Development of Malayalam Text Generator for translation from English

Abstract

The paper presents a strategy for developing Malayalam Text Generator for developing English Malayalam Machine Translation System using AnglaBharati technology developed by IIT Kanpur. AnglaBharati uses Interlingua approach for translation. It is a language independent, unambiguous representation of the input text. From this the text generator converts it in to the target language (Malayalam). In this paper we examine the major tasks in the development of Malayalam Text Generator for translation from English.

1. Introduction

Malayalam belongs to one of the four major languages of the Dravidian family. Morphologically, Malayalam is richly inflected by the addition of suffixes with the root/ stem word and Malayalam is agglutinative in nature. Malayalam uses a different word order as compared to English. Malayalam is a verb final language and all the noun phrases in the sentence normally appear to the left of the verb. Malayalam verb can be inflected to different forms. The inflection includes finite, infinite, adjectival, adverbial and conditional forms of words. These vary from one set of verbs to another. Malayalam has postpositions instead of prepositions and the adjectives and relative clauses precede their head nouns in a sentence. There is no one-to-one mapping between prepositions of English and the postposition in Malayalam. When suffixes or postpositions or any other base word itself is added to one base word, changes can occur to words. These changes are generally called sandhi changes. In English the two non-finite clauses are infinitive clause and gerund clause. The mapping patterns of these clauses in Malayalam are different depending on their syntactic functions and semantic roles. The other issues are in handling different types of sentences like interrogative or imperative sentences and mapping the patterns for pronouns. Figure: 1 illustrates the major tasks in the development of Malayalam Text Generator. All the blocks are mutually exhaustive.

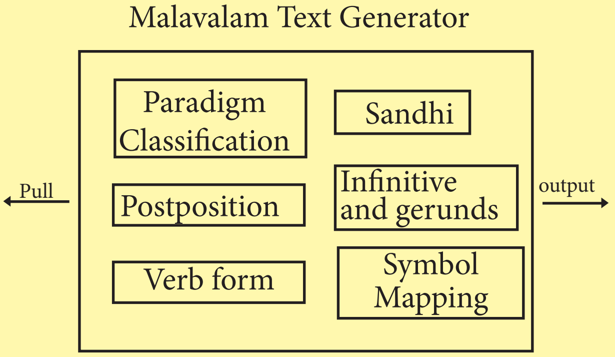

Figure 1 : Malayalam Text Generator

AnglaBharati technology uses a pseudo Interlingua rule based approach for Machine Translation. The input English sentences are transformed to an intermediate form called PLIL (pseudo lingua for Indian languages). The text generator will transform this PLIL in to the target language (Malayalam).

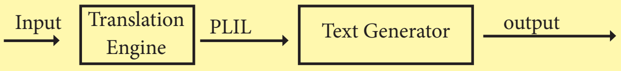

Figure 2: English to Malayalam Translation process

A sentence in PLIL is defined in terms of NP, VP and other constructs. In addition to that several keywords/terms are used to denote the nature of the sentence and other indicators that help to invoke the desired functions in the process of target language synthesize. Given below is an example of PLIL and its transformation.

(1) a. The prince came in front of the king.
PLIL:
<aff {sub_np (the det [] [anda] [A]) (prince noun dont_care singular third [human] [rAjakumAran_:4] [] []) } {pp (the det [] [anda] [A]) (king noun dont_care singular third [post_holder] [rAjAv:3] [] []) (infrontof prep [infrontof]) } {main_vp_active (come verb_3 normal normal dont_care singular third [var] 27 [] []) } > . sviram

Malayalam Output in Roman Notation:
rAjakumAran_ rAjAvinZemunnil_ vannu.
(prince) (king)(in front of) (come PAST)

(1) b. He walked along the street.
PLIL:
<aff {sub_np (he noun masculine singular third [human] [avan_:p 2] [] []) } {pp (the det [] [anda] [A]) (street noun neuter singular third [place] [weruv:10] [] []) (along prep [along]) } {main_vp_active (walk_1 verb_3 normal normal masculine singular third [nata] 16 [] []) } > . sviram
Malayalam Output in Roman Notation:
avan_ weruvilUte natannu.
(He) (street along) (walk PAST)

We can see the syntactic and semantic information's in the PLIL. 'aff' indicates it is an affirmative sentence. Also *'(street noun neuter singular third [place] [weruv:10] [] []) '* indicates the parts of speech of the word with its gender, number, person, semantic category, Malayalam meaning along with the paradigm.

2. Malayalam Text Generator

As pointed out, the PLIL or intermediate representation contains the syntactic and semantic information. The Text Generator takes PLIL as input and will perform morphophonemic additions required in the root word based on the syntactic information present in the PLIL structure. In Malayalam Text Generator, we give description of rules of Malayalam grammar which are relevant for translation. To solve the complexity in the inflections of verb, classification of Malayalam verbs are done. Similarly for other categories noun, pronoun and adjective the paradigm classifications are made. Preposition in English will be changed to postpositions in Malayalam. They are written after the noun. In section 2.2 we discuss some common patterns for disambiguation of postpositions in Malayalam. The changes called sandhi changes can occur, when suffixes or postpositions or any other base word itself is added to one base word. The different types of sandhi are described in section 2.3. In section 2.4 we will discuss with few examples on synthesizing verb forms. The two non-finite verb forms in English are: infinitive and gerund. They express a wide range of functions. In section 2.5, the rules that can be used to determine the mapping pattern of infinitive and gerund in Malayalam are illustrated. The mapping patterns of symbols or keywords present in PLIL, mainly the relative pronouns are described in section 2.6.

2.1 Paradigm classification

A paradigm defines all the word form of a given root/stem and also provides a feature structure with every word form.

2.1.1 Paradigm classification for noun.

The classification of Malayalam nouns are based on three features namely, number markers, case markers and word endings. The number markers (singular or plural markers) each with seven cases (case markers) are combined together to form the different generated forms and paradigm numbers. Let us consider one example "kutti", the singular and plural form of the noun "kutti" {(kutti - singular), (kuttikaL_ - plural)} will be combined to the seven case features. The seven case markers with suffix in Malayalam are:

1) Nominative case (kutti = Φ)
2) Accusative case (kuttiye = e)
3) Dative case (kuttikk = kk)
4) Sociative case (kuttiyOt = Ot)
5) Locative case (kuttiyil_ = il_)
6) Instrumental case (kuttiyAl_ = Al_)
7) Genitive case (kuttiyude = ute)

The variant forms of plural marker {kaL_ (kuttikaL_), mAZ_ (rAjakkan_mAZ_) and aZ_ (manuRyaZ_)} will mostly have a common case forms. For place names and proper names the word endings can also be considered for classification. The following example shows the various forms of a paradigm.

(3)

kutti, kuttikaL_, kuttiye, kuttikaLe, kuttiykk, kuttikaL_kk, kuttiyOt, kuttikaLOt, kuttiyil_, kuttikaLil_, kuttiyAl_, kuttikaLAl_, kuttiyute and kuttikaLute

2.1.2 Paradigm classification for pronoun.

A pronoun is a kind of noun, but it function is different from noun. A pronoun is a word which refers to a person or a thing that has already talked about. Pronoun can take the place of a noun in a sentence and function as a noun. The various word formation of the word 'avaZ_' based on case is shown below:

(4)

avaZ_, avare, avarOt, avarAl_, avaZ_kk, avarute and avaril_

2.1.3 Paradigm classification for adjectives.

Adjectives in Malayalam language are generally derived from either noun or verb by the process of suffixation. Most adjectives can occur both before and after a noun. Adjective before the noun is called attributive adjective and the adjective after the noun is called predicative adjective. Most of the adjectival forms in Malayalam will alter due to the influence of its gender and number distinction. The semantic information of head noun will also have a major role in the formation of adjectives in Malayalam. The following shows the various forms of a paradigm.

(5)

prasixXanAya, prasixXayAya, prasixXan_mArAya, prasixXakaLaya, prasixXamAya, prasixXarAya, prasixXan_, prasixXa, prasixXan_mAZ_, prasixXakaL_, prasixXaM and prasixXaZ_.

2.1.4 Paradigm classification for verbs.

The Malayalam verb morphology is very complex. Dr.A.R. Raja Raja Varma (AR) has listed and classified verbs into different classes in his monumental work Keralapaniniyam.

Dr.Suranad Kunjan Pillai (SKP) has classified Malayalam verbs into different classes based on the verb endings in its future form. We have classified verbs based on the works of AR and SKP for computational purpose. The following example shows how the various formations of a verb affected with the help of paradigm. Any verb having the similar suffixation will fall in this category and will assign the paradigm number as that of the verb root *akal*.

(6)

Root : *akal*

a) The three tense forms
   akalunnu, akaluM and akannu.

b) Transitive forms
   akaZZunnu, akaZZuM and akaZZi

c) Intransitive forms
   akalunnu, akaluM and akannu

d) Causative forms
   akaZZikunnu, akaZZikkuM and akaZZiccu
   akaZZippikkunu, akaZZippikkuM and akaZZippiccu

e) Noun generating form
   Akal_cca

2.2 Disambiguation of postpositions

Postpositions in Malayalam are certain forms, which occur immediately after nouns and establish some grammatical relations between the nouns and the verbs of the sentences. The semantic distribution of a single preposition will be varying in different context due to the influence of nouns and main verbs that follow. For instance, a preposition *to* can have multiple mapping patterns in Malayalam.

(7) [to = Ekk]

The procession goes to Kottayam.
jAWa kOttayawwEkk pOkunnu.
(procession) (kottayam PP) (go PR CONT)

(8) [to = Ot]
I have spoken to him already.
FAn_ iwinakaM wanne avanOt saMsAriccittuNt.
(I) (already) (him SOC) (speak PR PERF)

(9) [to = atuwwEkk]
I am going to the king.
FAn_ rAjAvinZe atuwwEkk pOkukayAN.
(I am) (king GEN) (PP) (go IMPERF PR)

(10) [to = e]
Please listen to him.
xayavAyi avane SraxXiykkU.
(Please) (him ACC) (listen FUT)

(11) [to = kk]
He is going to the meeting.
avan_ kUtikkAYcaykk pOkukayAN.
(He) (meeting DAT) (go IMPERF PR)

We can notice that the mapping of preposition varies in different context. In (7) "to" is mapped into "*Ekk*" and formed as "*kOttayawwEkk*". Where as in (8) "to" is mapped into sociative case "*Ot*" and combined with the noun as "*avanOt*". The mapping pattern of "to" in (9) is "*atuwwEkk*", where as in (10) accusative case is added to the noun to form "*avane*". In (11) dative case '*kk*' is added to the noun to form "*kUtikkAYcaykk*". From the above given examples we can see that the two major factors that determine the meaning of a preposition are the semantic type of the main verb and the nominal elements that occur with the prepositions. Rules that are used to disambiguate prepositions are illustrated for 'to'. We use the semantic category of verb and noun to disambiguate the multiple patterns of the prepositions.

a. to-NP (place) = NP- Ekk
E.g. (7)
b. to- NP(human) = NP-Ot
E.g. (8)
c. verb(motion) to-NP = NP- atuwwEkk
E.g. (9)
d. verb(mental) to-NP = NP-e
E.g. (10)
e. to- NP (activity/concept) = NP-kk
E.g. (11)

2.3 Sandhi

The general meaning of the term sandhi is union. We can group sandhi into vowel sandhi, vowel consonant sandhi, consonant vowel sandhi and consonant sandhi.

(12) Vowel Sandhi (Vowel+Vowel)
mala + alla => malayalla

(13)Vowel Consonant Sandhi (Vowel +Consonant)
wAmara + kuLaM => wAmarakkuLaM

(14) Consonant Vowel Sandhi (Consonant +Vowel)
minnunna + AkASaM => minnunnAkASaM

(15) Consonant Sandhi (Consonant + Consonant)
nel_ + maNi => nen_maNi

In (12), we can see that while combining the words, '*y*' has been generated to form '*malayalla*'. Similarly in (13) a sound '*k*' has been geminated to form '*wAmarakkuLaM*'. Where as in (14) '*a*' is deleted while combining the two words to form *minnunnAkASaM*. And in (15) chillu '*l_*' is displaced by another chillu '*n_*' resulting in the word *'nen_maNi'*. So the Sandhi can also be classified on the basis of lopa (elision), agama (addition), dvitva (gemination) and adesha (displacement). Examples of these classifications are given below.

(16) Lopa Sandhi
In this case one character is deleted while sandhi formation. In the given example the 'a' at the end of first word is deleted and the first word is combined with the second word.
oYukkunna+aruvi => oYukunnaruvi

(17) Agama Sandhi
This sandhi have the property of evolution of new character while sandhi. Here the consonant 'v' is generated while combining the vowels 'u' and 'O'
wiru+ONaM => wiruvONaM

(18) Dvitva Sandhi
During the sandhi formation one of the character geminates then it is Dvitva sandhi. The character 'p' is geminated in the example
avite + pOyi => aviteppOyi

(19) Adesha Sandhi
In this case one character is replaced with another character while sandhi formation. The character 'M' is replaced with 'F'
maraM+cAti => maraFcAti

2.4 Synthesizing verb form

The verb will take different forms according to tense, aspect and modality. These factors express wide range of functions. For example, consider the following sentences.

(20) a. He went to the market.
avan_ canwayilEkk *pOyi*.
(He) (market PP) (go PAST)

(20) b. He is going to the market.
avan_ canwayilEkk *pOkukayAN*.
(He) (market PP) (go IMPERF PR)

(20) c. They want to go.
avaZ_ *pOkAn_* Agrahikkunnu.
(they) (go INF) (want PR CONT)

(20) d. Can i go?
enikk *pOkuvAn_* sAXikkumO ?
(i) (go INF) (can QP)

(20) e. I got my jacket cleaned.
enikk enZe jAkkeZZ vqwwiyAkki *kitti*.
(I DAT) (my ACC) (jacket) (clean PAST) (get PAST)

(20) f. she made her children do their homework.
avaL_ avaLute kuttikaLe avarute gqhapATaM *ceyyiccu*.
(she) (her ACC) (children ACC) (their ACC) (homework) (make CAUS)

We notice that in (20 a) and (20 b), the difference in the formation of the verb. Here the type of the verb has been used for deciding the correct form. However, in (20 c) the type of the verb and the pattern type 'non-finite' are used to determine the desired formation. In (20 d) the type of the verb, pattern type and the auxiliary has to be considered for its correct formation. However in (20 e) and (20 f) causative verb (get, make) are taken into consideration for appropriate formation.

2.5 Resolution of infinitive and gerund

Infinitive clause and gerund clause are the two non-finite clauses in English. In infinitive clauses the main verb occurs in its root form preceded by to whereas in gerunds it has verb-ing form. Let us go through few example patterns and its rules to determine the mapping pattern of infinitive and gerund in Malayalam.

P1: If the verb has auxiliary as "to" and type of the verb in present plural form:

Rule 1: If the verb comes under the infinitive phrase in PLIL, infinitive verb forms as
Verb root + Link Morph (if any) + An_

(21) a. I like to play football.
FAn_ PutbOL_ kaLikkAn_ IRtappetunnu.
(I) (football) (play INF) (like PR)

(21) b. I like to sing a song.
FAn_ oru pAtt *pAtAn_* IRtappetunnu.
(I) (a) (song) (sing INF) (like PR)

Here we can see that in (21 a) the link morph "*ikk*" and the suffix '*An_*' is added to the root word to form the infinitive verb as "*kaLikkAn_*". Where as in (21 b), only the suffix '*An_*' is added to form the infinitive verb as "*pAtAn_*".

Rule 2: If the verb comes under the subject phrase in PLIL, infinitive verb forms as
Verb Root + Link Morph (if any) + uka

(21) c. To respect our parents is our duty.
nammute mAwApiwAkkan_mAre
Axarikkuka nammute kaZ_wwavyaM AN.
(our ACC) (parents ACC) (respect PR

INF) (our ACC) (duty) (be PR)

(21) d. To err is human.
weZZ paZZuka mAnuRikaM AN.
(err PR INF) (human) (be PR)

We can notice that in (21 c), the link morph "*ikk*" and the suffix "*uka*" are added to the root word to form "*Axarikkuka*" and in (21 d) the suffix "*uka*" is added to form "*weZZ paZZuka*".

P2: If the verb has auxiliary as "to" and type of the verb in past participle form:

Rule 3: If the verb comes under the infinitive phrase in PLIL, infinitive verb forms as
Verb Root + Link Morph (if any) + *ENta*

(21) e. This is a book to be read.
iw oru vAyikkENta puswakaM An.
(this) (a) (read INF) (book) (be)

(21) f. This is a song to be sung.
iw oru pAtENta gAnaM An.
(this) (a) (sing INF) (song) (be)

Here in (21 e), the link morh "ikk" and the suffix "*ENta*" are added to the root word to form "*vAyikkENta*". And in (21 f), the suffix "*ENta*" is added to the root word to form "*pAtENta*".

Rule 4: If the verb comes under infinitive phrase, and the sentence is in active voice, infinite verb forms as

Verb Root + Link Morph (if any) + appetAn_
(21) g. He wants to be loved.
avan_ snEhikkappetAn_
Agrahikkunnu.
(He) (love INF) (want PR)

Here we can notice that in (21 g), the link morph "ikk" and the suffix "*appetAn_*" are added to the root word to form "*snEhikkappetAn_*".

(22) a. The old man was tired of walking.
muwiZ_nna manuRyan_ *natakkunnawinAl_*
kRINiccaw Ayirunnu.
(old) (man) (walking INF) (tired) (be)

(22) b. He is fond of hoarding money.
avan_ paNaM *SEKariccu vaykkunnawil_*
wAwparyamuLLavan_ AN.
(He) (money) (hoarding) (fond) (be)

Similarly for gerunds, like infinitive patterns (P1 and P2), for (22 a) and (22 b) similar approach can be used to map the appropriate patterns. These are some of the rules that can be used to determine the mapping pattern of infinitive and gerunds in Malayalam.

2.6 Symbol Mapping

The symbols/keywords like karak, relative pronoun should be mapped to its correct form. The karak (case) that should be added to the noun is determined by considering the semantic type of the noun and the type of the verb it follows. For example:

(23) a. I told him.
FAn_ avan*Ot* paZaFFu.
(I) (he SOC) (tell PAST)

(23) b. I need you.

enikk niffaL*e* AvaSyamuNt.
(I DAT) (you ACC) (need PR PERF)

In (23 a), the communicative verb ‘tell’ is used to determine the appropriate case ‘sociative’ that should be added to the noun *‘avan_’* to form ‘*avanOt*’. Whereas in (23 b), the accusative form has to be added to form ‘*niffaLe*’.

In the resolution of relative pronouns, it can be mapped in Malayalam in multiple ways. For illustration, consider the set of sentences given below:

(24) a. I am feeling regret for what i did.
FAn_ ceyw*awil_* enikk paScAwwApaM wOnnunnu.
(I) (did) (I DAT) (regret) (feel PR)

(24) b. Carefully listen to what i say.
SraxXayOte FAn_ paZayunn*aw* SraxXiykkU.
(Carefully) (i) (say PAST) (listen FUT)

(24) c. I am the person that is to blame.
KuZZappetuwwENt*awAya* vyakwi FAn_ AN.
(blame) (person) (I am) (be PR)

(24) d. It is dogs that bark.
kuraykkunn*aw* pattikaL_ AN.
(bark PAST) (dogs) (be PR)

Here in (24 a), the mapping pattern of ‘what’ is *‘awil_’* in the verb form *‘ceywawil_’*, where as in (24 b) it is ‘*aw*’ in the verb form *‘paZayunnaw’*. Similarly we can see the different mappings of ‘that’ “*awAya*” in (24 c) and “*aw*” in (24 d).

An imperative sentence is a sentence which gives a command or makes a request. For example,

(25) Go to your room.
niffaLute muZiyilEkk pOkU
(you ACC) (room to) (go FUT)

Here the type of the sentence and the pattern type as “command” are used to get the appropriate meaning.

A sentence that asks a question is called an interrogative sentence. They may ask for information or for confirmation or denial of a statement.

(26) a. Is he working?
avan_ jOli ceyyukayANO?
(he) (work) (do be QP)

(26) b. Is he not working?
avan_ jOli ceyyunnillayO?
(he) (work)(do be NEG)

From these examples we can notice that “not” in the negative interrogative sentence (26 b) made the difference in the translation. Here the suffix “*illa*” has to be added to form “*jOli ceyyunnillayO*”.

The disambiguation strategies discussed above are based on the syntactic information present in the PLIL structure.

3. Summary

In this paper we have discussed some of the major tasks involved in the Malayalam text generator for translation from English to Malayalam. We have seen the paradigm classification of various categories. On the basis of the semantic category of preceding and following nouns of the preposition and the category of verb, we have made an attempt to disambiguate the multiple meanings of selected prepositions. We have seen the basic rules for combining the connectable words. Synthesis of the verb and to determine the mapping pattern of infinitive and gerund in Malayalam are illustrated in detail. Some of the mapping patterns of symbols/keywords that help to invoke the desired functions in the process of text generation are also gone through.

The system can handle simple sentences with considerably good accuracy rate. More research is needed to explore language rules to handle all types of sentence patterns for translation.

Abbreviations/ Acronyms

cat: Category, **main**: Main verb, **mot**: Motion verb, **mnt**: mental, **Noun1**: Noun after preposition, **Noun2**: Noun after preposition, **pst_hldr**: post_holder, **PROG** –Progressive, **QP** – Question Particle, **PAST**- Past tense, **PP** – Postposition, **PASS**-Passive, **INST** – Instrumental, **NEG** – Negative, **FUT** – Future, **CONT**- Continuous, **CONDT**-Conditional, **CAUS**- Causative, **LOC**- Locative, **INF**- Infinitive, **IMPERF** – Imperfective, **ACC**- Accusative

References

1. R.M.K. Sinha, An Engineering Perspective of Machine Translation: AnglaBharti-II and AnuBharti-II Architectures, Proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS- 2004), November 17-19, 2004, Tata Mc Graw Hill, New Delhi.
2. R. M. K. Sinha, Machine Translation: AnglaBharati and AnuBharati Approaches, Communications of CSI, October 2005.
3. R.M.K. Sinha, A Pseudo Lingua for Indian Languages (PLIL) for Translation from English. Technical Report, Language Technology Lab, Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur (2004).
4. R.E. Asher, T.C. Kumari, Malayalam - Descriptive grammars, 317-319, Routledge(2007)
5. A R Rajaraja Varma, Keralapaniniyam, Pages 177-269, DC Books(2000)
6. Suranad Kunjan Pillai, Volume 1, Malayalam Lexicon, Appendix(1-105), Introduction xviii The University of Kerala(2000)