TRANSFER GRAMMAR COMPONENTS FOR MALYALAM TO HINDI AND ENGLISH MACHINE TRANSLATION SYSTEM

Stephy Joseph*[1], Shabina Bhaskar*[2] , Elizabeth Sherly*[3]

*[1]Indian Institute of Information Technology and Management – Kerala
stephyjoseph.mphilcs2@iiitmk.ac.in[1]
*[2] Indian Institute of Information Technology and Management –Kerala
shabinabaskar.mphilcs2@iiitmk.ac.in[2]
*[3]Indian Institute of Information Technology and Management - Kerala
sherly@iiitmk.ac.in[3]

Abstract: A study on divergence between or among languages plays a significant role to formulate a generic approach to Machine Translation (MT) in Indian Languages. Transfer Grammar acts as a bridging component between the source and target languages, through which various divergences between two languages can be analysed. In this paper, a word order based changes in Malayalam-English and Malayalam-Hindi is analysed and applied rule based and Machine Learning techniques to extract relevant data structures.

INTRODUCTION

Machine Translation (MT) is a thrust area of Computational Linguistics and it is the process of decoding any text from the source to the target language and then re-encoding it. Transfer based Machine Translation is now a days attaining greater attention as it defines some intermediate representation between two independent languages. A Transfer based MT mainly uses three different approaches such as Direct Transfer, Interlingua and Transfer Grammar (TG). Transfer Grammar (TG) consists of appropriate rules for transferring the structure of one language into corresponding sentence of another language when its structure is given. In this work we use Malayalam, a Dravidian language as source and Hindi and English as the target language.

MT mainly uses rule based or statistical approach and the proposed system is based on hybrid approach. The major steps in the development of this system include Tokenization, POS tagging and chunking, morphological analysis and transfer grammar generation. A hybrid approach based on rule based and statistical machine learning is proposed in this work. Rule based approach is performed on tokenization, morphological analysis, and transfer grammar generation. Statistical machine learning approaches is used in POS Tagging and chunking.

METHODOLOGIES

The divergence between two languages occurs at syntactical or lexical level and these divergences are handled by the Transfer Grammar. For that we have developed a data structure for the transfer grammar using rule based approach.

1) Data structure for proposed method

The data structure for the proposed system contains morphological, syntactical and semantic information.

Steps for Data structure Implementation

Step 1: read the input

Step2: split the sentence into tokens

Step3. Invoke POS tagger and chunker

Step3.1: Read the tokens label with translation tagset and chunck set.

Step 3.2: identify subject, object, masculine and feminine information from tagged result.

Step3.3: store the tagged and chunked tokens

Step4: invoke morphological analyzer

Step4.1: read the chunked tokens

Step4.2: identify the root from paradigm

Step 4.3: identify TAM and PNG information from the suffix analyzer

Step4.4: store the morph output

Step5: combine the information from the above steps and store it in a data structure.

Step6: generate transfer grammar rules.

Step7: combine the data structure and transfer grammar rules

Step8: word replacement and sentence replacement using cross-lingual dictionary.

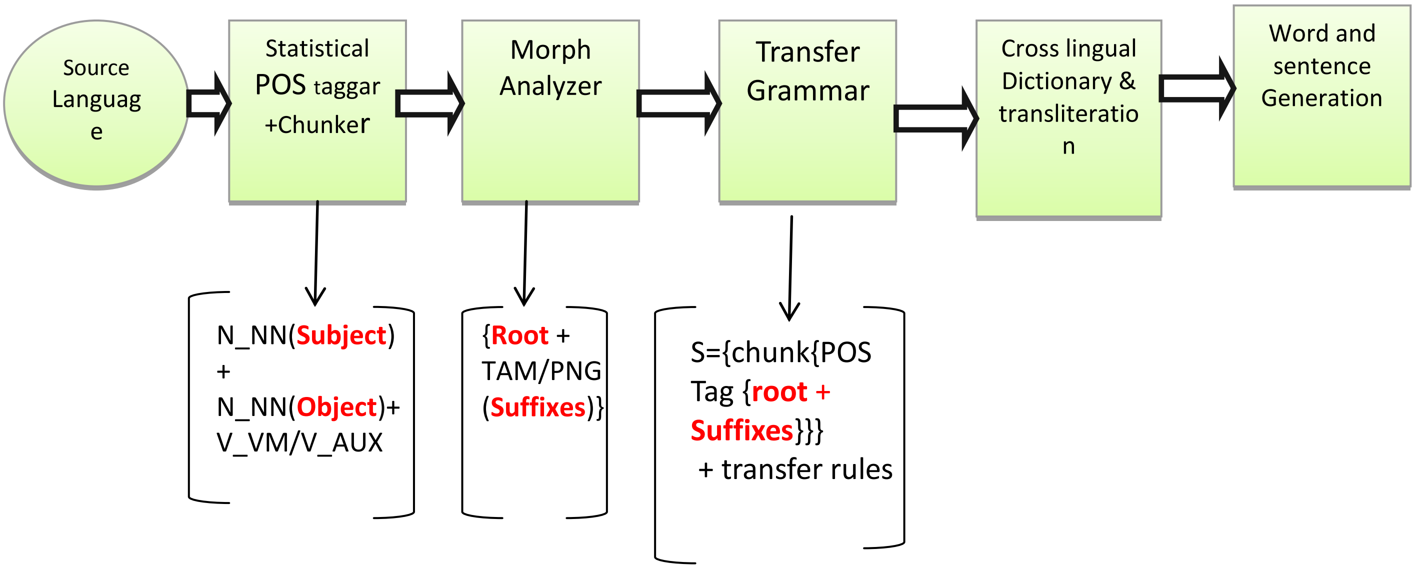
Source Languag e
Statistical POS taggar +Chunker
Morph Analyzer
Transfer Grammar
Cross lingual Dictionary & transliteratio n
Word and sentence Generation
N_NN(Subject) + N_NN(Object)+ V_VM/V_AUX
{Root + TAM/PNG (Suffixes)}
S={chunk{POS Tag {root + Suffixes}}}
+ transfer rules

Fig1: Data structure for the proposed method

2) Architecture of Malayalam to Hindi and English Machine translation system.

The proposed system architecture divided into three phases, source language, intermediate language and target language. In the first phase analysis of source language is done by using tokenization, morphological analysis, POS tagging and chunking. Identification of morphological properties is done by suffix stripping approach and paradigm approach. Also in Malayalam, nouns exhibit subject and object properties. So subject and object identification is very difficult in Malayalam. Instead of identifying a word as simply a noun or subject ,the subtypes in the different levels of hierarchy clearly determines the word type.ie, a noun which act as subject or object in a sentence. For that we used a special tagset called Translation tagset and performed POS tagging using SVM Tool. This SVM based Tagger is robust and flexible for feature modelling and is able to tag thousands of words per second.

In intermediate phase, developed transfer grammar rules which are required for mapping syntactic representations of language. In transliteration module, a word or words to be rendered in the script of the reader. And in the final phase corresponding Hindi sentence is generated and evaluated.

3) Transfer Grammar

In this module various types of transfers for Malayalam to Hindi and English translations are described. The following transfers are included in the transfer grammar module.
I)TG for Malayalam-Hindi translation

1) Subject - verb agreement in Hindi.

There is subject verb agreement in Hindi .i.e., In Hindi verb changes according to the subject.
1. Identify a subject is male or female.

a. If it is masculine then verb changed to root+आ .

b. If it is feminine then verb changed to root+ई

Example :राजिव को देखा

राजि को देखी

2) Copula transfer

It is verb or verb like word in Hindi copula verbs are expressed as possession, location and essential construct.

- ഇന്ത്യഎൻറെ രാജ്യമാണ്
- भारत मेरा देश है
- നിൻറെ പേര്എന്താണ്
- तुमारा नाम क्या हैं
- അവൻ വന്നിരുന്നു
- वह आया था
- നീ എന്താണ്ചെയ്യുന്നത്
- तुम क्या कर रहे हो

Copula corresponding to each noun is given below

Subject	Copula
मैं	हुँ
तु	है
तुम	हो
यह	है
हम , आप ,वे	हैं

3)Case transfers

In Malayalam case markers are agglutinated with nouns and pronouns. But in Hindi case markers are written as separate word with noun, and are suffixed with pronouns.
But there are some exceptions in nominative case, if the sentence is past tense and it is a sakarmaka verb then we have to add ने otherwise nothing is added. In genitive case it depends on the word that comes after the case marker. If it is singular and masculine then we add का its plural then, के and if it feminine word then we add ,के.
Following steps included in case transfer

1. Identify the subject in the sentence using tagging.
2. For each noun phrase in the sentence do the following.
   a. If the case is nominative and verb is past and sakarmaka verb then case is transferred into ने.
   b. Otherwise no change.
   c. If the noun is accusative or dative it is transferred into को
   d. If the case is instrumental then it is transferred into से
   e. If the case is genitive check the word that comes after the case marker.
      a. If it is singular and masculine then we add का its plural then ,के and
      b. If it feminine word then we add, के.
   f. If it is locative and it is ഇൽ then it is transferred into में otherwise पे.
   g. If it is sociative then it is transferred into को or से.

The case mapping between Malayalam and Hindi is presented in the table given below.

Cases	Malayalam	Hindi
Nominative	null	null,ने
Accusative	യെ,നെ	को
Instrumental	അൽ	से
Dative	ക്ക്,ന്	को
Genitive	ഉടെ,ന്റെ	का, के, की
Locative	ഇൽ,കൽ	में,पे
Sociative	ഓട്	को,से

4)Conjunction

ഉം– replaced with –और

Conjunction is used to connect two words or sentences. Here Conjunction ഉം– replaced with –और. Here the sentence form is noun/verb+ഉം +noun/verb+ഉം then it is replaced to noun/verb+ और+noun/verb.
Example.....

- രാമനുംസീതയുംമാങ്ങകഴിച്ചു
- राम और सीता आम खाया
- പുകവലിയുംമദ്യപാനവുംനിരോധിച്ചിരിക്കുന്നു
- धूम्रपान और शराब पीने निषिद्धहै

5 Handling of negative verbs
Negative verbs are which indicate that an action is not happen. In some cases when a negative Malayalam sentence is translated to Hindi reordering of words occurs. The following steps included in this.

1. Identify the type of negative verb.

a. if the negative verb is അല്ല,ഇല്ല then it is translated as नहीं.

b. if it is അരുത് then it is translated as मत.

c.if it is കഴിയില്ല then it is translated as नहीं सकते.

d. if it is വേണ്ട then it is translated as नहीं चाहिये
examples :

- അവൻവന്നില്ല.
- वहनहींआया
- നീമാങ്ങകഴിക്കരുത്.
- तूआममतखाओ

If it is verb + negative form in Malayalam then translated into negative + verb in Hindi. If Noun+ negative then it is noun + negative in Hindi.

II) TG for Malayalam-English translation

1)CASE OF CONNECTORS [ഉം- AND] ഉം – replaced with – AND

- രാമനും സീതയും മാങ്ങ കഴിച്ചു
- Raman and Seetha ate Mango.
- അവനും അവളും വന്നു.
- He and she came
- പുകവലിയും മദ്യപാനവും നിരോധിച്ചിരികുന്നു .
- Smoking and drinking is prohibited.

2) NEGATIVE WORDS. NOT ′ അല്ല,ഇല്ല replaced with NOT Example.....

- അവൻ വന്നില്ല
- He not came
- രാമൻ തിന്നില്ല
- Raman not Ate
- അവൻ മാങ്ങയല്ല കഴിച്ചത്
- He not ate mango

3)Copula Transfer : ആണ് replaced with IS

- ഇന്ത്യഎൻറെ രാജ്യമാണ്
- രാജ്യമാണ് = രാജ്യം + ആണ്
- India is my Country
- അവരുടെ വീട് ഭംഗിയുള്ളതാണ്
- ഭംഗിയുള്ളതാണ്= ഭംഗി + ഉള്ളത് + ആണ്
- Their house is pretty
- നിങ്ങളുടെ മകൻ നല്ലവനാണ്
- നല്ലവനാണ് = നല്ലവൻ + ആണ്
- Your son is good

4) HAVE/HAD

In the case of have - consider the gender feature Noun and pronoun

1. female [ക് + ഉണ്ട്]
2. Male [ന് + ഉണ്ട്] ′

- അവന് രണ്ട് മാങ്ങയുണ്ട്
- He have two mango
- സീതക് ഒരു പേ നയുണ്ട്
- Seetha have one pen
- രാമന് സഹോദരങ്ങളുണ്ട്
- Raman have brothers′
- എനിക് ഒരു തൊപ്പിയും കുടയും വേണം

  I want a hat and Umbrella
- നിങ്ങളുടെ മകൻ നല്ലവനാണ്
- Your son is good.

RESULTS

The accuracy of each stages calculated using the functional parameters such as precision and recall. The accuracy of each level in the Malayalam – Hindi machine translation system is calculated and the accuracy percentage is given below

	Precision (%)	**Recall (%)**
Tokenization	**87**	**82**
POS Tagger	**86**	**80**
Chunker	**86**	**91.5**
Morph Analyzer	**75**	**80.8**
Transfer Grammar	**70.1**	**72.2**
Word and Sentence Transfer	**75**	**79.4**

CONCLUSION

In this work we have explained various transfer grammars such as case transfer, handling of negative words and copula generation. The transfer grammar presented here captured the structural differences between source and target language in machine translation system. The proposed transfer grammar will be suitable for Malayalam to Hindi Machine translation system.To incorporate various other lexical transfers we have also include the statistical machine learning approaches in transfer grammar.

REFERENCES

[1] ShobhaLalitha Devi, SindhujaGopalan and Vijay Sundar Ram, *Transfer Grammar in Tamil-Hindi MT Systems*, International Conference on Asian Language Processing ,2013.

[2] V Goyal, G S Lehal, "Advances in Machine Translation Systems", Language In India, Vol. 9, No. 11, 2009, pp. 138-150 .

[3]Jisha P. Jayan, R.R. Rajeev,"Parts Of Speech Tagger and Chunker for Malayalam –Statistical Approach",Computer Engineering and Intelligent Systems ISSN

[4] Jisha P. Jayan, R.R. Rajeev and S. Rajendran, "Morphological Analyser for
Malayalam - A Comparison of Different Approaches", in Int. Journal of Computer Science and Informaiton Technology (IJCST), Vol. 2, No. 2, pp. 155-160, 2009.

[4]Saranya D Krishnan,Rajeev R R,Sherly Elizabeth,Mary Priya Sebastian,"Subject and Object Identification in Malayalam Text",ICACCI'12, 2012.