International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 1, Issue 1, May-June 2012 ISSN 2278-6856

POS Tagger for Kannada Sentence Translation

Mallamma V Reddy[1], Dr. M. Hanumanthappa[2]

[1,2]Department of Computer Science and Applications,
Bangalore University, Bangalore, INDIA
{[1]*mallamma_vreddy*,[2]*hanu6572*}*@bub.ernet.in*

Abstract: ***Syntactic analysis of a sentence is performed by parsing technique. The major used language in the*** Dravidian ***languages of India is Kannada Language. On the basis of computational linguistic Kannada is lagging compared to Telugu. Writing the grammar production for any south Indian language is bit difficult. Because the languages are highly inflected with three gender forms and two number forms.In the Majority of Indian language including Kannada, we can identify the gender of a person(Noun/ Pronoun) by a verb ends with a token. This paper highlights the process of text preprocessing, translation of English to Kannada and then generating and implementing Phrase Structure Grammar (PSG) for Kannada sentences. We have used our own Part-Of-Speech (POS) tagger generator for assigning proper tags to each and every word in the training and test sentences. The proposed POS tagger used for parsing Kannada sentences and is implemented using supervised machine learning approach.***

Keywords: Machine Translation (MT), Natural Language Processing (NLP), Phrase Structure grammar (PSG), Part-Of-Speech (POS).

1. INTRODUCTION

Kannada or Canarese is one of the 1652 mother tongues spoken in India. Forty three million people use it as their mother tongue. Kannada has 44 speech sounds. Among them 35 are consonants and 9 are vowels. The vowels are further classified into short vowels, long vowels and diphthongs. It is also one of the 18 Scheduled Languages included in the VIII Schedule of the Constitution of India. It belongs to the Dravidian family of languages. Within Dravidian, it belongs to the South Dravidian group. The Dravidian languages stand apart from other family of Indian languages like Indo Aryan, Sino Tibetan and Austro Asiatic by having distinctive structural differences at phonological, morphological, lexical, syntactic and semantic levels. It is recognized as the Official Language of the state of Karnataka [1]. The task of Part of speech (POS) [9] tagger is simply assigning a part of speech to a word. Each word is a noun, adjective, verb or any one of the other parts of speech. Many words have more than one POS tag. Some of the words are listed in Table 1.

Table 1: words with highest numbers of POS from WordNet

Word	Number of Parts of Speech
Out	5
Round	5
Still	5
Down	5
Over	4

The most likely tag for a word is selected, based on context and other information. This task is like other ambiguity problems in Natural Language. A single word can have multiple meanings and senses. Identifying the most likely meaning or sense and is computed using statistical method.

Information Retrieval (IR) deals with natural language text which is not always well structured and could be semantically ambiguous. The main objective of an IR system is to retrieve all relevant and few non-relevant documents according to user query. We perform Text Preprocessing for IR to tag POS.

1.1 Text Preprocessing

Text preprocessing [5] is a procedure which can be divided mainly into five text operations (or transformations):

1. Lexical analysis categorizes a text into digits, hyphens, case of letters and punctuation marks.
2. The objective of filtering out words with very low discrimination values for retrieval purposes is possible by Elimination of stop words. We have used 850 stop words in our implementation.
3. Stemming of the remaining words with the objective of removing affixes (i.e., prefixes and suffixes) and allowing the retrieval of documents containing syntactic variations of query terms (e.g., connect, connecting, connected, etc). We have implemented suffix stripping and Suffix joining algorithm to make the stems as root words. Rules to remove suffixes are listed in Table 2.

Table 2: Rules to remove suffixes

Noun Rules	Verb Rules	Adjective Rules
s->null	s->null	er->e or null
ses->s	ies->y	est->e or null
xes->x	es->e or null	
zes->z	ed->e or null	
ches->ch	ing->e or null	
shes->sh		

4. Selection of index terms to determine which words/stems (or groups of words) will be used as an indexing element. Usually, a particular word become an index term, the decision of it is depends on syntactic nature of the word. The frequency of noun words has more semantics than adjectives, adverbs, and verbs.

5. The categorization structures construction defined as "extraction of structure directly represented in the text, for allowing the extension of the original query with related terms (a usually useful procedure)". Fig 1 gives an idea about all above steps.

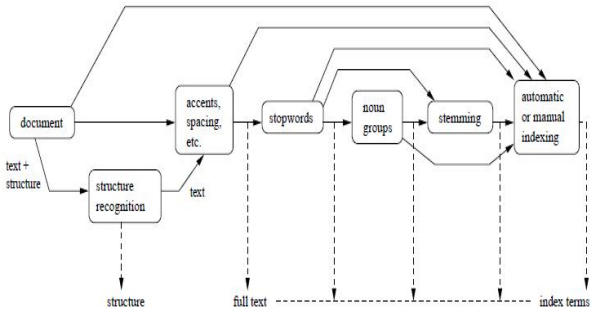


Figure 1 A logical view of document [5] throughout the various phases of text preprocessing.

2. TRANSLATION

Machine translation [7] is the process of translating from source language text into the target language. We have implemented the above for machine translation using 'BUBShabdasagar-2011' [6] Machine Readable Dictionary (MRD) as a translation lexicon resource for our research. The dictionary was available in the ISCII character encoding form and in the plain text format. The entries were converted into UTF-8 encoding. The English→Kannada bi-lingual dictionary has around 14,000 English entries and 40,000 Kannada entries.

3. PARSING

Syntactic analysis is the process of analyzing a text or sentence that is made up of a sequence of words called tokens, and to determine its grammatical structure with respect to a given grammatical rules then tag the token with proper POS.

4. PHRASE STRUCTURE GRAMMAR FOR KANNADA

Language is a collection of many components when all these are arrange in a systematic order. The syntax of a language contains a phrase structure (PS) component and a transformational component [8]. In phrase structure the assumed largest unit of grammar, the sentence [S] is progressively expanded by the application of rules into 'strings' of smaller units because in Transformational grammar (TG) sentence is the syntactic system fundamental unit. The techniques of actual sentences generating structural as per descriptions of sentences are set forth in PS rules. Each rule gives an idea about a symbol representing a component of a sentence to the left of an arrow and a symbol or series of symbols to the right. The following are the symbols used in PS rules: these rules are used as POS tagger for Kannada Sentences. Table 3 includes some of them.

We use a decision tree model to explain rule-based tagging. At the top of the tree all are untagged words as shown in Fig 2. For every word, we traverse s path through a tree till a leaf node. Each node in the tree corresponds to an attribute. The attribute value of a word determines the path from the root to a leaf node. The first attribute at the root node is a flag to indicate whether the word exists in the dictionary. Words that occur in the dictionary can have one or more tags.

Table 3: PS rules: these rules are used as POS tagger for Kannada Sentences

POS Tagger	Meaning	POS Tagger	Meaning
S	Sentence	Be	The verb Be
NP	Noun phrase	Pred	Predicate (noun,adjective,adverb)
VP	Verb phrase	Vt	Transitive Verb
N	Noun	Vi	Intransitive verb
VB	Verb	Vl	Linking Verb
T,art or D	Determiner	Comp	Complement (noun or adjective)
Pron	Pronoun	Adj	Adjective
Aux	Auxiliary	Adv	Adverb
M	Model Auxiliary	PP	Prepositional phrase

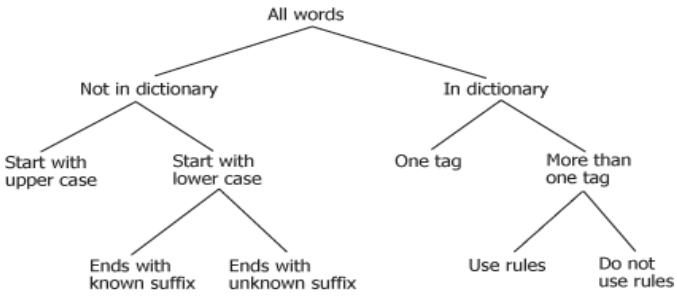


Figure 2. Decision tree to assign POS tags

For words that begin with lowercase letters. We cannot easily assign a tag. The suffixes of these words may provide clues to assigning a tag. For Example, when the suffix "in" is seen, there is a high probability that the untagged word is a verb. Similarly, a word with the suffix "s" is most likely a noun. If the word does not have a known suffix, then we assign a default noun tag.

Words can be classified into two word classes or part of speech (POS) [10]. The eight standard parts of speech are adjectives, adverbs, conjunctions, determiners, nouns, prepositions, pronouns, and verbs. Two other word classes are-interjection and punctuation marks-are sometimes included among POS. we refer four of the eight parts of speech-nouns, verbs, adjectives and adverbs –as a *content words.* The remaining four part of speech-

conjunctions, determiners, pronouns, and prepositions-are called *function words.* Most of the words in the lexicons (Dictionary) are content words are shown in Table 4.

Table 4: Frequencies of word classes taken from WordNet

Type	Number	Type	Number
Noun	114,400(75%)	Preposition	133(.08%)
Adjective	21,438(14%)	Pronoun	118(.077%)
Verb	11,341(7.4%)	Conjunction	89(.05%)
Adverb	4662(3%)	Determiner	14(.009%)

Algorithm: POS tagging

Input: Untagged English Sentence
Output: Tagged Translated Kannada Sentence

Tag<= First Word in Sentence
For each word in the Sentence Do
  If the Word is tagged
    Stop
  Else
    Tag<=Word
  End if
End For
Return Tagged Translated Kannada Sentence

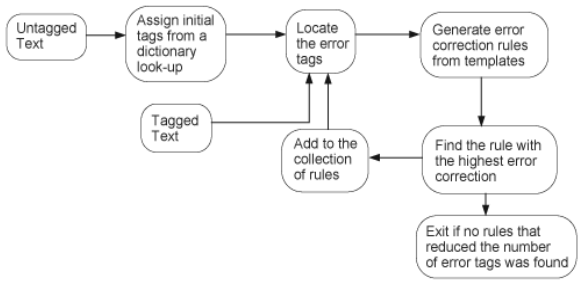


Figure 3 Training a rule based tagger

Example: Parser for the Kannada input sentence ರಾಮ ಚೆಂಡನ್ನು ಎಸೆದನು " Rama threw the ball" .

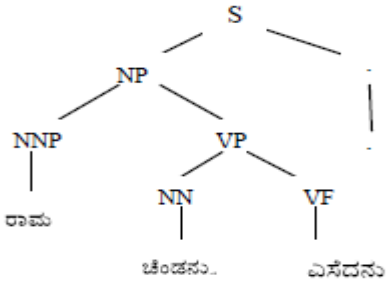


Figure 4 Parse tree structure

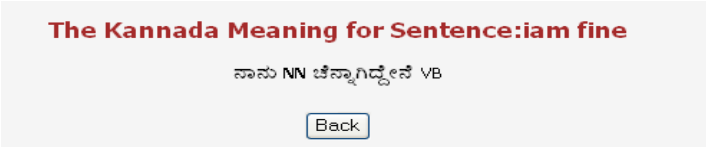
5. EXPERIMENTAL SETUP

Initially, a very limited lexicon and rules were present in POS tagger. So, when we performed POS tagging in our POS tagger, 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 and rules will be added for those new words. After some time, the accuracy level will also be increased as the numbers of lexicons are increased in Bilingual Dictionary. The precision of any part of speech tagger is measured in terms of percentage i.e. the percentage of words, which are accurately tagged by the tagger. This is defined as in (1)

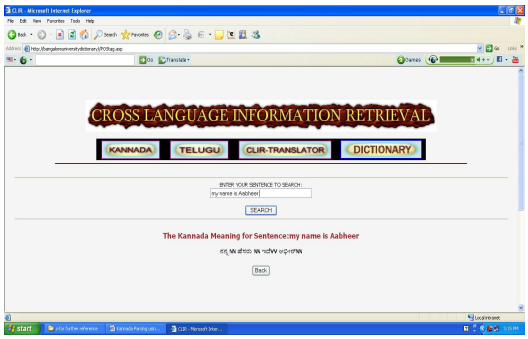
$$Accuracy = \frac{Correctly\ Tagged\ Words}{Total\ No\ of\ Tagged\ words} \quad (1)$$

6. RESULTS AND PERFORMANCE

Cross Language Information Retrieval Tool [8] is built by using the ASP.NET as front end and database as backend. English and Kannada are the source language and the target language, respectively, in our query translation. All the experiments carried out here involve the same set of English queries and the same query expansion, translation and retrieval method. The only difference between the experimental conditions is in what dictionaries are used in the query translation. We have trained the systems with corpus size of 500, 1000 and 1500 sentences respectively for POS tagging. Performances of the systems were evaluated with the same set of 500 distinguished sentences that were out of corpus. From the experiment we found that the performances of our systems are significantly well and achieves very competitive accuracy by increasing the corpus size.



Result 1: POS tagged English-Kannada



Result 2: POS tagged English-Kannada

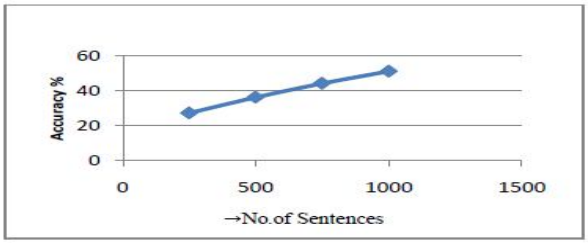


Figure 4 Performance Graph

7. CONCLUSION AND FUTURE WORK

Part of speech tagging plays a vital role in natural language processing. This paper presents a reasonably accurate POS tagger for Kannada language. Part of Speech tagging helps in the creation process of a parser. In future we can also use this parser for tree to tree translation. This will be very useful for bilingual machine translation from English to Kannada language. One of the major challenges is that English has Subject Verb Object (SVO) structure while Kannada has Subject Object Verb (SOV) structure in Machine translation has been unraveled by this parser.

Acknowledgement: I owe my sincere feelings of gratitude to Dr. M. Hanumanthappa, for his valuable guidance and suggestions which helped me a lot to write this paper. This paper is in continuation of the major research project entitled ***Cross-Language Information Retrieval*** sanctioned to Dr. M. Hanumanthappa, PI-UGC-MH, Department of Computer Science and Applications by the University Grant Commission carried out at the Bangalore University, Bangalore, India. We thank to the UGC for financial assistance.

References

[1] The Karnataka Official Language Act". *Official website of Department of Parliamentary Affairs and Legislation*. Government of Karnataka. Retrieved 2007-06-29.

[2] B M Sagar, Dr. Shobha, Dr. Ramakanth Kumar, *"Context Free Grammar (CFG) Analysis for simple Kannada sentences"* published in Special Issue

[3] of IJCCT Vol.1 Issue 2, 3, 4; 2010 for International Conference [ACCTA-2010], 3-5 August 2010

[4] Antony P J, Nandini. J. Warrier, Dr. Soman K P, *"Penn Treebank-Based Syntactic Parsers for South Dravidian Languages using a Machine Learning Approach"* published in International Journal of Computer Applications (0975 – 8887) Volume 7– No.8, October 2010

[5] Mallamma. V. Reddy, Dr. Hanumanthappa. M, *"Interlingual Machine Translation"* is published in *"UACEE International Journal of Artificial Intelligence and Neural Networks"* Volume 2: Issue 1 -25th April 2012. ISSN 2250 – 3749, PP: 19-23. Available online at http://ijainn.uacee.org/vol2iss1/files/ijainn-vol2-issue1-622.pdf

[6] Richard Baeza-Yates, Berthier Ribeiro-Neto book name "*Modern Information Retrieval"*.

[7] Mallamma. V. Reddy, Dr. Hanumanthappa. M ***"****Kannada and Telugu Native Languages to English Cross Language Information Retrieval"* Published in the International Journal of Computer Science and Information Technologies (IJCSIT) volume-2 issues-5 September-October 2011. ISSN: 0975-9646. PP: 1876-1880. Available online at www.ijcsit.com/docs/Volume%202/vol2issue5/ijcsit2011020510.pdf.

[8] S. Kereto, C. Wongchaisuwat, Y. Poovarawan. 1993. Machine translation research and development. In *proceedings of the Symposium on Natural Language processing in Thailand*, pages 167-195, March

[9] Noam Chomsky's book on "*syntactic structures" in* 1957

[10] Manu Kochady's book on " Text Mining Application Programming"

Mallamma .V Reddy received MCA degree from Visvesvaraya Technological University, Belgum, India in 2007. Pursuing PhD in computer science and applications, Bangalore University, Bangalore, India. she has over 3 years of teaching experience. she has published nearly 10 Research Papers in National and International Journal/ Conferences. The areas of interest are Cross Language Information Retrieval, Data Mining, Data Base Management System and Programming Languages.

Dr. M. Hanumanthappa is currently working as a faculty as well as chairman in the Dept. of Computer Science and Applications, Bangalore University, Bangalore. He has over 16 Years of teaching (Post Graduate) as well as Industry experience. His area of Interest includes mainly Data Mining, Information Retrieval and Programming Languages. Besides, He has conducted a number of training programmes and workshops for Computer Science students. He is also the Principle Investigator of UGC-Major Research Project; he has published nearly 50 Research Papers in National and International Journal and Conferences. Currently he is guiding students for Ph.D in Computer Science, under Bangalore University. He is also one of the member of Board of Studies as well as Board of Examiners for various Universities of Karnataka