Vinod P M, Jayan V, Bhadran V. K.

DIVERGENCE PATTERNS IN HINDI - MALAYALAM MACHINE TRANSLATION

Vinod P M[*1], Jayan V[*2] and Bhadran V K[*3]

[1,2,3]Language Technology Centre, Centre for Development of Advanced Computing(C-DAC), Trivandrum, Kerala, India.
{ vinodpm,jayan,bhadran}@cdac.in

Abstract: This paper describes about some lexicon as well as structural mappings pertaining to Hindi-Malayalam machine translation (MT) system. The system is developed as part of the Hindi-Malayalam Speech to Speech (S2S) interactive Voice Response System (IVRS). Malayalam and Hindi are both Indian languages and their structure is almost similar to each other. In most cases translated sentence follows the same word order of the source sentence. But rarely, it is required to modify the structure of sentence in order to get proper target language output. Moreover some lexicon patterns need to get special considerations. Pattern Restructuring module introduced to handle and solve these dissimilarities in language family.

INTRODUCTION

Speech to speech Indian language to Indian language Machine Aided Translation(MAT) based dialogue system is an on going project in CDAC. This project is an Integration of Automatic Speech Recognition(ASR), MAT and Text To Speech(TTS) modules. The Hybrid approach will be used for translating the text generated through the ASR system. Tree Adjoining Grammar(TAG) based machine translation enriched with Semantic role labelling and Statistical approaches will be used.

TAG based tree-to-tree translation model makes use of syntactic tree for both the source and target language. As in the tree-to-string model, sets of operations apply, each with some probability, to transform one tree into another. Once the structure of the source language is transferred to a structure, which is close to the target language, sentence/word generator will generate the sentence in the target language. The Translation module is generalized for all languages. For each language pair derivational tree structure and transfer grammar can be added to this module. But these rules are much generalized ones. Here comes the importance of pattern restructuring module to modify the target language patterns which are having exceptions from the source language. Even though Malayalam and Hindi are having similar sentence structure, there are many exceptions due to their family of origin. Hindi belongs to Indo-Aryan and Malayalam is in Dravidian language family. Consider an example given below:

(1) I bought a new car that is very fast.

मैंने एक नयी कार खरीदी जो बहुत तेज है

വളരെ വേഗമേറിയ ഒരു പുതിയ കാർ ഞാൻ വാങ്ങി.

In the above example the structure of Malayalam and Hindi is different. When relative clause coming in the sentence the structure gets modified. The Hindi sentence follows the English pattern with structural changes in the phrases prior to the relative clause 'that' and Malayalam having different pattern as well as the structural changes which is not similar to Hindi.

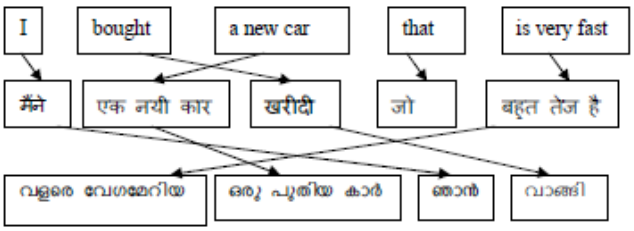

Figure 1. Phrase level mapping between translations.

Malayalam synthesizer module generates the final result by combing the suffixes in the appropriate word category based on sandhi rules in Malayalam.

The schematic diagram of translation system is shown below.

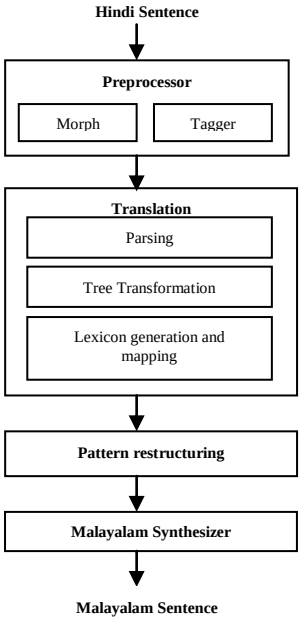

Figure 2. Schematic diagram of Translation System.

TRANSLATION DIVERGENCES

A translation divergence occurs when a translation of one language results in to a very different form than that of the original. There are syntactic and Lexical divergence types. Constituent order, Preposition stranding, Long distance movement, Null object and dative divergence types are the syntactic translation divergences that are accounted by means of syntactic parameterization. Conflational, structural, thematic, categorical, demotional, promotional, lexical divergence types are lexical semantic translational divergences that are accounted for by means of parameterization of the lexicon.

The classification captures the major grammatical issues in translation divergence across languages. However, it also misses a number of points pertaining to a particular set of translation languages. The issue of divergence between a set of languages is associated with many factors ranging from linguistic to socio - and psycho-linguistic aspects of the languages involved. Although Dorr's classification takes into account many of the major linguistic factors associated with translation divergence, there still remains a number of points related to both linguistic and extra-linguistic factors that may exist in different sets of translation languages. Furthermore the parameters of the classification does not take into account subtle semantic factors to the extent they are relevant for the classification of translation divergences in various languages. Without going into a detailed discussion of the different classes and categories of translation divergences as proposed in Dorr (1993, 1994), we discuss English and Malayalam translation examples that present new sources and topics of translation divergence in Hindi-Malayalam MT. We have examined some lexical semantic divergences between Hindi and its equivalent Malayalam and the same is described with the help of some examples. Also mentioned how the same affects during the translation process. In this paper we describe an alternative approach to MT that provides a general solution to the specific sub problem of cross linguistic divergences.

Dependency of Verb with Subject

In Hindi the main verb formation is depending up on two features, person and gender of the subject. In Malayalam neither the subject's gender nor the number influence the verb formation.

(2) वह पाठ पढ़ता है (Subject is male)
അവൻ പാഠം പഠിക്കുന്നു.
He is studying.

(3) वह पाठ पढ़ती है (Subject is female)
അവൾ പാഠം പഠിക്കുന്നു.
She is studying

Here sentence (2) says that the subject is male and which is expressed by its verb formation "पढ़ता है". While translating the target word "അവൻ" is opted with the help of above information. The lexicon generation module will take care of these things. In sentence (3) similarly the "पढ़ती है" verb formation directs to choose the target word "അവൾ".

(4) तुम पानी पीते हो (talking to a male)
നിങ്ങൾ വെള്ളം കുടിക്കുന്നു.
You are drinking water.

(5) तुम पानी पीती हो (talking to a female)
നിങ്ങൾ വെള്ളം കുടിക്കുന്നു.
You are drinking water

In sentences (4) and (5), the subject is "तुम" so the gender difference in the Hindi sentence is not required in Malayalam. Both of these sentences have the same target equivalent. The same is applicable when the subject is "मैं", "आप", "तू", "वे", "हम" and proper nouns. Some examples are given below to clear the idea.

(6) मैं अंग्रेज़ी बोलती हूँ (A speaker is female)
ഞാൻ ഇംഗ്ലീഷ് സംസാരിക്കുന്നു.
I am speaking English.

(7) मैं अंग्रेज़ी बोलता हूँ (A speaker is male)
ഞാൻ ഇംഗ്ലീഷ് സംസാരിക്കുന്നു.
I am speaking English.

(8) आप अंग्रेजी बोलते हैं (to a honourable male)
താങ്കൾ ഇംഗ്ലീഷ് സംസാരിക്കുന്നു.
You are speaking English.

If we observe these translations it is clear that the word order in both sentences (Hindi and its Malayalam translation) are same. No structural modification is required for this type of sentences. When we consider negation in this type of sentences we need a reordering of words.

Case Marker

Hindi is relatively free word-order language. Constituents can be moved around in the sentence without changing the core meaning. The case marker information will help to identify the subject and object (Ananthakrishnan et al., 2009). The sentences (9) and (10) convey the same meaning. Since Malayalam also shows the same property there is no issue in the translation process.

(9) जोन ने मेरी को देखा

ജോൺ മേരിയെ കണ്ടു

(10) मेरी को जोन ने देखा
മേരിയെ ജോൺ കണ്ടു

The word order in the source and translated sentences are same. The Malayalam Synthesizer module will take care of appending the case markers to the subject and object. While doing this process the sandhi rules are taken into consideration. The case marker "को" is mapped to "എ)". Then the process is as follows.

ജോൺ മേരി എ) കണ്ടു(Output of Pattern Restructuring module)
(മേരി + എ) → മേരിയെ) (Done by Malayalam Synthesizer)
ജോൺ മേരിയെ കണ്ടു (Final result)
John met Mary.

Hindi case marker details are listed below (Mohanan, 1994a).

a) Nominative
b) Ergative (ने)
c) accusative (को)
d) dative (को)
e) instrumental (से)
f) genitive (का)
g) locative1 (में)
h) locative2 (पर)

Each of these postpositions is invariable except for "का", which agrees with the possessed noun. Based on the number and gender "का", "के" and "की" is using. While translating to Malayalam it is mapped to "ൻറെ" or "ഉടെ" with respect to the mapped word's ending. This is handled by the Malayalam synthesizer module.

The possible mappings of the case markers are listed below.

ergative(ने)	null
accusative (को)	ൊ
Dative (को)	്
instrumental (से)	നിന്ന്, കൊണ്ട്
locative1 (में)	ൽ
locative2 (पर)	ൽ, പുറത്ത്

Table I. Mapping of Hindi case markers

Mapping the case marker to Malayalam is a tedious task. Malayalam case marker is very much depending upon the object part of the Hindi sentence. Some examples are shown below.

(11) बच्चे को दूध दो
കുഞ്ഞിന് പാൽ കൊടുക്കുവിൻ
Give milk to the child.

(12) चोर को भगाओ
കള്ളനെ ഓടിക്കുവിൻ
Make run the thief.

Sentences (11) and (12) shows that Hindi has the same marker for different case inflections (Sobha et al., 1999).

Pronoun Handling

Hindi pronoun, especially in the 3rd person level, expressing the number information only (Sobha et al., 1999). Whereas in Malayalam, 3rd level pronouns contains both the number and gender information.

Singular	Plural
वह वे (honorific)	वे

Table 2. 3rd person Hindi pronouns.

Singular	Plural
അവൻ അവൾ അത്	അവർ

Table 3. 3rd person Malayalam pronouns

By analyzing the verb part we can determine the exact Malayalam mapping of the 3rd person Hindi pronouns. The Hindi morph analyzer gives this information to the lexicon generation and mapping module. So that the module can generate the correct Malayalam 3rd person pronoun. Consider the case of reverse translation i.e. Malayalam to Hindi.

(13) അവൾ ഭക്ഷണം കഴിച്ചു.
वह खाना खायी
She had food.

In the Malayalam to Hindi translation case the gender information obtained from the subject is used to generate the Hindi verb. Sentence (12) shows an example for the same. Here the Hindi verb is generated based on the gender information obtained from the Malayalam pronoun.

Question Type Sentences

Question type sentences mainly sentences which contains "क्या", shows some structural modifications while translating to Malayalam.

Question sentence starting with "क्या"

(14) क्या इसके लिए कोई प्रवेश पास है
ഇതിനു പ്രവേശനപാസ് വല്ലതും ഉണ്ടോ?
Is there admission fee for this?

Question sentence contains "क्या"

(15) कहिये क्या जानकारी चाहिये
എന്തു വിവരമാണ് വേണ്ടതെന്ന് പറഞ്ഞാലും?
പറഞ്ഞാലും, എന്തു വിവരമാണ് വേണ്ടത്?
What information do you want?

The translations of sentences (14) and (15) show some structural modifications. The word order of the source and translated sentences are different. Hindi sentence (14) the first word "क्या" says that which is a question type sentence. But in its Malayalam translation the last word modifies the particular sentence to a question type. Rest of the sentence structure is same in both Hindi and Malayalam. Sentence (15) Hindi and its Malayalam translation show high degree of structural change.

Step 1: कहिये क्या जानकारी चाहिये
Step 2: क्या जानकारी चाहिये कहिये
Step 3: എന്ത് വിവരം വേണ്ടത് പറഞ്ഞാലും?
Step 4: എന്തു വിവരമാണ് വേണ്ടതെന്ന് പറഞ്ഞാലും
Tell, what information do you want?

Other question type sentences keeps the same sentence structure during the translation process and Sentence (16) is an example for the same.

(16) कौन सा रास्ता है, जहाँ से हम पहूँचेंगे
ഏത് വഴിയിൽ കൂടിയാണ് അവിടെ നമ്മൾ എത്തിച്ചേരുന്നത് ?

Conjunctions

The presence of some conjunctions in Hindi sentence also force to do structural modification in the translated sentence. Let us consider the case of "कि". The same conjunction acts differently in different sentences. One needs structural modification and the other does not.

(17) कहा जाता हैं कि सूरदास जन्म से अन्धे थे
സൂർദാസ് ജൻമനാ അന്ധനായിരുന്നെന്ന് പറയപ്പെടുന്നു
It is said that Surdas was blind by birth.

(18) आप चाय पीते हैं कि नहीं ?
താങ്കൾ ചായ കുടിക്കുന്നോ ഇല്ലയോ ?
Are you taking tea or not?

Consider sentence (17), it needs some structural modification in order to convey the meaning in Malayalam. But the sentence (18) follows the same structure in the source and translated sentences. While automating the translation this kind of behavior is very difficult to handle. To deal with this the Pattern restructuring module first check whether the sentence contains "कि " then check if it is a question or not. If it is a question then no structural modification is done, else go for structural modification.

The "और" and "एंव" conjunctions stands for "and" and the corresponding Malayalam mapping is "ഉം" for both. In Hindi these conjuncts are appears in between two words or between last two if more than two words. Whereas in Malayalam the "ഉം" conjunction is appending with each of the words. This case has to be handled in the Malayalam synthesizer module. Two examples are shown below.

(19) राम और सीता
രാമനും സീതയും
Ram and Sita

(20) राम सीता और लक्षमण
രാമനും സീതയും ലക്ഷ്മണനും
Ram, Sita and Laxman

Numbering System

Mapping of Hindi numbers, written in words, to Malayalam is a tedious task. Hindi is based on Vedic numbering system1 in which numbers over 9,999 are written in two-digit groups (or a mix of two- and three-digit groups) rather than the three-digit groups used in most other parts of the world. 1947 is written in both languages is shown below.

(21) उन्निस सौ सैंतालीस
ആയിരത്തിതൊള്ളായിരത്തി
നാൽപ്പത്തിയേഴ്

It is very difficult to map numbers written in Hindi to Malayalam and which is not included in this system.

Time

In Hindi and Malayalam people commonly follows the 12 hour system except for train and flight timings. In the case of 24 hour system it is very easy to be mapped to Malayalam. But in 12 hour system some additional modifiers are required to mention the correct time. Then we have to map these

modifiers correctly to express the time. The Hindi time modifies are listed below with its corresponding Malayalam mapping.

रात	രാത്രി
सुबह	രാവിലെ
दोपहर	ഉച്ചക്ക്
शाम	വൈകിട്ട്

Table 4. Time modifiers

Republication

In Hindi reduplication, both verbal and nominal, is important part of lexicon as well as grammatical structures. It belongs to the core of the language (Annie, 2009).

(22) बच्चों को एक-एक टोफी दो
കുട്ടികൾക്ക് ഓരോ മിഠായി വീതം നൽകൂ.

In sentence (22) "एक-एक" is a "numeral numeral" type reduplication. Here the meaning is 'to give a toffee to each child'. While translating to Malayalam it is required to add the meaning of "एक" before toffee and after toffee add the word "വീതം".

(23) तुम ने क्या-क्या देखा?
നീ എന്തൊക്കെ കണ്ടു ?
What all are you seen?

The sentence (23) shows an example of "क्या-क्या", and which belongs to noun duplication. This can be mapped as follows,

Meaning (क्या) + ഒക്കെ → എന്ത്+ഒക്കെ → എന്തൊക്കെ

The same method can be used for to solve noun and pronoun duplications like "कोन-कोन", "किस-किस", "जो-जो" etc.

(24) यहाँ महिलायें-महिलायें बैठेंगी
സ്ത്രീകൾ മാത്രം ഇവിടെ ഇരിക്കും
Only ladies should sit here.

Reduplication of plural noun is the other case. They used to express the meaning of exclusiveness or restrictiveness. Sentence (24) shows the example. Here we take the meaning of the plural noun and append "മാത്രം" to the meaning.

(25) खाते-खाते मत बोलो
കഴിച്ചുകൊണ്ട് സംസാരിക്കരുത്.
Don't talk while you eat.

(26) सोये-सोये मर गया
ഉറങ്ങിക്കൊണ്ട് മരിച്ചു.

Sentences (25) and (26) contain verb reduplication. In sentence (25) the reduplication expresses the meaning "while eating" and in the next sentence the reduplication means "in the sleep". The behavior is not predictable in these cases and so they are difficult to handle. In some other cases the reduplication reflects in the Malayalam sentence also. E.g. "करते-करते", "रोते-रोते" etc.

करते-करते → ചെയ്ത് ചെയ്ത്
रोते-रोते → കരഞ്ഞ് കരഞ്ഞ്

(27) उसके बाल काले-काले थे
അവളുടെ മുടി വളരെ കറുത്തതായിരുന്നു.

Example for adjective reduplication is shown in sentence (27). Here the reduplication expresses high degree of black color. In these situations "വളരെ", "നല്ല" etc. words are added in front of the Malayalam meaning.

Echo Words

Partial reduplication or echo constructions, formed with a "व" substitution to the initial consonant or with other forms of alliteration, shows that it modifies the notion itself by de-centering it, and reshapes it by taking into account various forms of heterogeneity, particularly the conflicting viewpoints of speaker and hearer (Annie, 2009).

An instance of the mere extension of the notional domain is the classical "चाय-वाय" (tea and other eatable and drinkable), "शादि-वादि" (marriage and so on), "पेन-वेन" (pen and the like) etc. These are identified as a phrase and directly replaced with the Malayalam meaning. And these meaning cannot be generated programmatically. In Malayalam there is no echo word formation as in Hindi. Normally such kind of words can be identified in the preprocessing stage and identify it as echo word and replace based on target language equivalent.

(28) तुम को पैसा-वैसा चाहिए क्या?
നിനക്ക് പൈസയോ മറ്റോ വേണമോ?

In this case the phrase 'पैसा-वैसा' have a target language equivalent is 'പൈസയോ മറ്റോ' which is not the source language equivalent in meaning. Any phrases like this can be replaced with first word followed by 'മറ്റോ'.

SYSTEM DESIGN

Only a subset of examples is mentioned in the above section. The current translation engine is not capable of handling all mentioned divergence patterns. The Pattern Restructuring module and synthesizer modules can handle some issues pretty well. These two are processing the translation module output. The pre processing information is also made available for these modules to perform well. Lexicon database also plays an important role in the system. The lexicon should be carefully added according to the category and with relevant feature information. So that it can be accessed and processed easily.

CONCLUSION

Lot of patterns and their issues are identified. Many of them are solved but still there are to be solved. The users expectation is very high always. They expects beautiful sentence with fluency. So the synthesizer module also needs to be improved to generate fluent Malayalam sentences. An example based translation module can help a lot to resolve some patterns.

ACKNOWLEDGMENTS

The authors thank all the members of S2S project who worked in its development processes, Dr. Hemant Darbari and the members of CDAC Pune who realized the Mantra System. The work is supported by the Ministry of Communication and Information Technology, Government of India, sponsored project.

REFERENCES

[1] Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya, "Case markers and morphology: addressing the crux of the fluency problem in English-Hindi smt" In Proceedings of ACL-IJCNLP,. 2009.

[2] B. Dorr., "Classification of Machine Translation Divergences and a Proposed Solution", Computat. Linguistics.20(4):597-633.DOI: http://aclweb.org/anthology-new/J/J94/J94-4004.pdf, 1994

[3] Jayan, V., R. Sunil, G. Sulochana Kurambath, and R. Ravindra Kumar, "Divergence patterns in machine translation between Malayalam and English", In Proceedings of the International Conference on Advances in Computing, Communications and Informatics, pp. 788-794. ACM, 2012.

[4] L. Sobha, B.N Patnaik, "VASISTH- An Anaphora Resolution System", Unpublished Doctoral dissertation. Mahatma Gandhi University,Kottayam,Kerala. (1999).

[5] Annie Montaut, "Reduplication and echo words in Hindi/Urdu", Annual Review of South Asian Languages and Linguistics , 21-91, 2009.

[6] Scott Gimm. "Subject Marking in Hindi/Urdu: A Study in Case and Agency", ESSLLI Student Session. Malaga, Spain, 2007.

[7] Spencer, Andrew "Case in Hindi. In Miriam Butt and Tracy H. King (eds.)", Proceedings of LFG '05. CSLI Publications, 429–446., 2005.

[8] V. Geethakumari, "A contrastive Analysis of Hindi and Malayalam, Langauge in India", Volume 2, Chapter 3, 2002.