Disambiguation of pre/post positions in English – Malayalam Text Translation

Jayan V, Sunil R, Bhadran V K
Language Technology Centre, Centre for Development of Advanced Computing (C-DAC),
Thiruvananthapuram, Kerala, India
jayan@cdac.in, bhadran@cdac.in, sunilrkp82@gmail.com

ABSTRACT

This paper presents the disambiguation of preposition in English to the corresponding post positions in Malayalam while translating text from English to Malayalam. Preposition in English will be replaced with post position in Indian Languages during translation. The polysemous nature of prepositions in English increases the difficulty in determining its exact meaning/function. This makes the task of mapping a preposition in English to an Indian language difficult, particularly for machine translation. Research or works on one to one mapping of English prepositions to Malayalam postpositions are not done for Malayalam. In this paper we illustrate the patterns of preposition mapping from English to Malayalam and present context disambiguation rules for determining the right mapping of different prepositions in English to postposition in Malayalam.

Keywords: Postposition, cases, suffixes

Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 93–102,
COLING 2012, Mumbai, December 2012.

1. INTRODUCTION

It has been observed that almost all postpositions in Malayalam function as case endings. Postpositions in Malayalam occur after the nominal. They perform similar to inflectional markers. But unlike inflectional markers, postpositions in Malayalam are free forms. Being invariants, they can stand alone or alongside another free or bound morpheme. Postpositions (PP) in Malayalam [1], [2], [3] are certain forms, which occur immediately after nouns and establish some grammatical relations between the nouns and the verbs of the sentences. In English, the semantic distribution of a single preposition will be varying in different context due to the influence of nouns and main verbs that follow. When an English preposition is translated into Malayalam, the following transformation takes place:
(preposition) (object) ← → (object) [(inflection)][(postpositional-word)].

I played **with** *Ram.* ← → ŋa:n *ra:man***re ku:te** kaliccu (ഞാൻ രാമന്റെ കൂടെ കളിച്ചു)

In the example given above the object is modified with a suffix and postposition after it. It is not necessary that both suffix and the postpositional equivalent word occur simultaneously. They may occur independently or together. It will depend on the type of object and verb.

The factors that determine the generation of appropriate postposition in Malayalam for a preposition in English is discussed in detail in this paper.

The correspondence between English prepositions and Malayalam postpositions (inflections and postpositional words) is not direct. The reference object plays a major role in determining the correct preposition sense. Reference object will decide whether the preposition is used in a spatial sense or other sense. A noun phrase (NP) denoting a place gives rise to a spatial postposition. Similarly, an object referring to a time entity produces a temporal expression.

For instance, a preposition *with* can have multiple mapping patterns in Malayalam, as shown below.

Example (1)

a. [with = kont(കൊണ്ട്)]

I wrote with the pen.

ŋa:n pe:na **kont**[1] eYuwi.

ഞാൻ പേന കൊണ്ട് എഴുതി

(I) (pen) (PP) (Write PST)

b. [with = kute(കൂടെ)]

Radha went with Gopi.

ra:dha go:piy**ute kute** pOyi.

രാധ ഗോപിയുടെ കൂടെ പോയി

(Radha) (Gopi) (PP) (go PAST)

c. [with = pakkal/vaSaM(പക്കൽ/വശം)]

[1] One of the reviewer mentioned that '*kontu*' is an instrumental case and not postposition, but in the paper "An Introduction to postposition in Malayalam for POS tagging" by Dr. S Radhakrishnan Nair mentioned that it is a postposition.

Sita has no money with her.

si:wayute **pakkal/vaSaM** panamilla.

സീതയുടെ പക്കൽ/വശം പണമില്ല.

(sita) (PP) (money NEG)

d. [with = Ayi]

A bus collided with a car.

bass ka:rum**a:yi** kuttiyiticcu.

ബസ് കാറുമായി കൂട്ടിയിടിച്ചു

(bus) (car PP) (collide PAST)

e. [with = nZe(ന്റെ)]

What is the problem with you?

ni**nre** praSnaM entha:N ?

നിന്റെ പ്രശ്നം എന്താണ്?

(you PP) (problem) (what QP)

From the above said examples we notice that the preposition with in English can be mapped in Malayalam in multiple ways. This is one of the challenging tasks in machine translation. The different functions of the prepositions need to be disambiguated for their correct mapping in machine translation. In example (1) above, the mapping patterns between English preposition and Malayalam postpositions can be resolved by taking into consideration semantic type of the main verb and that of the nominal elements in the sentence. In (1a) we can see that '*write*' is a verb that requires an instrument to carry out the action. So the agent is given the post position '*kont*'. Similarly for (1b) the verb '*went*' is a directed motion verb. Here this will map with the post position '*kute*'. In (1c), '*have*' is a contain verb. We can map this with the PP '*vaSaM/pakkal*'. In (1d) '*collide*' is a correspond verb, this can be mapped with the PP '*a:yi*'. Here we pointed out the disambiguation of the preposition '*with*' in English with the corresponding postposition equivalent in Malayalam in the translation context. This will be applicable to all Indian languages. The first step in disambiguation is identifying the class of verbs and the semantics of noun phrase (NP) coming as a subject and/object. Then map the preposition accordingly. A similar approach can be followed to other polysemic prepositions also.

In section 4 we present some common patterns of mapping of prepositions from English to Malayalam with respect to selected prepositions. In section 5 we present the major strategies that can be used to handle different preposition mapping patterns that we present in section 4.

2. RELATED WORKS ON POSTPOSITIONS

There is not much works has been taken up for the disambiguation of postpositions in Malayalam in the computational aspect. *Lilaatilakam* of 14th century A.D is considered as the earliest work which describes Malayalam grammar. Gundert does not use the term 'postpositions' but identifies some 'Noun Particles which are used with case suffixes. Caldwell has described various aspects of case system in Malayalam. Every postposition in his opinion affixed with a case will express a new case relation. The number of cases in Malayalam therefore depends upon the requirements of the speaker and the different shades of meaning he wishes to express. He

also points out that postpositions are in reality separate words and they retain traces of their original character as auxiliary nouns. Rev. George Mathan includes postpositions in taddhitaavyaya (derived connectives) ie, particles derived from nouns or verbs. Among the 150 taddhitaavyayas which he listed, only a few are seemed to be postpositions. Raja Raja Varma defines postpositions (gati) as a particle which modifies cases. He differentiated between case affixes and postpositions and uses the term 'misravibhakti' (mixed case) to indicate cases with postpositions. He also observes that the postpositions are not originally particles (nipaata) but particles derived from nouns or verbs (avyayas). Seshagiriprabhu describes the case system of Malayalam elaborately. He defines postpositions as certain words added to the case affixes to indicate special meaning and he lists eighty postpositions and classified them on the basis of the case affixes.

Similar work has been done for English-Hindi language pair as part of AnglaBharati Machine Translation system development by RMK Sinha et al. In that work the authors clearly mentioned the different prepositions in English and their effect on the target language as postpositions. The paper mainly focused on the language Hindi. For the language like Malayalam, an agglutinative language, some more factors must be taken in to consideration.

3. AN OVERVIEW OF THE ENGLISH MALAYALAM MACHINE TRANSLATION SYSTEM(ANGLAMT)

AnglaMT is a rule based machine translation system based on AnglaBharti technology developed by IIT Kanpur. It takes the English input sentences and passes through the preprocessor module for handling different formats like date, time, acronyms, abbreviations, etc. After preprocessing input text, the system will take all the syntactic and semantic information from the lexical database and processed further in the Morphological analyzer module. There after it goes through the rule base module. In rule base module, based on the information fetched from the morphological analyzer, the sentence gets parsed and generates an interlingua representation. It is generally known as PLIL (Pseudo Lingua for Indian Languages). PLIL will have the word order as that of target language. Here English is having subject-verb-object (SVO) pattern where as Malayalam is having the subject-Object-Verb (SOV) pattern. PLIL contain the root words of the source and target language along with their syntactic and semantic information. This PLIL is going to the text generator as input and text generator is adding the necessary suffixes and other target language dependant words. Depending on the complexity and multiple meaning for a word the system will generate alternate translations. This can be post edited manually as per the user requirement.

4. FEATURES OF POSTPOSITIONS

Postpositions in a Malayalam indicate case relations. It can be separated from noun phrases by coordinate conjunctions. They can be followed by case affixes. Normally they are disyllabic or polysyllabic and cannot take auxiliary verbs. Postpositions cannot be separated from noun phrases by morphemes other than coordinate conjunction. They cannot replace the present participial *–e* by temporal *um + po:l.* They are not having the grammatical inflection that is they are indeclinable. Prepositions will not be head of an endocentric construction and cannot be modified by adjectives. They occur only after nouns. They will be deleted in relativization. They cannot occur in the initial position of a sentence and cannot occur immediately after another postposition. They will be an immediate constituent of noun phrases.

5. MAPPING PATTERNS OF COMMONLY USED PREPOSITIONS

Consider some of the very commonly used prepositions *to* inorder to show the patterns of their mapping in Malayalam.

to: The preposition to in English can be mapped by different postpositions in Malayalam. The examples are listed in (2) to (6).

(2) [to = e:kk]

The procession goes to Kottayam.
Ja:tha ko:ttayathil**e:kk** po:kunnu.
ജാഥ കോട്ടയത്തേക്ക് പോകുന്നു.
(procession) (kottayam PP) (go PRS)

(3) [to = o:t]

I have spoken to him already.
ŋa:n iwinakaM thanne avan**o:t** saMsa:riccittunt.
ഞാൻ ഇതിനകം തന്നെ അവനോട് സംസാരിച്ചിട്ടുണ്ട്.

(I) (already) (he PP) (speak PAST)

(4) [to = atuthe:kk]

I am going to the king.
ŋa:n ra:ja:vinre **atuthe:kk** po:kukaya:N.
ഞാൻ രാജാവിന്റെ അടുത്തേക്ക് പോകുകയാണ്.

(I) (king) (PP) (go PRS)

(5) [to = e]

Please listen to him.
dayava:yi avan**e** Sradhiykku.
ദയവായി അവനെ ശ്രദ്ധിക്കൂ

(please) (him) (listen)

(6) [to = kk]

He is going to the meeting.
avan kutikka:Ycay**kk** po:kunnu.
അവൻ കൂടിക്കാഴ്ചയ്ക്ക് പോകുന്നു.
(he) (meeting+PP) (go PRS)

6. RULES FOR DISAMBIGUATION OF MULTIPLE PATTERNS OF PREPOSITIONS

In this section we propose rules for the disambiguation of the above said (section 2) examples having multiple mappings of prepositions. The examples clearly show the difficulty in mapping English prepositions to Malayalam postpositions in machine translation. The two major factors that determine the meaning of a preposition are the semantic type of the main verb and that of the nominal elements that occur with the prepositions. The syntactic and semantic information are extracted from the bilingual lexical database in the system. The dictionary contains more than 50000 entries. The structure of the dictionary entry is as given below:

1. abdication
2. 46 noun
3. ~G abdication
4. [activity]
5. pariwyAgaM:2
6. ***

First line is the English root word, second line contains the POS category and its inflection information as a category number, third line is the English meaning, fourth line is the semantic information, fifth line is the Malayalam root word and its paradigm number that represents the inflection of a noun based on its case and number.

We present some of the sample rules that are used to disambiguate prepositions. We use the semantic category of verb and noun to disambiguate the multiple patterns of the prepositions.

Disambiguation of *to*: As mentioned in section 2, the preposition *to* have different mappings in Malayalam. Rules for generating the correct mappings are illustrated below for the examples presented in section 2.

a. to-NP (place) = NP- e:kk (2)

b. to-NP(human) = NP-o:t (3)

c. verb(motion) to-NP = NP- atuthe:kk (4)

d. verb(mental) to-NP = NP-e (5)

e.to-NP(activity/concept) = NP-kk (6)

In a similar manner, rules for disambiguation of from, at, in, through, for, by, of and on are given below:

f. from-NP = NP-ninn (7)

g. from-NP(place/time) = NP-muthal (8)

h. at-NP(human) = NP-ne:rkk (9)

i. at-NP(place) = NP-il (10)-(11)

j. in-NP(time) = NP-il (12)

k. in-NP(currency) = NP-a:yi (13)

l. in-NP(natural) = NP-ath (14)

m. through-NP(place) = NP-kuti (15)

n. through-NP(thing) = NP-ute (16)

o. for-NP(human) = NP-ve:nti (17)

p. for-NP(time) = NP-e:kk (18)

q. NP1-by-NP2 (time) = NP2-o:te (19)

r. NP1 (place)-by-NP2 (place) = NP2-vaYi (20)

s. by-NP(quantitative) = NP-kaNakkin (21)

t. of-NP(human_group) = NP-il (22)

u. of-NP(concept) = NP-ulla (23)

v. verb(resultive) of NP(illness) = NP-a:l(24)

w. on-NP(concept/activity/thing) = NP-il (25)

x. on-NP(topic) = NP-parri (26)

The disambiguation methods discussed above are based on information about the type of verb and the noun in the relevant structure. In the above explained disambiguation rules, we can see that the case markers (genitive, dative, sociative, locative, instrumental, accusative and nominative) are handled appropriately based on the context. A sample example on how the above constraints can be formulated is illustrated in a tabular form (Table 1).

Table 1

Verb category			mov	mnt	
Verb form	main	main	main	main	main
Noun1 category	place	human	pst_hldr	human	activity
Noun2 category					
preposition	to	to	to	to	to
	e:kk	o:t	atuthe:kk	e	kk

7. RESULTS

All the examples illustrated above are taken from the output generated by the machine translation system developed by us. The performance of the system evaluated with a set of about 3000 sentences. The sentences are selected randomly from the corpus collected in the tourism and health domain from different articles and from the internet. The sentences are translated using the MT system developed and then evaluated using the five point scale manually. The ranking points are described below:

0 - No output provided by the engine concerned.

1 - The translated output is not comprehensible.

2 - Comprehensible after accessing the English text.

3 - Comprehensible with difficulty.

4 - The text is comprehensible.

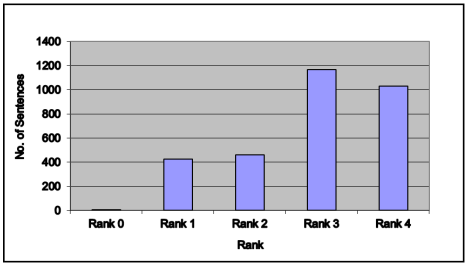

Figure 3. Analysis chart

Depending on the complexity of the sentence the output may contain all the possible alternate translations. This may vary from one sentence to hundreds of sentences. If the sentence is not grammatically correct and is extremely complex in nature, then the possibility of getting its translation is less. We have considered the first five translations for the evaluation. The evaluation done based on the methodology developed by C-DAC Pune. Fig. 4 below shows the sample output of the MAT system.

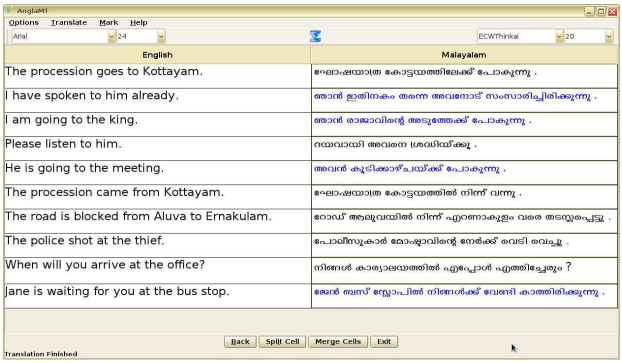

Figure 4. Sample output of the MAT system

The system will give more than 70% accuracy for the simple sentences and about 50-55% accuracy for the complex sentences. The accuracy of the system customized for Malayalam is at par with the system that is originally designed for Indo-Aryan language family.

Conclusions

In this paper we have examined the contexts for the multiple meanings of the prepositions in English and their mapping patterns in Malayalam. 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. The work is implemented in the English Malayalam Machine Aided Translation system using AnglaBharati Technology developed by IIT Kanpur.

However, it is difficult to find out the translation equivalence for some of the prepositions in English, when we try to do the translation process especially for machine translation. For comprehensive rule we need further finer semantic categorization of verbs and nouns.

Abbreviations/Acronyms

main: Main verb, **mov:** Motion verb, **mnt:** mental, **Noun1:** Noun after preposition, **Noun2:** Noun before preposition, **pst_hldr:** post_holder, **PAST**:Past, **FUT**: Future, **det**: Determiner, **QP**: Question Particle, **PP**: Postposition

Acknowledgement

We extend our sincere thanks to Prof. RMK Sinha and the people at IIT Kanpur, involved in the development of AnglaBharati Technology and to all the members of consortia involved in customizing AnglaBharati to different languages. The work is supported by the Ministry of Communication and Information Technology, Government of India sponsored project.

References

R.E. Asher, T.C. Kumari(1997), *Malayalam*, Routledge London and New York.

Suranad Kunjan Pillai(2000), *Malayalam Lexicon*, The University of Kerala, India.

A.R.Raja Raja Varma, *Keralapaanineeyam*(2000), D. C Books Kottayam-12, India

RMK Sinha, Anil Thakur(2004): *Pre/Post-positions Selection in Text Generation for Hindi and other Indian Languages for Translation from English*, Proceedings of International Symposium on Machine Translation NLP and TSS(iSTRANS-2004), pp 40-45, Tata Mc Graw Hill, New Delhi, India

R.M.K. Sinha(2004), *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), Pages 10-17, Tata Mc Graw Hill, New Delhi, India

R. M. K. Sinha(2005), *Machine Translation: AnglaBharati and AnuBharati Approaches*, Communications of CSI, India

R.M.K. Sinha(2004), *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, India

R. Ravindra Kumar, K G. Sulochana, V. Jayan(2011), *Computational Aspect of Verb Classification in Malayalam*, Information Systems for Indian Languages, International Conference, ICISIL 2011, India, Springer

Sudip Kumar Naskar, Sivaji Bandyopadhyay(2006), *Handling of Prepositions in English to Bengali Machine Translation*, Proceedings of the Third ACL-SIGSEM Workshop on Prepositions, pages 89–94, Trento, Italy.

Dr. S.Radhakrishnan Nair(2012), *An Introduction to Postpositions in Malayalam for POS Tagging*, Proceedings of the workshop on POS Annotation for Indian Languages: Issues & Perspectives, LDC-IL, CIIL, Mysore, India

Ravi Sankar S. Nair(2012), *Semantics of the Dative Case in Malayalam*, Language in India, Volume 12, India