A Strategy for Morphological Analysis and Synthesis of Malayalam

1. Introduction

Malayalam is Dravidian language with about 35 million speakers. Like other languages in Dravidian family, Malayalam is adhesive (“stuck-together”) in nature. The nouns and verbs in the language are inflected for plural making, case, tense, aspect and modality and also multiple words are combined to form complex words. Thus, indefinite number of word forms are possible in Malayalam. To develop a machine translator for such a morphologically rich language, we need a comprehensive lexicon, which listed all the word forms of all the roots. It is definitely the wastage of memory space. Every form of the word is listed which contributes to the large number of entries in such a lexicon. Data redundancy occurs when two roots following the same rule, is stored in lexicon. It fails to generalize relationship among different roots that have similar word forms. Linguistic generalization is necessary if the system is to have the capability of understanding an unknown word. Morphological analysis is the primary step for predicting the syntactic and semantic categories of unknown words. Morphology is seen as ‘the study of words that are formally and semantically related’. The study of morphophonemic changes in the word formation of Malayalam can be accomplished by Morphological analysis. Identification of generalized rules for relationship between different suffixes in words is possible by this analysis and that can be used in Morphological generator for synthesis of Malayalam words.

2. Morphological Analysis

The term morphology comes from ancient Greek (morphe) and means shape or form. The general definition of morphology is “the study of form or pattern”, i.e. the shape and arrangement of parts of an object, and how these “conform” to create a whole. In linguistics, a morpheme is the smallest component of word, or other linguistic unit, that has semantic meaning. Morphology is the study of morphemes and their arrangements in forming words. The morphological analyzer segment a word to a sequence of morphemes (root/stem and affixes), tag the part-of-speech (POS) of those morphemes and identify the morpho-syntactic relation between them. It provides the grammatical information depending upon the word category. For nouns it will provide gender, number, and case information and for verbs, it will be tense, aspects, and moods. The general format of morphological analysis of Malayalam is word = root/stem+suffix

For Example (1) Rama called Sita
rAman_ sIwaye viliccu
rAman_ - rAman_+ Φ (noun, singular, nominative)
sIwaye - sIwa + e (noun, singular, accusative)
viliccu - vili+ ccu (verb, past)

The morphemes of a word cannot occur in a random order. The order in which morphemes follow each other is strictly governed by a set of rules called Morphotactics.

Noun + Number + Case as in kuTTikaLuTe
=> kuTTi+kaL_+uTe

Another phenomenon that is of concern...
is **Morphonology**. Morphophonology or Sandhi explains the mutations in spelling when morphemes concatenate. The mutations are of three kinds: additions, deletions and substitutions. These mutations occur at morpheme boundaries during concatenation. These mutations are facilitated by the context in which the morphemes concatenate

\[
\text{mar}aM + \text{kal}_- \Rightarrow \text{maraffal}_-
\]

### 2.1. Word Formations in Malayalam

Malayalam is having a complex morphological process in the sentence formation. A complete sentence can be combined to form a single word because of the agglutinative nature of the Malayalam language. Consider the sentence 'avanaviteyewwiyittuNtAvilla', here the words avan_, avite and ewwiyittuNtAvilla are combined to form a single word. A verb in Malayalam can have up to ten suffixes. Different parts of speeches used in Malayalam are verb, noun, pronoun, adjective, adverb, preposition, conjunction and interjection. Some of these are words which have literal meaning of its own (Root/Stem) and others are words of relation which has no individual meaning (Affixes). Noun, verb, adjectives, adverbs and pronouns belong to first category. Preposition, conjunction, and interjunction belong to second category. The affixes are joined with root words to make complex words.

#### 2.1.1. Verb Formations

Verb is a grammatical category, which takes tense, aspect and modular information with it. These three types of information are generally called TAM (Tense Aspect and Modality). Tense indicates the time of action of the verb. Verb in present tense denotes the action at the time of speaking or writing. Verb in past tense indicates the completion of an action. Verb in future tense is the action performed before the time of speaking or writing. Aspect expresses the nature of action described by the verb. Four types of aspects are there simple, continuous, perfect, and perfect continuous. A Continuous verb shows the action in progress and perfect verbs shows the action completed. The following table 1 shows how the root word eYuw is inflected by different tense and aspects.

<table>
<thead>
<tr>
<th>TENSES</th>
<th>ASPECTS</th>
<th>EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>simple</td>
<td>eYuwunnu</td>
</tr>
<tr>
<td></td>
<td>continuous</td>
<td>eYuwikkoNtirikkunnu</td>
</tr>
<tr>
<td></td>
<td>perfect</td>
<td>eYuwunnuNt</td>
</tr>
<tr>
<td></td>
<td>Perfect</td>
<td>eYuwikoNtirikkunnuNt</td>
</tr>
<tr>
<td></td>
<td>continuous</td>
<td>eYuwi</td>
</tr>
<tr>
<td>Past</td>
<td>simple</td>
<td>eYuwiayirunnu</td>
</tr>
<tr>
<td></td>
<td>continuous</td>
<td>eYuwukayAyirunnu</td>
</tr>
<tr>
<td></td>
<td>perfect</td>
<td>eYuwiyittuNtAyirunnu</td>
</tr>
<tr>
<td></td>
<td>Perfect</td>
<td>eYuwikontirikkunnuNtAyirunnu</td>
</tr>
</tbody>
</table>
Moods (Modality) – Mood is the mode or manner in which the action of verb is shown. Different kinds of moods are Indicative mood, Imperative mood and Subjunctive mood

Indicative mood – It shows a simple statement of a fact or question
(2) Rama writes a letter.
   rAman_oru kaww eYuwunnu.

Imperative mood – This expresses a command, exhortation or prayer
(3) Write this letter soon.
   I kaww pettenn eYuwU.

Subjunctive mood – It expresses a wish, purpose or condition
(4) If you write this letter now I can send it.
   niffal_I kaww ippOL_eYuwiyiruneffkil_enikk aw ayakkAmAyirunnu.

Voice - There are two special forms of verb called Active and Passive. In active voice object receives the action of verb Rama wrote a letter
(5)rAman_oru kaww eYuwi.
In passive voice subject receives the action of verb.
(6) The letter was written by Rama.
   kaww rAmanAl_eYuwappettu.

Causative verbs: A form of the verb that shows that someone or something caused the action in the verb to happen rather than doing it directly.
(7) John ran the horse
   JON_kuwiraye Oticcu (i.e. John did not run, the horse ran; John made the horse run).

Transitive and Intransitive verbs: A verb that needs a direct object to complete its meaning. The action in the verb passes over to affect its object. In contrast, the action in an intransitive verb is limited to the agent or subject.
Transitive: (8) Children fly the kites.
   kuttikal_pattaM paZappikkunnu
Intransitive: (9) He ran
   Avan_Oti

Table 2. Inflection of number and case on noun

<table>
<thead>
<tr>
<th>case</th>
<th>Affix</th>
<th>singular</th>
<th>plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominative</td>
<td>nil</td>
<td>kutti</td>
<td>KuttikaL_</td>
</tr>
<tr>
<td>Accusative</td>
<td>e</td>
<td>kuttiye</td>
<td>KuttikaLe</td>
</tr>
<tr>
<td>Sociative</td>
<td>Ot</td>
<td>kuttiyOt</td>
<td>KuttikaL_Ot</td>
</tr>
<tr>
<td>Dative</td>
<td>kk,n</td>
<td>kuttikk</td>
<td>KuttikaL_kk</td>
</tr>
<tr>
<td>Instrumental</td>
<td>Al_</td>
<td>kuttiyAl_</td>
<td>KuttikaL_AL_,</td>
</tr>
<tr>
<td>Gentive</td>
<td>ute,nZe</td>
<td>kuttiyute</td>
<td>KuttikaL_ute,</td>
</tr>
<tr>
<td>Locative</td>
<td>il_</td>
<td>kuttiyil_</td>
<td>KuttikaL_l_</td>
</tr>
</tbody>
</table>
2.1.2. Noun Formations

The word formation of nouns is mainly based on Case and Number. The number markers (singular and plural) together with seven case markers add suffixes to nouns to make different noun forms.

2.1.3. Pronoun Formations

Pronoun is a word used instead of a noun and 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 different forms of pronoun FAn_(I): enne, ennOt, ennAl_, enikk, enZe, ennil_

2.1.4. Adjective Formations

Adjective is a word that qualifies a noun. 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 and subject will have a major role in the formation of adjectives in Malayalam. The two different types of adjectives are Attributive adjective and Predicative adjective.

Attributive adjective:

(10) nalla AN_kutti
    Good boy

Predicative adjective:

(11) I AN_kutti nallavan AN.
    This boy is good.

3. Synthesizer for Malayalam

Morphological generator and lexicon plays the major role in development of Machine Translation system for Malayalam. Morphological generation is the reverse of morphological analysis. Morphological analyzer is designed to analyse the constituents of words and it segments the word into stem and inflectional markers. The morphological synthesizer combines a stem with its suffixes based on the syntax of a language. Morphological synthesizer requires all inflected word forms are to be organised into paradigms and which are defined by set of syntactic rules.

3.1. Paradigm

A paradigm is a complete set of related word forms associated with a root word. All inflected forms of verb, noun, adjective, adverb and pronoun are classified into certain paradigm class based on their morphological behaviour. For example kutti belongs to one paradigm class and amma belongs to another paradigm class in noun.

Example for inflection List:

kutti, kuttiKal_, kuttiiye, kuttikaLe, kuttiyOt, kuttikaLOt, kuttikkk, KuttiKal_kk, kuttiyAI_, KuttiKal_kitiiute, KuttiKalute, kuttiyiIL, KuttiKalI.l,

amma, ammamAZ_, ammaye, ammAmAre, ammayOt, ammAmArOt, ammayk, ammAmAZ_kk, ammAmArI_kk, ammAmArute, ammAyil_, ammAmAril_

In the above example two number markers (singular and plural) and seven case markers are the feature values for noun formation. Thus 14 word formations of a root word can be obtained from one paradigm. Any word with similar inflections as that of the word amma can be put in to same paradigm class. For example amma and gayika can belong to same paradigm class as both of them have same word formations.

Word formation for adjective ‘sunxaraM’ based on gender (masculine, feminine, neuter), number (singular, plural) and type (attributive, predicative) is given below:
Eg: sunxaranAya, sunxariyAya, sunxaran_ mArAya, sunxarikaLAda, sunxaramAya, sunxaranAya, sunxaran_, sunxari, sunxaran_ mAZ_, sunxarikaL_, sunxaraM, sunxaraZ_

In this way we can generalise the relationship between different roots having same word formation. Verbs and pronouns can also be classified like this by taking features which makes classification easier.

3.2. Postpositions

Postpositions in Malayalam are suffixes which occur immediately after nouns and establish some grammatical relations between the nouns and the verbs of sentences. Prepositions in English language will be transformed to post positions in Malayalam. The semantic distribution of a single preposition will be varying in different context due to the influence of nouns and main verbs that follow. Some of the prepositions in English can be directly translated and attached to the Malayalam nouns as postpositions, while some need special rules.

For Example
(12) Radha went with Gopi

rAWa gOpiyute kUte pOi.

In Malayalam translation the preposition with is replaced by kUte. But a case marker ‘ute’ is to be added to the noun gOpi. These type word formations are possible by adding some rules in Text generator.

3.3. Sandhi Rules:

Malayalam has the property of combining of adjacent words. There are some specific rules in Malayalam to join two words called sandhi rules. These rules are used for joining roots and suffixes. On the basis of the sound involved, Sandhi can be grouped into Vowel Sandhi,

<table>
<thead>
<tr>
<th>Sandhi</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel Sandhi</td>
<td>pana+Ola = panayOla</td>
</tr>
<tr>
<td>Vowel Consonant Sandhi</td>
<td>wAmara + kuLaM = wAmarakuLaM</td>
</tr>
<tr>
<td>Consonant Vowel Sandhi</td>
<td>kaNN+illa = kaNNilla</td>
</tr>
<tr>
<td>Consonant Sandhi</td>
<td>Nel_+maNi = nenmaNi</td>
</tr>
<tr>
<td>Lopa Sandhi( Elision)</td>
<td>kEttu + illa = kEttilla</td>
</tr>
<tr>
<td>Agama Sandhi( Augmentation)</td>
<td>wiru + ONaM = wiruvONaM</td>
</tr>
<tr>
<td>Dvitva Sandhi(Reduplication )</td>
<td>paNa+ petti = paNappetti</td>
</tr>
<tr>
<td>Adesha Sandhi(Substitution )</td>
<td>kaN_+nIZ_ = kaNNIZ_</td>
</tr>
</tbody>
</table>
Vowel Consonant Sandhi, Consonant Vowel Sandhi and Consonant Consonant Sandhi. Sandhi can also be classified on the basis of the change occurring, namely lopa (elision), agama (Augmentation), dvitva (Reduplication) and adesha (Substitution) sandhis. Lopa is that in which one of the sounds is lost, agama is that in which a new sound is added, dvitva is that in which one of the sounds geminates and adesha is that in which one of the sounds is displaced by gives an idea about suffixes and inflections in word formation of Malayalam. Morphologicl analysis segments a word and identifies the grammatical information. This information can be transformed into general rules for Malayalam Text Generation. Paradigm classification is one of the generalized methods for Text Generation.

4. Summary

Morphological analysis and synthesis play major role in Machine translation. In morphologically rich language like Malayalam word formation is more complex. Morphological analysis of Malayalam gives an idea about suffixes and inflections in word formation of Malayalam. Morphological analysis segments a word and identifies the grammatical information. This information can be transformed into general rules for Malayalam Text Generation. Paradigm classification is one of the generalized methods for Text Generation.