Design & Development of Kannada to Telugu Translator: A Rule based Approach

P. Aparna
M.Tech student CSE Department,
JNTUCEA,
Ananthapur, India.

***Abstract*: MachineTranslation is the task of translating the sentences or words from one language to another language and it is one of the interesting applied research areas thatdraw ideas and techniques from Linguistic, Computer Science, Artificial Intelligence, Statistics and Translation Theory. Machine Translation plays an important role for sharing the information from one language to another language like English to Hindi, Malayalam to English etc.,which are life transforming stories available in India. There is a huge demand for machine translation between English and various Indian languages.**

The fundamental activity of machine translation application is to manage the vocabulary of words. Inthe existing literature, it has varioustypes of machine translationsystems they are Directbasedmachine translation system, Interlingualmachine translation system, Transfer based approach,Corpus based machine translationsystem,Hybrid based approachetc., In this work, translation of Kannada language as source language to Telugu language as target language has been considered. TheTransfer based approach has been used for this purpose. It is observed that, this method is possible to improve the performance and accuracy of Machine translation. In this paper tagging for adverbs and adjectives are also performed.

Keywords:- Machine Translation, Transfer based approach, Source language, Target language.

I. INTRODUCTION:

Natural language processing is one of the major,oldest and the most active research area for Computer Science, Artificial Intelligence,Linguistics etc., Machine translation is mainly design to analyse and understand the languages that humans use naturally. Itis the task of translation for sentences or words from one language to another language automatically without any human intervention or assistance. Even though Machine language was proposed as a computer application in the 1950s research has been made for sixty years. The research on machine translation has happening Worldwide and it was most successfully providing and promising Machine Translation Systems.

Machine Translation is a significant technology for Localization.Itparticularlyrelevant in a linguisticallydiverse country like India. Eighteenfundamental languages are composed in ten different scripts in India. And those languages are highly inflectional with rich morphology; it has Dravidian language and Indo Aryan language. The languages that are derived from Dravidian language areTelugu, Kannada, Malayalam, and Tamil. Telugu is one of Dravidian language in India.So, the translation among these languages isvery important and it is not possible to manually translate the required resources among these languages. Telugu is second most popular language and official language of Andhra Pradesh.Kannada is a language spoken in India mainly in the state of Karnataka. It is official language of Karnataka and given birth to many Indian languages like Tulu,Kodava etc., Kannada and Telugu are most widely used in southern part of India. Only 7% of population speaks English now the translation can be done manually, automationis restricted to word processing there is problem for large volume of data through manual translation like sport news that are translated from Kannada to local languages and government department annual reports,public sector units can betranslated to Hindi, English and local languages these all are translated manually from Kannada to local languages. By using this human translation it requires more time and cost. This is one of the disadvantages, by using the machine translation system, optimization in fastness and cost is achieved when compared to the human translator. The main scheme of machine translation system is to enhance the accuracy and boost the speed of the translation.

II. LITERATURE SURVEY:

various approaches for machine translationare rule based or linguistic approach,directmachine translation,transfer based approach,interlingualmachine translation, example based approach, non-linguistic or rule based approach, hybrid approach etc.,

II-1 RULE BASED MACHINE TRANSLATION:

Rule based approach requires thelinguistic knowledgeat the time of translation and uses grammar rules.

a.DIRECT BASEDMACHINE TRANSLATION:

As stated by the name, this system directly translates the sentence or wordswithout any intermediate representation.This is done by word to word translation by using bilingual dictionary followsthe syntactic rules.

b. INTERLINGUALMACHINE TRANSLATION:

In theInterlingualmachine translation, it transforms input text into a common representation with the help of

common independent representation;text can be generated in the target language.

c. TRANSFER BASED APPROACH:

Transfer based approach has three phases. Analysis phase, the input language sentence is parsed,and thenstructure and the constituents of the sentence are identified, sentence can be generated as parse tree form. In the Transfer phase,grammar rules are applied to parse tree which is generated from input to be converted into structure of output language. In generation phase, translation of words that are generated from parse tree and expresses the tense, number, gender etc.,

II-2 DICTIONARY BASED APPROACH:

In Dictionary based approach,ituses dictionaries for the language pair and it translates thetext from the input language to output language. In this approach word level translation will be done by using large number of dictionaries for storing all types of words.

II-3 EXAMPLE BASED APPROACH:

Example based approach is basedon solving the problemsand interpretion of humans. It requires large bilingual dictionariesof the language pair which is having the sentences in both languages. The main drawback of example based approach is it requires more depth of analysis.

In the past,Transfer based approach is used by different machine translation systems. In MANTRA machine translation system, the languages are used for the translation fromEnglish to Hindi. Itwas developed in the year 1997 and used byBharathi for information preservation. Further version of MANTRAmachine translationsystem, translates English to Hindi language developed in the year 1999 for the purpose of application proceedings in Rajyasabha. MAT developed a translation systemfor English languagetoKannada languagewhich was developed in the year 2002 by using morphologicalanalyzer and generator for Kannada language. In 2002, English language to Hindi language machine translation systemwas developed by using the Transferbased approach which mainly applicable to the weather narration. SHAKTHImachine translation system was developed fortranslations of English toIndian languages in the year 2003. MATRA Machine Translation system has developed a system by using Transfer based approach in the year 2004,2006. English language to Kannadamachine aided translation was developed in the year 2009 and is funded by Karnataka government. Punjabi to Hindi machine translationwhich was developed in the year 2007, 2008 which can be applicable andis used for general purpose.

III. KANNADA TO TELUGUMACHINE TRANSLATION SYSYTEM:

In this scheme, machine translation system is developed forKannada language to Telugulanguage by using Transfer based approach. In the existing literature, if the structure of both languages is similar,then it uses directmachine translation system. If the structure of both source and target languages are dissimilar, then it uses Transfer based approach. If the language structureis similar it did not usesdirectmachine translation approach, it uses Transfer based approach. Therefore, Transfer based approach is used for improving the performance of translationsystem.

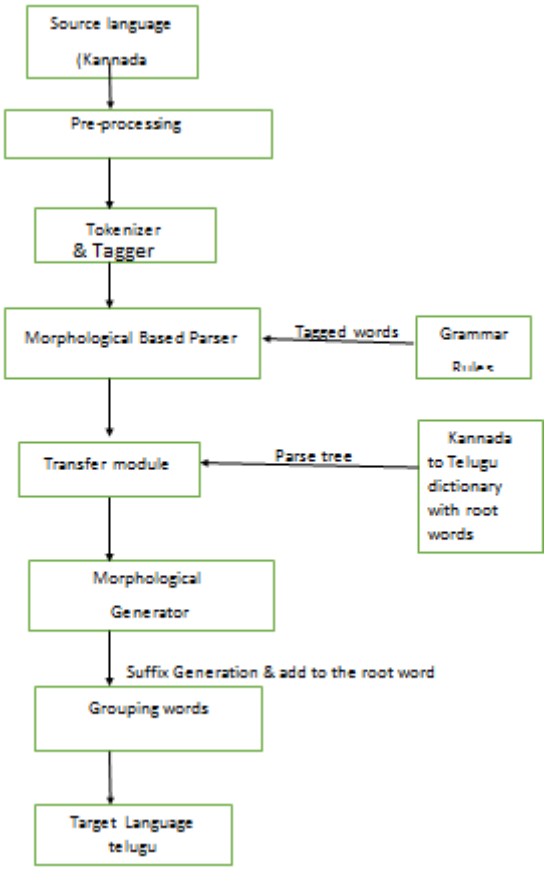

Fig:1 Block diagram forKannada to Telugu Translation by using Transfer Based Approach

III-1 PREPROCESSING:

In preprocessing phase, numbers of operation are applied to input data to make it possible by translation system. It includes treatment of punctuation; special charactersand transliterates the Kannada sentence into Romanized form.

III-2 TOKENIZATION:

Tokenizer is also known as lexical analyzer or word Segmenter. It takes the output of preprocessing phase as an input and segments a sentence into units known as tokens. In Tokenization phase, Kannada paragraph or sentence can be taken from the source file and it can be tokenizes the

sentence into wordsand for each word the root word are derived as shown in the below example.

Ex:Raamanu||Manega||Hoodanu

Raamanu| | raama
Manega| | mane
Hoodanu| |hoogu

III-3 TAGGING:

In tagging phase, tags must be assigned to words, thatwords can be tagged for each sentence and the output gives Kannada words with tagging.

NOUNS:
Lakshmanu||Lakshma||N-PRP-PER-M.SL-NOM
Raamanu||raama||N-PRP-PER-M.SL-NOM

The above sentence shows, the root word ofraamanu as raama. In most number of words V-IN-ABS is common so it did not classify that tag in the parser. Some tags are

N(NOUN)
- COM(Common) -PRP(proper)
-PER(personal)-LOC(Location)
-ORG (Organization)-OTH(others)
-LOC (Locative) -NOM(nominative)
-M (male)-SL (singular)
VERBS:
Banda| | V-PAST-P3-M.SL
Adalu| |V-IN-ABS-PRES-P3-M.SL

- IN (Intransitive) -TR(Transitive)
-BI (Bitransitive) -DEFE(defective)
-P1 (First person) -P2(Second person)
-P3 (Third person)-M(male)
-F (female) -SL(singular)
- PL(plural)

ADVERBS:
Tvaritavagiyu| | ADV-TIM
Aadudarimda| |ADV-CONJ
Mele| |ADV-PLA

-ADV (Adverbs)
-MAN (Manner)- NEG (Negative)
-CONJ (Conjunctive)-QW (Question Word)
-PLA (Place)-INTF (Intensifier)
-TIM (Time) -ABS (Absolute)
-POSN (Post-Nominal modifiers)

ADJECTIVES:
Prabala| |ADJ-ABS
Ivu| |ADJ-DEM

-ADJ (Adjective)
-ABS (Absolute) -DEM (Demonstrative)
-QNTF (Quantifying) -ORD (Ordinal)

III-4 MORPHOLOGICAL BASED PARSER:

In Morphologicalbased parser,the structure of words and parts of speech (POS) tagsfor given Kannada sentence are going to generate. When compare to the other parsers, it gives better results for generating the word with POS category. In Morphological based parser, tagged output is taken from the Tokenization and Tagging phase and this parser generates the parse tree for each tagged word using the Brute Force parsing mechanism from the grammar Rules and it gives the obtained output parse tree for each tagged word structure.

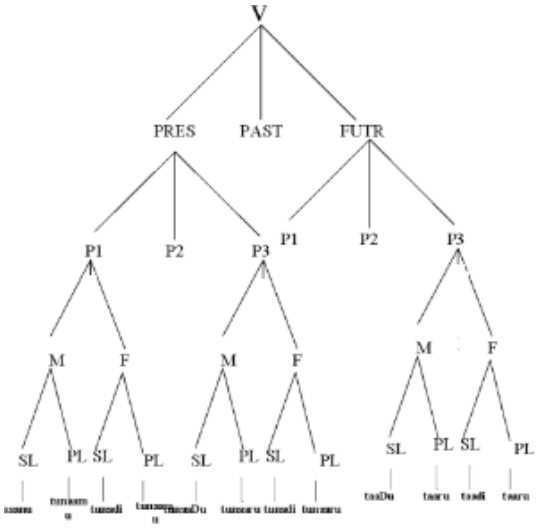

Fig:2 Parse Tree For Verb Structure

III-5 CROSS LINGUAL DICTIONARY:

Cross lingual dictionary contains the meaningsof root words for Kannada to Telugu languages and hasmost occurring root words of nouns, verbs, adverbs and so on. It has the two fields, one field is for Kannada root words and anotherfield forTelugu root words in the Romanized form for most common occurring verbs, nouns and so on.The bilingual dictionary is collected through various resources like Internet, books etc.,

TABLE 1: Cross Lingual Dictionary

Kannada root word	Translation of Telugu root word
niiDu	Iccu
Negu	GeMtu
Banda	Vaccu

III-6 TRANSFER MODULE:

Transfer module has three phases namely Analysis,Transfer and Generator phases.In the first phase, the input or source language is parsed the sentence structure and thenconstituents of the sentence is identified. In the Second phase, transformations are applied to the parse tree and in generation phase, itconvertsthe structure and generates the target language.

III-7 MORPHOLOGICAL GENERATOR:

Morphological generator indicates the generation of morphological words which is nothing but generation of Telugu words. It was developed using data driven approach and it has three modules. In the first module, it takes the input as POS and gives the output as lemmas paradigm number and word stem. The second module takes input as morph-lexical information and gives output as index number. In the last module, suffix table is used for generating the word with the information from the first and second modules.And then the suffix can be added to the root word. In the next step combining the all words which are generated from Morphologicalgeneration (Romanized words in Telugu). In the last phase, Romanized Telugu sentence can be taken as input and gives the output as exact Telugu sentence using TeluguSaara System.

IV .IMPLEMENTATION:

In implementation phase, input sentences are taken from a text file and output can be stored in another text file.The system can be implemented by using the python programming language and with the help of the Saara systems for getting the tagged words. And then, the text files is tested which contains Kannada sentenceand it translates into Telugu sentence.The system can be implemented by using the different modules.

a) Transliteration of Roman Kannada sentences.

b) Tokenization of sentences into words and words can be tagged.

c) Tagging of tokenized words generate parse tree.

d) Root words can be translated from Kannada to Telugu by using the bilingual dictionary.

e) Suffix which can be generated by the parse tree.It can be added to the rootword.

f) All the words can be grouped and translate sentence from Romanized Telugu to Telugu transliteration.

V. CONCLUSIONS:

There are several types of machine translation approaches are exist. In this paper,Transfer based approach has been selected for translation of sentences or words from Kannada to Telugulanguages and research mainly focused on tagging of adverbs and adjectives. It is found that accuracy has been improved by using Transfer based approach over other existing methods.

Transfer based approach can be extended to multilingual environment with more entries which gives the better performance.By using thisapproach,Kannada to Telugu sentences is tested for machine translation.

REFERENCES:

[1] G V Gharaje and G K kharate “Survey of Machine Translation System in India” International Journal on Natural Language Computing (IJLC) Vol. 2, No.4, October2013.
[2] Latha R. Nair and David Peter S “Machine Translation Systems for Indian Languages” International Journal of Computer Applications (0975 – 8887) Volume 39– No.1, February 2012.
[3] Kavi Narayana Murthy and Srinivasan Badugu “A New Approach to Tagging in Indian Languages” Research in computing 2013.
[4] Kavi Narayana Murthy and Srinivasan Badugu developed a paper on “Roman Transliteration for Indian scripts”
[5] Latha R Nair, David Peter & Renjith P Ravindran “Design and Development of a Malayalam to English Translator- A Transfer Based Approach” International Journal of Computational Linguistics (IJCL), Volume (3): Issue (1): 2012
[6] T.Suryakassnthi Research Scholar, and Dr. S.V.A.V. Prasad Translation of Pronominal Anaphora from English to Telugu Language *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013*.
[7] T. Venkateswara Prasad1, G. Mayil Muthukumaran2 Telugu to English Translation using Direct Machine Translation Approach International Journal of Science and Engineering Investigations vol. 2, issue 12, January 2013
[8] David Peter S. School of Engineering Cochin University of Science and Technology MachineTranslation Systems for Indian Languages International Journal of Computer Applications (0975 – 8887) Volume 39– No.1, February 2012.
[9] A Punjabi To Hindi Machine Translation System Gurpreet Singh Lehal by Professor, Dept. of Comp. Sci.,Punjabi University Patiala.
[10] Saara System is an integrated system that includes monolingual and bilingual dictionaries, stemmer, morphological analyzers and generators, etc., developed by Dr. Kavi Narayana Murthy at University of Hyderabad in Natural Language Engineering Lab.
[11] Machine Translation, Doug Arnold, University of Essex,doug@essex.ac.uk
[12] Natural language processing with python
[13] https://www.nltk.org
[14] https://www.google.co.in/
[15] http://en.wikipedia.org/wiki/
[16] Mallama V Reddy, DR. M. Hanumathappa “NLP challenges for Machine Translation from English to Indian Languages” International Journal of Computer Science and Informatics, ISSN (PRINT): 2231–5292, Volume-3, Issue-1, 2013
[17] Mantra Machine Translation System from English to Hindi which was developed by C-DAC