Rule Based Case Transfer in Tamil-Malayalam Machine Translation

S. Lakshmi and Sobha Lalitha Devi

AU-KBC Research Centre, MIT Campus of Anna University, Chennai,
India

sobha@au-kbc.org

Abstract. The paper focuses on the rule based case transfer, which is a part of the transfer grammar module developed for bidirectional Tamil to Malayalam Machine Translation system. The present study involves two typologically close and genetically related languages, namely Tamil and Malayalam. We considered the basic construction of sentences which is highly dependent on the case systems. The rules were written by taking into consideration the Postpositions and cases in the languages. A parallel corpora was chosen and a deep analysis of the case transfer patterns were done and rules were written to sort out the case changes that happens when translating from one language to another. We have also considered copula transfer in our approach. Web data was used for evaluation and the results were encouraging.

Keywords: Case suffixes, Dravidian languages, machine translation.

1   Introduction

One of the main components of the machine translation system is the transfer grammar that transfers an intermediate representation of the source language to an intermediate representation of the target language. The transfer grammar constitutes of lexical level transfer and structural transfer. In our approach case transfer is taken into consideration. Cases have been used in theChomskyan framework to trigger movement. In Dravidian languages, grammatical relations and semantic roles are usually explained with the help of case suffixes. Case is most easily observed and studied in languages that have a rich case morphology.

Tamil and Malayalam are closely related to each other in grammar and vocabulary than the other two Dravidian languages, Kannada and Telugu. Malayalam is highly influenced by Sanskrit language at lexical, grammatical and phonemic levels were as Tamil is not. The Noun morphology is same in both the languages as the word may contain the root alone or root with suffixes attached to it. Agglutination is widely seen in Tamil and Malayalam. In Tamil and Malayalam the case markers are seen attached to the noun and pronoun information. Postpositions are also seen attached to it. In traditional analysis, there is always a clear distinction made between postpositional

morphemes and case endings. Both the languages belong to the category of nominative-accusative languages. The Tamil verbs inflect for person, number and gender whereas Malayalam verbs do not take person, number and gender termination. Hence the gender marking of the noun is not a relevant feature when Malayalam language is considered. Tamil nouns inflect for case, number (singular and plural) and gender. So when translating from Tamil to Malayalam the verb PNG marker is subdued. A variety of case changes have been observed in the two languages and rules have been formulated. Consider the following example

An accusative dropping was noted when moving from Tamil to Malayalam.

1. *Ta: avan   panthai   eduthaan*
      he   ball-acc   take-past+3sm
   Ml: avan   panth   eduthu
      he   ball-nom   take-past
      (He took the ball.)

In the above example 1 the accusative marking in Tamil is being mapped to nominative case in Malayalam. Malayalam is a language in which only animate objects are marked with accusative case [9]. Rules have been written to handle the accusative drop.

The syntactic difference between languages can be studied to identify an underlying word order in the source language that might be similar to the target language word order. Many approaches have incorporated syntactic information within statistical machine translation systems to obtain better results. Lavie has presented a Stat-XFER, a general search based and syntactic driven framework for developing MT systems [6]. Carbonell, J. G. et al., [1] have developed knowledge based MT by combining syntactic and semantic information to produce an intermediate knowledge representation of the source text which is then generated in the target language. Dave, S., et al., [2] studied the language divergence between English and Hindi and its implication to machine translation between these languages using the Universal Networking Language (UNL).Koehn et al., [4] showed heuristic learning of phrase translations from word-based alignments and lexical weighting of phrase translations leads to significant improvement in translation accuracy. To handle syntactic differences, Melamed [8] proposes methods based on tree-to-tree mappings.Sobha et al., [16] described syntactic structure transfer in a Tamil-Hindi Machine Translation system using hybrid approach where they learned the structures from clause identified parallel data and incorporated it into a rule based system. Sobha et al., [17] has also used a rule-based approach to transfer nominal constructions from Tamil to Hindi. Case transfers from English to Hindi and vice versa has been approached by Sinha [13,14] and case transfer pattern analysis from Hindi to Tamil MT was done by P. Pralayankar et al.,[10].

The paper is organized as follows. In the next section we give a detailed description of various transfers that happen in the Tamil-Malayalam Machine Translation system such as syntactic structure transfer, case transfer and copula transfer. Then we have briefly explained our approach and the computational aspect. The results for the case transfers and conclusion section follows.

2 Types of transfers

Following transfers can happen in transfer grammar module.

1. Syntactic Structure Transfer,
2. Case Transfer, and
3. Copula Generation.

2.1 Syntactic Structure Transfer

The goal of this syntactic structure transfer is to improve the translation grammatically and to give the naturalness to the target language structures [16]. Tamil and Malayalam has similarity at the basic structure level, hence we have given more importance to the lexical level transfers.

2.2 Case Transfer

Lehmann classifies the Tamil case system into 9 cases [5] and Malayalam has been classified to 7 cases [12]. We have done a mapping of the case systems in the two languages and represented it in the table below.

Table 1. Case mapping.

Case	Tamil	Malayalam
Nominative	NULL	NULL
Accusative	Ai	e
Dative	Kku	kk,n
Instrumental	aal, kontu	aal,kont
Locative	il, itam	il,thth
Ablative	Iliruntu	ilninn
Benefactive	Ukkaaka	kkaayi
Sociative	ootu, utan	ot
Genitive	utaiya, in, atu	nte,ute

To analyse the case transfers we have chosen a parallel corpora. In the sections below a detailed description of case transfers is considered by looking into each specific case.

(a) Nominative Case

The nominative case in Tamil and Malayalam is unmarked. A nominal case is identified by the subject of a sentence in its unmarked form. Nominative noun can function as agent and experiencer as shown in example 2.

2. *Ta: avaL   aluthaaL*
    she-nom   cry-past+3sf

> *Ml: avaL   karanju*
> *she-nom   cry-past*
> *(She cried.)*

(b) Accusative Case

The accusative marker usually follows the object. The accusative case in Tamil marks the direct object noun phrase of a transitive verb. The accusative marker is 'ai' in Tamil and 'e' in Malayalam.

> 3. *Ta: meri   avanai   paarthaaL*
> *Mary-nom   him-acc   see-past+3sf*
> *Ml: meri   avane   kandu*
> *Mary-nom   him-acc   see-past*
> *(Mary saw him.)*

An accusative drop was noted when moving from Tamil to Malayalam. Consider the example given below.

> 4. *Ta: avan   panthai   eduthaan*
> *he-nom   ball-acc   take-past+3sm*
> *Ml: avan   panth   eduthu*
> *he-nom   ball-nom   take-past*
> *(He took the ball.)*

In Malayalam the accusative suffix is usually dropped in a sentence where the subject-object distinction is clear [11]. In Tamil when the direct object is human, the accusative marker is obligatory, but when non-human object occurs accusative marker signals definiteness [19]. Mohanan has observed that in Malayalam language only animate objects take accusative markers. In the above examples we can see that in example 3 accusative case in Tamil is mapped to accusative in Malayalam and in example 4 the accusative case in Tamil is being mapped to nominative case in Malayalam.

Consider the example 5 given below.

> 5. *Ta: avaL   ammaavai   velai ceyyavethaaL*
> *she-nom   mother-acc   job   do-past-caus+3sf*
> *Ml: avaL   ammaye   koNt joli ceyyiccu*
> *she-nom   mother-acc   psp   job do-past-caus*
> *(She made her mother work.)*

Here the accusative case in Malayalam is marked by the addition of a postposition (koNt) which represents an agentive role.

(c) Dative Case

The dative suffix 'kku' in Tamil is transferred to 'kk' or 'n' in Malayalam. A case divergence has been noted for dative and genitive markers in Malayalam. It was observed by Asher et al., that in Malayalam language dative 'n' occurs with noun roots

ending in 'an' and dative 'kk' occurs with other singular nouns and all plurals. Consider example 6 given below.

6. *Ta: ainth manikin   pujai   natakkum*
   Five o'clock-dat   worship-nom happen-fut
  Ml: anchu maNikk   puja   natakkum
   Five o'clock-dat   worship-nom happen-fut
   (The worship will happen at five'oclock.)

Here the dative case in Tamil is mapped to dative in Malayalam. Given below are some exceptions.

7. *Ta: naalaikku   paritcai   thotankukirathu*
   Tomorrow-dat   exam   start-fut
  Ml: naale   pareeksha   thutangum
   Tomorrow-nom   exam   start-fut
   (Exam will start tomorrow.)

In example 7 dative case gets mapped to the nominative case. In Malayalam the subject acquires a dative case only if there is no nominative noun in the sentence.

8. *Ta: raaman   vittirkku   centran*
   Raman-nom   house-dat   go-past+3sm
  Ml: raaman   vittil   poyi
   Raman-nom   house-loc   go-past
   (Raman went home.)

In example 8 dative case of the object gets mapped to locative case. Here the object denotes destination and hence in Malayalam locative marker is used to give the sense.

9. *Ta: amma   kathavukku pinnati   nintral*
   Mother-nom   door-dat   behind-adv   stand-past+3sf
  Ml: amma   kathakinte   pinnil   ninnu
   Mother-nom   door-gen   behind-adv   stand-past
   (Mother stood behind the door.)

In example 9 in Tamil when a dative noun is followed by a place adverb it is found that in Malayalam the dative case gets mapped to genitive. The place adverbs can indicate the static location or movement.

(d) Locative Case

The locative case in Tamil is marked by 2 markers 'il' and 'itam'.Here 'il' specifies 'the place in which' and 'itam' is used for animate nouns to indicate 'with the person'. In Malayalam the locative marker is il and in some cases specified using the marker 'thth'.

10. *Ta: tiraivar   hinthiiyil kettaar*
    driver-nom hindi-loc ask-past+3sm
     Ml:drivar   hindiyil   codiccu

    driver-nom hindi-loc ask-past
    (Driver asked in Hindi.)

There were some specific cases were locative marker in Tamil was mapped to Dative and also Genitive +place adverb which is provided in the examples below.

11. *Ta: kaikeyiyitam varam kotuththan*
     Kaikeyi-loc boon give+past+3sm
   Ml:kaikeyikku varam koduthu
     kaikeyi-dat boon give+past
    (Gave boon to kaikeyi.)

In example 11 the locative case marker in Tamil changes to dative in Malayalam. In Malayalam there is no distinction for locative marking in animate or inanimate nouns.

12. *Ta: puvitam   vantu vanthathu*
     flower-loc bee   come-past+3sn
   Ml: puvinte   atuth   vant vannu
     Flower-gen near-adv bee come-past
     (The bee came near the flower.)

In example 12 locative marker in Tamil gets transferred to genitive +adverb in Malayalam.

    (e) Genitive

The genitive marker is used to denote possession, relationship and many such semantic relationships. The genitive case is realized by the markers '-in', '-uTaiya' or '-atu', in Tamil. In Malayalam 'nte' and ute are two genitive markers. Genitive nouns are followed by the noun which it modifies.

    Given in example 13 is a genitive transfer from Tamil to Malayalam.

13. *Ta: ithu indyaavin vekamaana rayil*
     this india-gen fast-adj   railway
   Ml: ith   indyayute vegamulla rayilaaN
     this india-gen fast-adj railway-is
     (This is India's fastest railways.)

Consider the example below.

14. *Ta: raaja aranmanaiyin araiyil   ninRaar*
     king palace-gen   room-loc stand-past+3sm
   Ml: raajav kottarathile   muriyil   ninnu
     king   palace-loc+acc room-loc stand-past
     (King stood in a room inside the palace.)

In the example 14 above genitive case in Tamil gets replaced to locative+accusative suffix better known as locative copula which is derived from hidden copular verb 'ulla' [11] in Malayalam.

(f) Instrumental

The instrumental suffix 'aal' in Tamil as well as Malayalam is used to specify the means of the cause. Given below is an example for this case suffix transfer. Hence the transfer was from instrumental 'aal' in Tamil to 'aal' in Malayalam.

15. *Ta: raamu   penaavaal   ezhuthinaan*
        ramu   pen-ins   write-past+3sm
   Ml: raamu penayaal   ezhuthi
        ramu   pen-ins   write-past
        (Ramu wrote with a pen.)

(g) Sociative

The sociative case in Tamil, realized by the markers '-ootu' or '-utan' expresses association or means by which action is done. In Malayalam single marker for sociative case is 'ootu'.
   Consider the examples given below

16. *Ta: raaman puthakathotu vanthaan*
        raman   book-soc   come-past+3sm
   Ml: raaman pusthakathote vannu
        raman   book-soc+acc come+past
        (raman came with a book.)

Here the sociative case in Tamil is being mapped to sociative+accusative case in Malayalam.

17. *Ta: kantrutan pacu vanthathu*
        calf-soc   cow come-past
   Ml: kitaavinte_koote pasu vannu
        calf-gen   psp   cow come-past
      (The cow came with the calf.)

Here sociative case 'utan' known as bound postposition [5] in Tamil gets changed to genitive case followed by post-position 'koote'.

(h) Ablative Case

The ablative case, marked by '-iliruntu' or '-itamiruntu', is transferred to 'ilninn' in Malayalam. In Tamil 'iliruntu' represent the motion from inanimate object and '-itamiruntu' represents the motion from animate object. It was noted by Asher that in Malayalam there is no such case like ablative but it is locative marker il + postposition ninn which provides the meaning of a 'source'.
   An example is given below

18. *Ta: ithu manaaliyilirunthu 50 ki.mI. thoorathil   irukkirathu*

      this Manali-abl   50 k.m. distance-loc   is-present+3sn
   Ml: ith   manaaliyilninn   50 ki.mI. doorathil   aan

> *this   Manali-abl   50 k.m   distance-loc   is-present*
> *(It is 50 k.m from Manali.)*

(i) Benefactive Case

The benefactive case in Tamil, realized by the case marker '-ukkaaka', which is transferred to 'kkaayi' in Malayalam. Given example 19 illustrates the benefactive case transfer.

> 19. *Ta: enakkaaka tiraivar kaathirunthaar*
>   *me-ben   driver wait-past+3sm*
> *Ml: enikkaayi   drivar kaathirunnu*
>   *me-ben   driver   wait-past*
>   *(Driver waited for me.)*

2.3   Copula Generation

A copula can be defined as a verb or a verb like word. They are capable of functioning as the main verb, but are grammatically and semantically different from action verbs. Tamil lacks copula. It is included only to convey the meaning. In Malayalam "aak" (form of being) and "unt" are two copulas used. They have been defined as equative and existential copulas [12]. So copula generation have been considered when moving from Tamil to Malayalam.

> 20. *Ta: avanukku ammaavai   pitikkum*
>   *he-dat   mother-acc like*
> *Ml :avan   ammaye   ishtam aaN*
>   *he-dat   mother-acc   like be-pres*
>   *(He likes (his) mother.)*

3   Our Approach

3.1   Steps Involved in Case Transfer

In the following section we describe the various steps involved to handle the case transfer from Tamil to Malayalam.

– Identify the noun phrases in Tamil source sentence.
– Identify the verbs in the sentence.
– For each noun phrase in the sentence do the following.

1. If the noun word is having nominative case, then the case transfer is nominative itself.
   – If the noun is marked with accusative case and for a list of verbs, then change accusative case suffix into nominative-accusative

     ○ For another list of verbs and if the current verb is causative, then add postposition "kont" to the accusative marker.

      ○ else transfer accusative case marker "ai" to "e" .

2. If the noun is having dative case then

- For a list of verbs change dative case into locative.
- For a list of verbs if the dative case is followed by a place adverb then change the dative case+ adverb to genitive case+adverb.
- For another list of verbs change dative case to nominative.
- else

    (a) If the noun word is singular and masculine then dative case is transferred to 'n'.
    (b) If the noun is plural and masculine then dative case is transferred to 'kk'.
    (c) If the noun is singular or plural and feminine then dative case is transferred to 'kk'.

3. If the noun is having sociative case then

- For a list of verbs and if the sociative case marker is 'utan'

    (a) If the noun word is singular and masculine then change sociative case to genitive 'nte'+psp 'koote'.
    (b) If the noun word is plural and masculine then sociative case is transferred to genitive 'ute'+psp 'koote'.
    (c) If the noun word is singular or plural and feminine then sociative case is transferred to genitive 'ute'+psp 'koote'.

- For another list of verbs transfer sociative case to sociative +accusative
- else transfer sociative case marker in Tamil 'otu' to sociative case marker 'ot' in Malayalam.

4. If the noun is marked with genitive case then

- For a list of verbs transfer genitive case to locative +accusative.
- else

    (a) If the noun word is singular and masculine then dative case is transferred to 'nte'.
    (b) If the noun is plural and masculine then dative case is transferred to 'ute'.
    (c) If the noun is singular or plural and feminine then dative case is transferred to 'ute'.

5. If the noun is marked with instrumental case then transfer case marker 'aal|kontu' then transfers it into 'aal|kont'.
6. If the noun is having ablative case marker 'iliruntu' then transfer it into 'ilninn'.
7. If the noun is marked by benefactive case marker then transfer it into 'kkaayi'.

4   Computational aspect of the case transfer

To give a computational view of the case transfer an example is illustrated below with traces. We have preprocessed the data with morphological information, Parts-of-Speech tagging, chunking information. The example 21 shows how a dative to locative case transfer happens after applying the hand crafted case transfer rules.

21.   *seetha   kataikku   centraal*
       sitha-nom   shop-dat   go-past-3sf
       (Seetha went to shop.)

  After performing lexical analysis,
    Subject : seetha - sitha-NOM(NP1,sg,f)
    Object : kataikku- katai-DAT(NP2,sg,n)
    Verb   : centraal- go-PAST-3.sg.f
  On applying the transfer grammar rules,
    Subject: seetha (nom)
    Object: katai (il)
    Verb: cel (...)
  During lexical substitution,
    Subject: seetha
    Object: kata (il)
    Verb: pok (...)
  So, finally,
    seetha   katayil   poyi.
    sitha-Nom   shop-Loc   go-past
    (Sitha went to shop.)

5   Results and Discussion

We evaluated the system with 1000 sentences collected from web articles. The sentences were run through the Tamil-Malayalam Sampark system and the preprocessing errors were taken care as preprocessing errors such as POS tagging errors adversely affected the machine translation output. The correctness of the case transfer depends on the accuracy of the previous modules. Sampark is a platform for Indian language to Indian Language translation [10]. The sentences were evaluated for case transfer and the results are shown in the table below. The results shown are based on real translation system output.

Table 2. Table 2 : Performance of Case Transfer

Case	No. of Cases	Correctly transferred	Accuracy
Nominative	3485	3485	100%
Accusative	300	280	93.3%
Genitive	403	372	92.3%

Case	No. of Cases	Correctly transferred	Accuracy
Dative	213	198	92.95%
Sociative	28	18	64.28%
Locative	650	634	97.53%
Instrumental	76	75	98.68%
Ablative	99	99	100%
Benefactive	28	28	100%
Total	5282	5189	98.23%

From the results it is clear that the case transfer was properly handled and the transfer of sociative case was not properly handled.

Ta: vacista kiraamathil cutaana wannir nirurrukalil kulitha   pinpu nInkal viyaas
   vasisth village-loc hot   water   spring-loc   bath   after   you   beas

   nathiyotu celkintra poluthu colafka vaaykkaal vanthataiyalaam
   river-soc go-past-rp   solang tributary   can-reach

Ml: vasishta graamathil cutvella   aruvikalil   kulichathinu shesham ningal byaas
   vasisth village-loc hot-water spring-loc   bath   after   you   beas

   nathyilekk   ethunna   solanga kaivazhi   vannucheraam
   river-loc+dat   go-past-rp solang tributary   can-reach

(After taking a bath in the hot water springs of vasisth village you can reach the solang tributary of the beas river.)

Here "nathiyotu" in Tamil is having a meaning towards the river which is represented in malayalam using the locative+dative cases known as allative marker [7]. This was not transferred as per the rules we formulated. Some new rules is to be added where case transfer is not proper.

6   Conclusion

In this paper, we have given a rule based implementation of the case transfer for Tamil-Malayalam machine transfer. On applying the rules into the transfer grammar component and analysing the translation results an improvement was seen in the overall performance of the system. The case transfer has to be improved with more rules. Also various other lexical transfers can be considered to improve the transfer grammar performance.

References

1. Carbonell, J. G., Cullingford, R. E., and Gershman, A. V.: Steps toward knowledge-based machine translation. In: Pattern Analysis and Machine Intelligence, IEEE Transactions on, (4), 376-392 (1981).

2. Dave, S., Parikh, J., and Bhattacharyya, P.: Interlingua-based English–Hindi Machine Translation and Language Divergence. In : Machine Translation, 16(4), 251-304 (2001).
3. Dorr, B. J.: Machine translation divergences: A formal description and proposed solution. In: Computational Linguistics, 20(4), 597-633 (1994).
4. Koehn, P., Josef O.F, and Marcu, D.: Statistical Phrase-Based Translation. In: HLT/NAACL'03. pp. 127-133 (2003).
5. Lehmann T., A Grammar of Modern Tamil. Pondicherry Institute of Linguistics and Culture, Pondicherry, (1989).
6. Lavie, A. : Stat-XFER: A general search-based syntax-driven framework for machine translation. In: Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science. 362375, Springer, (2008).
7. Nirenburg, Sergei. : Knowledge-based machine translation. Machine Translation 4.1 5-24 (1989).
8. Melamed D., Statistical Machine Translation by Parsing. In: ACL, (2004).
9. Mohanan, K. P.,.: Grammatical relations and anaphora in Malayalam. In: Diss. Massachusetts Institute of Technology, (1981).
10. Pralayankar P., Kavitha V., and Sobha L.: Case Transfer pattern from Hindi to Tamil MT. In: PIMT Journal of Research. vol. 2. No. 1, pp. 26-31 March-August (2009).
11. Ravi Sankar S Nair. : A Grammar of Malayalam, www.languageinindia.com/ravisankarmalayalamgrammar.pdf.
12. R.E.Asher and T.C.Kumari. : MalayalamRoutledge, London and New York, (1996).
13. R.M.K. Sinha and Anil Thakur. : Translation Divergence in English-Hindi MT. In: EAMT, Budapest, Hungary (2005).
14. R.M.K. Sinha and Anil Thakur. : Pre-/post-positions Selection in Text Generation for Hindi and other Indian Languages for Translation from English. In: International Symposium on Machine Translation, NLP and Translation Support System, pp: 40-45 New Delhi, (2004).
15. Sangal R.: Project Proposal to Develop Indian Language to Indian Language Machine Translation System.IIIT Hyderabad, TDIL Group, Dept. Of IT, Govt.of India, (2006).
16. Sobha Lalitha Devi, R. Vijay Sundar Ram, Pravin Pralayankar and T. Bakiyavathi. : Syntactic Structure Transfer in a Tamil to Hindi MT System - A Hybrid Approach. In: A. Gelbukh (Ed), Computational Linguistics and Intelligent Text Processing, Springer LNCS Vol. 6008. pp 438 – 450, (2010a).
17. Sobha, Lalitha Devi., Kavitha V., Pralayankar P., Menaka S., Bakiyavathi T., and Vijay Sundar Ram R., Nominal Transfer from Tamil to Hindi. In: International Conference on Asian Language Processing (IALP), Harbin, China. pp. 270 – 273 (2010b).
18. Sobha Lalitha Devi, Sindhuja G., Vijay Sundar Ram R., Transfer Grammar in Tamil-Hindi MT System. IALP 79-82 (2013).
19. Steever, Sanford: The Dravidian Languages, London: Routledge, (1998).