Divergence Patterns in Machine Translation between Malayalam and English

Abstract

The translation divergence is the problem pertained with the Transfer and Interlingual Machine Translation (MT) because it requires a large combination of complex lexical and structural mappings. Divergence makes straight forward transfer from source language to target language impractical. The main aim of the divergence study is to know the source and target language pattern and morphophonemic changes upon translation of a sentence. This will enable the MT system to handle most of the anomalies coming out of divergence problem. For this we need a systematic mapping between the interlingual representation and the surface syntactic structure that accomodates all of the divergences. In this paper we try to list out different types of divergence occuring among English and Malayalam.

Introduction

The process of parallel text exploitation to extract transfer mappings between language pairs raises the capability of language translation. However, while this process can be fully automated, one thorny problem called "divergence" causes indisposed mapping extraction. A class of system analyzes sentences in parallel texts to obtain transfer structures or rules. The process of parallel text exploitation to extract transfer mappings (models or rules) between language pairs can raise the capability of language translation. It is still hampered by the difficulties in transfer mapping extraction of achieving accurate lexical alignment and acquiring reusable structural correspondences. Although automatic extraction methods of lexical alignment and structural correspondences are introduced, they are not capable of handling exceptional cases like "divergence".

1.Non-Configurational Nature of Malayalam

English is a configurational language that follows a rigid word order pattern as opposed to Malayalam which is relatively less rigid and exhibit free word order variation. This is one of the major sources of divergence between a pair of natural languages. In Dorr's classification, word order related translation divergences have been discussed under syntactic divergence. Malayalam verbs do not have the gender information unlike other Dravidian languages. For instance, one of the implications of the word order related divergence can be noticed with respect to the interpretation of the question particle *'ANO'* in Malayalam.

(1)Are you reading?
   wAfkaL_ vAyiccukoNtirikkukayANO?
   (You)   (read PROG be QP)

Malayalam, like most of the other South Asian languages, exhibits the phenomenon of replication of the lexical items to express different grammatical functions. The English counterparts of these Malayalam constructions do not resort to replicative structure. This distinction often results into a change in the syntactic category of the relevant elements. For instance, we notice that in Malayalam, as in (2),

the replication of the verb (in participial form) denote an adverbial function of cause. The English counterpart of this function is realized by a gerundive prepositional phrase. Malayalam is also having the replicative nature in lexical item. On reverse translation replication do not occur. Example given below:

(2)avan_ saMsAricc saMsAricc waLaZ_nnu pOyi
  (he) (speak) (speak) (tired PAST) (got)
  He got tired of speaking.

Reverse translation don't have divergence here. 'saMsAricc saMsAricc' is an adverbial clause that is not replicated in the reverse translation.

avan_ saMsAricc waLaZ_nnu pOyi
(he) (speak) (tired PAST) (got)

Here another factor that have divergence. For the phrase 'nOkkiyuM kaNtuM' have actual meaning in the sentence is 'be careful' but word meaning is 'look and see'. The reverse translation is entirely different as in the idioms.

(3)nOkkiyuM kaNtuM irunnAl_ ninakku koLLAM
  (look) (see) (be) (you DAT) (better)
  It is better if you are careful.
  SraxXiccirunnAl_ ninakku nallaw
  (careful Be) (you DAT) (better)

In normal conversation we use the term "maZZO" even if we don't indicate directly or indirectly some other things. 'maZZO' is having the independent meaning 'something else'.
(4)ninakk pQsayO maZZO vENO?
  (you) (money) (EW) (want QP)

Do you want money?
ninakk pQsa vENO?
(you DAT) (money) (want QP)

Expressive Elements

Expressive words exist in all natural languages and pose difficulty in processing, particularly in mapping onto another language. The reason is that these words do not have exact parallel in another language. Thus the word *powwO* is only distantly mapped by '*bump*' in English, as in (5). We use the expressive words in a sentence to convey the actual impact of a particular action performed by a *subject*. In the example the word 'powwO' gives a visual impact on the reader about the action. Getting exact translation for such words are difficult.
(5) She fell with a 'bump'.
  avaL_ powwO ennu vINu.
  (She) (bump) (PP) (fall PAST)
The expressive words usually originate from the sound associated with the semantics of the action verb and can be adverbial or verbalized action-verbs such as 'pata-pata' (crackle),'kilu-kilu (clink) etc.
(6)Engine made a crackling sound.
  eFcin_ patapata SabxamuNtAkki
  (Engine) (crackle) (sound make PAST)

Asymmetry in NP and Existential Clauses

English has (in) definite articles that mark the (in) definiteness of the noun phrase overtly whereas Malayalam lacks an overt article system and different devices are used to realize the (in) definiteness of a noun phrase in Malayalam. For instance, mapping onto articles *a-an/ the* in English is not lexically realizable from Malayalam. Eg:

(7) The boy is coming.
   AN_ kutti varunnu.
   (boy)   (come be PR)

In this connection, another point of divergence between Malayalam and English related to *there*- and *it*-sentences in English is worth examining. In English, *there*- and *it*-constructions are used to denote existential sentences (besides others). Malayalam does not have a pleonastic subject construction and the contrast between existential and non-existential (mostly definite) sentences is realized by several other ways such as the movement of the noun phrase from its canonical position and the use of demonstrative elements. Let us look at the examples given below:

(8) muZiyil_ pAmp uNt.
   (room)   (snake) (be)
   There is a snake in the room.

(9) pAmp muZiyilAN.
   (snake) (room be)
   The snake is in the room.

(10) There is a snake in the room
   muZiyil_ oru pAmp uNt.
   (room)   (a) (snake) (be)

Tense, Moods and Aspects (TAM)

Another important source of translation divergence in Malayalam and English MT is associated with the difference in the manifestation of different tense, moods and aspectual properties of the verb in these languages. For instance, Malayalam uses a certain type of passive construction that marks a kind of non-volition function. The English counterparts of such Malayalam sentences are only partially able to express the exact meaning.

(11) **rAmanAl_** weZZu saMBaviccu
   (Ram INST) (mistake) (be PAST)
   Ram made a mistake.
   rAman_ weZZu   ceywu.
   (Ram)   (mistake) (do)

The possible English counterpart of the Malayalam sentence in (11) is far from the actual sense in which the Malayalam impersonal passive has been used. The literal sense will be somewhat like: 'a mistake got made by Ram unintentionally'. Thus the reverse translation for the same translation sentence from English to Malayalam involves far more complex procedure. A somewhat similar dimension of divergence between Malayalam and English is manifested with respect to the negative impersonal passive constructions in Malayalam and the way they are realized in English.

(12) rAmanekkoNt natakkAn_ paZZukayilla.
   (Ram PP)   (walk)   (can NEG)
   Ram cannot walk.
   rAman natakkAn_ sAXikkilla.
   (Ram) (walk)   (can NEG)

In this case, too, translation divergence occurs in the case of the reverse translation and the source Malayalam sentence cannot be obtained.

In Malayalam, some of the aspectual features of the verb are realized by verbal inflection whereas English resorts to different non-inflectional ways such as phrasal verb or an adverbial element or a prepositional phrase with gerund as the head, to realize them. For instance, in (13-14), the aspectual property is identical in both the sentences and the difference is located only in tense. The habitual aspect of the tense is reflected by inflectional morphology on the

verb in both the tenses. However, this habitual aspect in English is realized by the use of a phrasal verb in the case of the past tense (13) and by the use of an adverbial word '*often*' in the case of the present (and future) tense (14). Thus the adverbial element in Malayalam is optional whereas the one in English cannot be optional. In (15), we notice that the non-terminative aspect is realized by verbal morphology in Malayalam whereas English uses a phrasal structure to realize this aspect.

(13) rAman_ varumAyirunnu
    (Ram) (come be FUT PAST)
    Ram used to come.

(14) rAman_ itaykkite varumAyirunnu
    (Ram) (often) (come be FUT PAST)
    Ram often comes.

(15) rAman_ saMsAriccukoNtEyirunnu
    (Ram) (speak be PAST CONT)
    Ram kept on speaking.

In certain types of conditional clauses in Malayalam, there is optainality between present and future/past tenses. But the English counterparts of these Malayalam sentences always require the verb to occur in the present tense.

(16) Del_hiyil_ pOkukayANefkil_ nI vijayikkuM
    (Delhi (go CONDIT (you) (successful)
    If you go to Delhi you will be successful.

The reverse translation from English to Malayalam will produce only the source Malayalam sentence that has the verb in the present tense form and hence will not involve any translation divergence.

Role of Conjunctions and Particles

Another source of divergence between Malayalam and English can be located in the case of the use of different conjunctions and particles in Malayalam. We take examples involving some of these particles in Malayalam such as *alla*, *ANO* and *awO*. It can be used to indicate alternate conjunction in an affirmative sentence (17) and an interrogative sentence (18) in Malayalam.

(17) sIwa enne kaNtu avaneyalla.
 (Sita) (me) (met) (he NEG)
 Sita met me not him.

(18) rAman_ paTikkukayANO awO uZaffukayANO?
 (Ram) (study+QP) (or) (sleep+QP)
 Does Ram study **or** sleep?

In another instance, *awO* ('or') is a coordinate conjunction particle in Malayalam that conjoins two clauses or phrases. Here the question particle '*ANO*' is attached to both the conjoining clauses and phrases.

For instance, when *awO* ('or') is used in a sentence that has an interrogative marker, '*ANO*' functions as an interrogative marking particle. The contrast is shown in (19-20).

(19) avan_ Del_hiyilEkkO kol_kkawwayilEkkO pOyi
(he) (Delhi COORD) (kolkata COORD) (go PST)
He has gone either to Delhi **or** to Kolkata.

(20) avan_ Del_hiyilEkkANO awO kol_kkawwayilEkkANO pOyaw?
    (he) (Delhi LOC QP) (or) (Kolkata LOC

QP)   (go PAST)

Has he gone to Delhi **or** Kolkata?

avan_ del_hiyilEkkO   kol_kkawwayilEkkO pOyittuNtO?

        (he)   (Delhi LOC COORD) (Kolkata LOC COORD)   (go QP)

In the reverse translation of the English sentence in (20) back into Malayalam, we notice that the question particle is combined with the verb instead of nouns in the first case. Here the question particle '*ANO*' is replaced with '*uNtO*'.

Asymmetry in Transitivity and Causativity

The divergence related to the morphology-syntax asymmetry for Malayalam-English translation pair can be located in the difference in the realization of certain transitive verbs and most of the causative constructions in Malayalam and English.

In Malayalam, there are three forms of a verb (in this case *uruNtu* 'roll') that are morphologically derived. (*uruNtu => urutticcu => uruttippiccu*). The English counterparts of these sentences show that in English, there is only one lexical verb '*roll*' and the other forms are realized by syntactic processes (such as resorting to different kinds of verbal constructions). In Malayalam, *urutticcu* is a transitive verb which does not have a lexical counterpart in English (English has only the intransitive form as a lexical item). In English, it is realized by using two verbs *make* and *roll*. *uruttippiccu* is a lexical causative verb in Malayalam which in English is realized by using three verbs *get*, *make* and *roll*, with separate argument structures of their own. The English counterpart of the Malayalam example in (22) appears to be a forced translation. In

certain cases, it is quite difficult to obtain an exact translation of a common Malayalam ditransitive verb. For instance, in (23), the English counterpart of the transitive verb *aZiyiccu* is *inform*. However, Malayalam also has a ditransitive verb *aZiyippiccu* derived from *aZiyiccu*. English does not have a counterpart of this ditransitive verb.

(21) rAman_ sIwayekkoNt panw urutticcu.
     (Ram)   (Sita PP)   (ball) (roll CAUS)
     Ram made Sita to roll the ball.

The verb '*give*' is used because there is no exact English counterpart of the Malayalam verb *uruttippiccu*. Thus the reverse translation involves a far more complex procedure. Gaps of this type are quite common between Malayalam and English.

(22) rAman_ mOhanAl_ sIwayekkoNt panw uruttippiccu.
     (Ram) (Mohan INSTR)(Sita PP) (ball) (roll CAUS)
     Ram **got** Mohan **make** Sita to roll the ball.

(23) avan_ FaffaLe vivaraM aZiyippiccu
     (he)   (we)   (news)   (inform CAUS)
      He informed us the news
      avan_ FaffaL_kk vivaraM wannu
     (he)   (we)   (news) (give PAST)

Stative and Progressive Aspect

English seems to lack an exact counterpart of Malayalam stative verb/adjective which is realized by the progressive aspect marker. In English, there is no distinction between the progressive aspect denoting sentence and its stative counterpart. The English verbs such as *sit*, *stand*, *sleep*, and *wake* fall in this category. In Malayalam, they are distinguished by different lexical form of the relevant verb.

(24) rAman_ kasErayil_ irikkunnu
   (Ram)   (chair LOC) (sit+PR CONT)
   Ram is sitting on a chair.
    rAman_ kasErayil_ irunnukoNtirikkunnu.
   (Ram)   (chair LOC) (sit PR CONT INF)

The divergence of this kind seems to involve both lexical and structural aspects of the languages involved. If '*sitting*' is entered in the lexicon both as an adjective and a (form of) verb, only then the source Malayalam sentence can be obtained in reverse translation.

Participle Modification

The participle modifiers in Malayalam are mostly realized by relative clauses in English. For instance, in Malayalam, *pOkunna* is a suffix that, besides denoting several other functions, also functions as an adjective.

(25) nALe   varAn_   pOkunna ALukaLe enikku kANAn_ sAXikkilla.
     (Tomarrow) (come INF) (who)   (people) (me)   (meet INF) (able NEG)
     I will not be able to meet the people who are coming tomorrow.

In (25), we notice that in Malayalam, the noun *ALukaL_* ('people') is modified by a participial adjectival phrase *nALe varAn_ pOkunna* that precedes the head noun. However, in English, the same is realized by a relative clause construction that follows the main clause. In this case, an (adjectival) phrase in Malayalam is realized by a clause in English.

Gerund and Participle Clauses

Another significant source of divergence in Malayalam and English MT can be located in the way the various clausal complements and adjuncts (such as verbal participles) in Malayalam are realized in English.

(26) avan_ vannitt sanwORavAnAyi
     (He)   (come) (happy be PAST)
      He got happy to come.

(27) avan_ ennOt saMsArikkAn_ vannu.
     (he)   (me)   (talk INF)   (come PAST)
      He came to talk to me

(28) avan_ iw ceyyAn_ prApwanalla.
     (He) (this) (do INF)   (able NEG)
      He is not able **to do** this.

(29) avan_ pOkAn_ Agrahikkunnu.
     (He)   (go INF)   (want PR)
      He wants to go.

In the Malayalam sentences in (26-29), the adjunct verbal clauses and complement verbal clauses are realized by different structures which in English are mapped by a single structure. Thus the reverse translation for this set of examples in (26-29) faces different type of difficulty. In the former case, it is many-to-one mapping whereas in the latter case, it is a one-to-many mapping.

Clausal Conjunction

Another difference that is manifested between Malayalam and English is with respect to clausal conjunction where the subordinate clause is used to express different types of clause in Malayalam. In English, they are not always realized by the same type of clause, rather they are realized by different devices such as a modal verb.

(30)   enwO   AvO   avan_   eviteyefkiluM pOyittuNtAvuM.

(May be) (he) (somewhere CONDT) (go FUT)

(31) He might have gone somewhere.

avaneviteyefkiluM pOyittuNtAvuM

here the reverse translation does not give the source Malayalam sentence.

Mapping of *have-verbs* in Malayalam

Certain English have-sentences are difficult to be exactly mapped onto Malayalam. Besides its polysemous nature, *have*-constructions also involve structural aspects and constitute a case of translation divergence.

(32) avanil_ sAhasamuNt
(he LOC) (courage be)
He has courage.

(33) avan mUnn kuttikaL_ uNt
(he) (three) (kids) (be)
He has three kids.

(34) avanZe kayyil_/(pakkal_) paNaM uNt.
(he-of) (hand) (near) (money) (be)
He has money.

Some of the representative examples, as in (32), can show the divergence issues involved in translating the Malayalam counterparts of the English have-constructions. In Malayalam, the subject NP occurs in different case forms but they all are mapped onto English by have-verb sentences. In the reverse translation from English to Malayalam, although there will be one-to-many mapping but the nature of the divergence will remain the same.

Had-Counterfactual Clause

In Malayalam, the counterfactual conditional clause is marked by a conjunction *efkil_*('if') which in English can be realized either by a had-clause or an if-clause. In the former case, translation divergence occurs.

(35) wAfkaL_ iviteyuNtAyirunnefkil_ FaffaLuM varumAyirunnu.
(you) (here been CONDT) (we) (come)
Had you been here we would have also come.

In reverse translation for the English sentence, the divergence remains the same.

Let-sentences

The Malayalam permissive-sentences are mostly translated into English by "let-sentences", as in (36a). However, there are certain wish-sentences that also occur in the form of a let-sentence (36b).

(36a) avane pOkAnanuvaxikkU
(him) (go permit FUT)
Let him go.

(36b) pOkAM, BakRaNaM kaYikkAM
(go FUT) (food) (eat FUT)
Let us go and eat now.
namukk pOyi BakRaNaM kaYikkAM
(we) (go) (food) (eat FUT)

The difference in the English sentences between (36a) and (36b) is only in the use of a pronoun. The use of first person plural pronoun 'us' in (38b) makes the sentence a wish-sentence rather than a permissive-sentence. Thus the reverse translation in (36a) does not involve translation divergence whereas in (36b), a translation divergence occurs. The nature of this translation divergence again pertains to the gaps in the realization of different verbal inflections and

functions between English and Malayalam.

1) News Headings

The news headings in English and Malayalam follow different grammar rules. In English, generally the present tense form of the verb is used whereas Malayalam uses past tense form of the verb.

(37) sunAmiyil_ lakRakkaNakkin AL_kkAZ_ mariccu

      (Tsunami-in) (millions PP) (people) (die PAST)

    Millions die in Tsunami

      Sunamiyil_ lakRakkaNakkin AL_kkAZ_ marikkunnu

      (Tsunami-in) (millions PP) (people) (die PRES)

In this case, translation in both the Malayalam-English and English-Malayalam cases involves divergence.

2) Optative Sentences

The verb in the subordinate clause in the sentences of optative mood in Malayalam occurs in different forms depending on the gender, number and person of the subject NP of the subordinate clause whereas its English counterpart remains constant (root form of the verb) in all the cases. In this case, the verb form does not indicate tense. The divergence is triggered by the similarity of the form of the verb with other tense forms in the case of English whereas it is not the case in Malayalam.

(38) rAman_ jayikkaNamenn FaffaL_ Agrahikkunnu

    (Ram) (success PP) (we) (want PR)

    We want that Ram succeed.

This type of divergence can be resolved by taking into account the semantic type of the verb.

3) Indirect Speech

The indirect speech sentences in Malayalam and English differ in both the form of tense and the use of pronominal elements.

(39a) rAman_ paZaFFu FAn_ pOvilla enn

    (Ram) (say PAST) (I) (go NEG)(that)

    Ram said that I would not go.

(39b) rAman_ paZaFFu FAn_ varunnu enn

    (Ram) (say PAST) (I) (come PROG PR) (that)

    Ram said that I was coming.

The use of the pronoun *FAn_* (I) in (39a) is ambiguous and can be translated either by 'I' or 'he' in English. The example in (39b) shows that the tense in the English indirect speech sentences is past but must be mapped by present tense in the Malayalam sentence.

4. Socio-cultural Factors

Different natural languages have different mechanisms to indicate socio-cultural features leading to a variety of divergence in translation. We examine two of these in case of Malayalam-English translation in the following sections.

Honorificity Markers

In Malayalam, the honorific feature is marked by the introduction of pronouns that are used for honorific use in the language. In the example we can see that the pronoun 'axXEhaM' is used for honorific purpose. Otherwise we use 'avan_' as the target language equivalent for 'he'.

(40) rARtrapawi vannu. axXEhaM ippOL_

saMsArikkum

(president) (come PAST).(He)　　(now) (lecture deliver FUT)

The President has arrived. He will deliver a lecture now.

In both Malayalam-English MT and English-Malayalam MT, the divergence caused by this socio-cultural aspect of the respective language arises.

Mappings of Time

Usually, people's perception of different objects in the world is dependent upon several socio-cultural beliefs. For instance, time is conceptualized in the Indian culture differently than that is done in the Western culture. These concepts are expressed through our respective languages and difference in concepts manifests itself in the language that is the source of translation divergence. For instance, in English, the concept of a.m. vs. p.m. cannot be exactly mapped in Malayalam. The Malayalam counterpart of a.m. and p.m. denote only a small part of time and the other parts of time is denoted by different other terms. The example in (41a) shows that the time at the 5 o'clock in the morning is denoted by a.m. in English but the exact translation of a.m. in Malayalam does not produce an appropriate Malayalam expression. However, in (41b), we notice that the time at the 11 o'clock in the morning which is expressed in English by a.m. can also be expressed in Malayalam by the exact translation of the term a.m. A similar situation is noticeable with respect to the mapping of p.m. in examples (41c-e).

(41a) He arrived at 5 a.m.

avan_ rAvile aFcumanikk vannu.

(He)　(morning) (5 o'clock) (come PAST)

He arrived at 5 o'clock in the morning.

(41b) He arrived at 11 a.m.

avan_ rAvile pawinonnu maNikk vannu.

(he)　(morning) (11)　(o'clock) (come PAST)

He arrived at 11 o' clock in the morning.

(41c) He arrived at 1 p.m.

avan_ uccaykk oru manikk vannu.

(he)　(afternoon) (1 o'clock) (come PAST)

He arrived at 1 o'clock in the afternoon.

(41d) He arrived at 5 p.m.

avan_ vQkunnEraM aFc manikk vannu.

(he)　(evening)　(5 o' clock) (come PAST)

He arrived at 5 o'clock in the evening.

(41e) He arrived at 10 p.m.

avan_ rAwri paww maNikk vannu.

(he)　(night) (10 o'clock)　(come PAST)

He arrived at 10 o'clock in the night.

These findings are based on the rules derived during the development of Machine Aided Translation system for English-Malayalam language pairs. Some more divergences can be observed between Malayalam and English language pairs by analyzing it in detail.

Conclusions

In this paper, we have examined the issue of translation divergence in Malayalam-English MT keeping in view the classification of translation divergence in some of the existing works on Hindi-English MT (Dave et al 2001, Gupta et al 2003, Sinha and Thakur 2005a). We have also discussed the reverse translations (English-Malayalam) for some of the translation

pairs to examine the nature of divergence (if any) in the case of reverse translation. We have observed that there are a number of areas in Malayalam-English translation pair that fall under translation divergence but cannot be accounted for within the existing parameters of classification. We propose that to capture these (other) types of translation divergences from Malayalam to English and vice-versa, we need to further modify the classification and augment it by new categories and subtypes. The topics of translation divergence discussed in this paper should provide insights into the complexity of translation divergences in Malayalam and English and give a direction for their further classification and resolution. However, due to constraints on space, it has not been possible to include discussions on issues of divergence related to several particles in Malayalam and mappings patterns of certain possessive constructions which in English are realized by 'have'-construction.

Further, due to space constraints, we have only briefly outlined strategies used our MT systems for translation from Malayalam to English with these translation divergences.

Abbreviations/Acronyms

PROG- Progressive, PR – present tense, QP- Question Particle, PAST – Past Tense, EW – Echo word, PP – Postposition, PASS – Passive, INST – Instrumental, NEG – Negative, FUT- Future Tense, CONT- continuous, CONDT – Conditional, COORD- Coordinator, INSTR – instrumental, CAUS – causative, LOC – Locative, INF – Infinitive

References

1. R. Mahesh K. Sinha and Anil Thakur, A Study of the Translation Divergence in English and Hindi MT, CSI Journal, Vol. 38, Issue No.3, Jul-Sept. 2008. (Updaed version of Divergence Patterns in Machine Translation between Hindi and English, 10th Machine Translation summit (MT Summit X), Phuket, Thailand, September 13-15, 2005).

2. R.E. Asher, T.C. Kumari, Malayalam - Descriptive grammars, 317-319, Routledge(2007)

3. A R Rajaraja Varma, Keralapaniniyam, Pages 177-269, DC Books(2000)

4. Suranad Kunjan Pillai, Volume 1, Malayalam Lexicon, Appendix(1-105), Introduction xviii The University of Kerala(2000)