. *Detection in POS Tagged Corpus. In Proceedings ofthe Third International Workshop on Text, Speech and Dialogue, LNCS/LNAI, pp27- 32. Spring-Verlag, September.*
9. *Church K W. 1988. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In Proceedings of. 2nd Conference on Applied Natural Language Processing, pp. 136-143, ACL.*
10. *Roger Garside. 1987. The CLAWS word-tagging system. In Garside, Leech and Sampson (eds.), The Computational Analysis of English. London and New York: Longman.*
11. *Green B, Rubun G. 1971. Automated Grammatical Tagging of English. In Department of Linguistics, Brown University.*
12. *RatnaParkhi A. 1996. A Maximum Entropy Model for Part-of-Speech Tagging. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-96) Philadelphia,PA, USA.*
13. *DeRose S J. 1988. Grammatical Category Disambiguation by Statistical Optimization. In Computational Linguistics 14(1), pp. 31-39, ACL.*
14. *User Manual of Brill Tagger*
15. *Jean-Pierre Chanod, Pasi Tapanainen. 1995. Tagging French: comparing a statistical and a constraint-based method. Proc. EACL'95. ACL, Dublin.*
16. *Jian Zhao, Xlao-LongWang. 2002. Chinese POS Tagging based on Maximum EntropyModel. Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing.*
17. *Kemal Of lazer,Ilker Kuruoz. 1994. Tagging and Morphological Disambiguation of Turkish Text. Proceedings of the Fourth ACL Conference on Applied Natural Language Processing.*
18. *Macklovitch E. 1992. Where the tagger falters. In Proceedings of the Fourth Conference on Theoretical and Methodological Issues inMachine Translation, pp 113-126.*
19. *Deaniel Jurafsky, James H. Martin. Speech and Language Processing: An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Pearson Education series in Artificial Intelligence.*
20. *Deshaspe L. 1997 Maximum Entropy modeling with clausal constraints.*
21. *In Inductive Logic Programming proceedings of the 7th International Workshop, Lecture notes in Artificial Intelligence , pp 109-124*
22. *Daelemans, Walter, Zavrel, Jakub, van der Sloot, Ko. 2002. TiMBL: Tilburg Memory Based Learner. Version 4.3. Reference guide. Technical Report ILK 02-10, Induction of Linguistic Knowledge Research*
23. *Group,Tilberg University, The Netherlands.*
24. *Megyesi B.1998. Brill's POS tagger with Extended Lexical Templates for Hungarian. Department of Linguistics, Stockholm University, Sweden.*
25. *Merialdo.B. 1994. Tagging English Text with a probabilistic model. Computational linguistics, 20:155-157.*
26. *Ma, Qing, Isahara, Hitoshi. 1998. A multi-neuro tagger using variable lengths of contexts. Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of*

ISSN 2349-6193
Impact Factor: 2.805

IJESMR

International Journal OF Engineering Sciences & Management Research

the Association for Computational Linguistics (COLING-ACL), pp 802-806.Montreal, Canada.

27. *Tumer, Kagan, Joydeep Ghosh. 1999. Linear and order statistics combines for pattern classification. Amanda Sharkey (ed.), Combining Artificial Neural Networks. Berlin: Springer-Verlag, pp 127-162.*
28. *Umamaheshwara Rao G. 2004. A Telugu Morphological Analyser. Paper presented at the National Seminar on Language Technology Tools& Implementation of Telugu; Vol – I. Phonology and Moprhology.CALTS, University of Hyderabad, Hyderabad.*
29. *Zhang Le. Maximum Entropy Modeling Toolkit for Python and C++.*
30. *Orphanos Giorgos, Kalles Dimitris,Papagelis, Thanasis and Christrodoulakis Dimitris. Decision Trees and NLP: A case study in POS Tagging. University of Patras.*
31. *..Pasi Tapanainen, Atro Voutilainen. 1994. Tagging accurately Don't guess if you know. Proceedings of the Fourth ACL Conference on Applied Natural Language Processing (13--15 October 1994, Stuttgart)*
32. *Voutilainen Atro. 1995. A syntax-based part speech analyser, Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics, March 27-31. Association for Computational Linguistics. Dublin. pp 157-164.*
33. *Voutilainen Atro. 1997. EngCG tagger, Version 2. Tom Brondsted and Inger Lytje (Eds.) 1997, Sprog og Multimedier. Aalborg Universitetsforlag, Aalborg.*