A RULE BASED APPROACH FOR ROOT WORD IDENTIFICATION IN MALAYALAM LANGUAGE

Meera Subhash[1,2], Wilscy. M[1], S.A Shanavas[2]
[1]Department of Computer Science, University of Kerala, Kerala
meera_subhash@yahoo.co.uk,wilsyphilipose@hotmail.com
[2]Department of Linguistics, University of Kerala, Kerala
drsasha2002@yahoo.com

ABSTRACT

Words are tools of life which is omnipresent in every language. All words in a language are unique having their own function and meaning. The syntactic and semantic knowledge about individual words can be encapsulated in a highly structured repository known as computational lexicon which is very essential for Machine Translation. For designing a computational lexicon, the first and foremost task is to identify the head words or root words in the language. The Root Word Identifier proposed in this work is a rule based approach which automatically removes the inflected part and derive the root words using morphophonemic rules. The system is tested with 2400 words from a Malayalam corpus to generate the linguistic information such as the root form, their inflected forms and grammatical category. The performance is evaluated using the statistical measures like Precision, Recall and F-measure. The values obtained for these measures are more than 90%.

KEYWORDS

Corpus, Computational lexicon, Morphophonemic rules, Root Word, Root word Identifier

1. INTRODUCTION

A computational lexicon plays an important role in Machine Translation since it is the place where all the information about the vocabulary of a language is recorded for proper usage. A Machine Translation system save enormous amount of human power and time for the translation of one language text into another when the source and the target languages have computational lexicons of their own. The proposed system developed a software system for automatic identification of root words with their grammatical properties. These linguistic information about the words are systematically stored as lexical entries in a computational lexicon.

The need for identifying the root form of a word is very important in Natural Language Processing. A root word can be taken as a key term for searching, indexing, translating etc. Statistical tools like frequency counter, concordance, keyword-in-context, n-gram etc. need root form of a word to know more about the vocabulary. For example, using a frequency counter, most frequently used word in the vocabulary can be found out. Also such information can help to predict the correct spelling of the word, if assuming that the most frequent use of a word is correct. Morphological analysers and generators also require root words. Lexicographers assign root word as the head word/lexical entry in the lexicon.

Malayalam is the official language of the state of Kerala situated in the southern half of west coast of India. Malayalam language is one among the 22 official languages of India and one among the four major languages of the Dravidian family [1]. The influence of other languages like Sanskrit, Tamil, Telugu, Tulu, Toda, Kota, Kodagu and Badaga is seen in phonemic,

DOI : 10.5121/ijcsit.2012.4313

morphemic and grammatical levels of language. Malayalam is a morphologically rich and agglutinative language. There is no distinction on upper and lower case characters.

Most of the words in Malayalam are occuring in its inflected form. For obtaining the root form of the words, the suffixes agglutinated with them are to be removed. Also the morphophonemic change (sandhi) occurring when a root word concatenates with a suffix should be analysed and generalized. Sandhi rules are phonological alternations that are triggered at junctures, at junctions of words or morphemes. Malayalam grammar has categorised sandhi rules into different types. According to its consonant-vowel pair based categorisation, there are *svara sanDhi* (svaraM~+svaraM~), *svara vyanjgana sanDhi* (svaraM~+vyanjanam), *vyanjgana svara sanDhi* (vyanjanam +svaraM~), *vyanjgana sanDhi* (vyanjganaM~ + vyanjganaM~). Here *svaram* is vowel and *vyanjganaM~* is the consonant [1]. The morphophonemic changes occurring in the end phonemes of the word and the initial phoneme of the suffix is used to derive morphophonemic rule.

Morphological analysers and parts-of-speech taggers developed for Malayalam attempted to find out the root form of the word. In our approach, no lexicon or dictionary is used. The surface structure of the words is studied using a corpus to derive the morphophonemic rule. The Root Word Identifier system uses these generalised rules to automatically identify the root words, the grammatical category and their inflected forms. A corpus is created for Malayalam language using documents from world wide web. 2400 words in this corpus are used to test the system. The results obtained are stored in a computational lexicon having the linguistic information about these words such as whether the word is a noun, pronoun, verb or postposition. It also gives the inflected forms of root word present in the corpus. Performance of the system is evaluated and obtained a high Precision, Recall and F-measure. Incremental development of computational lexicon can be attained by using the proposed system with richer corpus as the resource.

The paper is organized as follows. The related work done in this area is presented in section 2. Section 3 discusses about the methodology of Root Word Identifier, resource for this work and the linguistics analysis of Malayalam words. In the same section, morphophonemic rule is discussed. Section 4 gives the results and discussions about the work. The last section gives the conclusion about the work.

2. RELATED WORKS

The morphological analysis deals with the study of internal structure of words of a language based on its grammatical category. It is the process of segmenting a morphologically inflected word into its root word and its associated morphological components along with the features specifying the morphological structure [2]. Even though a full-fledged morphological analyser is not for Malayalam language, there are many attempts in this area, as discussed below.

Morphological analysis for Malayalam verbs using a hybrid approach (paradigm and suffix stripping method) is an attempt made to attain morphological generalisation of verbs [3]. There will be dictionary of lexical items of Malayalam, which contains lexical items, grammatical category and paradigm type. The program compares each inflected form. The verbs are categorized into 28 classes or paradigms based on the past tense marker. They identified around 1100 inflections of verb. Using the same hybrid approach, a Malayalam morphological analyser using Apertium Lttoolbox is developed at Language Technology Centre, Centre for Development of Advanced Computing (C-DAC), Thiruvanathapuram [4] as part of Machine Translation task. Lttoolbox is available with the Apertium toolkit, which is an open source shallow-transfer

Machine Translation system originated with in the project “Open-Source Machine Translation for the Language of Spain”. Lttoolbox can be customised to any language by including the required lexical dictionary [5]. It uses the FST approach for doing lexical processing. Certain other attempts using stochastic taggers like HMM are also in progress but they cannot give a high accuracy for Malayalam because the language is inflectionally rich and is relatively free-word order like Tamil.

Apart from these attempts, related works to attain the aim for developing full-fledged Machine Translation systems are also going on in this language engineering field. Developing Named Entity Recognizer system, Noun Phrase Chunkers, Computational Lexicon etc. for Malayalam language are on progress.

3. ROOT WORD IDENTIFIER METHODOLOGY

Malayalam words can occur in its root form, inflected form, derived form, compound form and in reduplicated form. Inflected words are formed by the affixation of grammatical features such a case, number, tense, aspect, mood etc. to the root word. The process of separating the affixes from an inflected word can provide the root of the word and its grammatical information. The root word should be the most basic form of a word that is able to convey a particular description, thought, or meaning. The definition given for root word is that *it is a real word that can make new words from root words by adding prefixes and suffixes* [6]. Here all forms which are affixed immediately after the root word is considered as suffix. No attempt was made to identify the prefix.

The Malayalam documents seen in the world wide web are collected and stores as Malayalam corpus. Corpus [7] is a large collection of written and/or spoken text samples available in machine-readable form, collected in a scientific way to represent the use of a language [8]. In this work, a corpus of 24,000 words in written form is used for linguistic analysis of Malayalam from different domains. Using the rule based approach, the rules that govern the suffixation are derived manually by analysing the words in the corpus. The Root Word Identifier system will separate the root words after removing the suffixes as per these rules.

Common grammatical categories for Malayalam are the noun, pronoun, verb, adverb, adjective, postpositions, indeclinables, clitics etc. In this work, the main grammatical categories such as noun, pronoun, verb, postpositions are analysed.

3.1. Noun Morphology

Nouns can occur in isolation or can take gender markers, plural markers, case suffixes, postpositions, clitics etc. It takes the form

W= noun root ± [plural suffix] ± [case suffix] ± [postpositions] ± [clitics] ± …

where W is any word having the properties of a noun. Some of them are shown in Table 1.

Table 1. Case markers in Malayalam with Example

Case	Marker	Example
Nominative	nil	mankan~
Accusative	-e	makane
Dative	-kkU/(n)U	makan
Sociative	-ootU	makanootU
Locative	-il~	makanil~
Instrumental	-aal~	makanaal~
Genitive	-ute/nte	makante

Plural forms of noun contain suffix '-maaR~' and 'kaL~'. But it also take some allomorphs 'ngngal' and 'kkaL~'. Apart from these case markers and plural suffixes there are suffixes which are agglutinated with nouns [9]. Some of them are allative marker (-ileekk), place locative suffix (-athth), optative suffix (-aakatee), reciprocity suffix (-tammil), sufficient suffix(-maththi) etc. Pronouns are those words which can be used instead of nouns. Since they are in free form, considering them as root word for easy analysis: njaan~ (I (personal pronoun, first person, singular)), nii (You (personal pronoun, second person, singular)), avan~ (He (Third person, remote, singular, Masc.) etc. Sixty two pronouns are identified manually.

3.2. Verb Morphology

The morphological structure of verb is complex. They are capable of taking tense markers. Verbs in Malayalam are not inflected for Person Number and Gender. All verbal forms in Malayalam both finite and non finite consists of verb stems followed by affixes which express various grammatical categories such as tense, aspect, mood, voice valency change [9][10][11] etc. Generally Tense is classified into past, present and future. Aspect as perfective, imperfective, progressive. Mood as indicative, interrogative, imperative, conditional, optative, debitive, potential. Two voices are active voice and passive voice. Valency change is classified into causative and passive.

Categorization of verbs [12] can be done according to the suffixes attached to verb forms. Identified fifty-one verb suffixes to retrieve the verbal forms of words. Some of them are listed below with suffixes and example

1. Past tense marker - /-njnju/, /-nnu/, /-RRu/, /-ththu/, /-thu/, /-i/, /-ccu/, /-Ntu/, /-ttu/
   Eg: paRanjnju (told), ezhuthi (wrote)
2. Present tense marker - /-unnu/
   Eg: varunnu (coming)
3. Debitive emphatic marker /-aNee/
   Eg: tharaNee (should give)

3.3. Postpositions

Robert Caldwell (1913), commented that every postpositions annexed to a noun constitutes, properly speaking, a new case. On the basis of above definition, postpositions are also considered here as markers. English language has prepositions instead of Malayalam postposition. Different types of classifications are given for postpositions according to their origin, morphological similarities such as particles, case indicators, co-ordinator, derived noun etc. Twenty-seven suffixes are identified as postpositional suffixes. Some of them are

1. Duration/Distance suffix - /-oolaaM~/
   Eg: raamanooLaaM~ (upto/till/ approximately raman)
2. Equality suffix - /-poole/
   Eg: kittiyepoole (like a child)

3.4. Morphophonemic Changes in Malayalam

The suffixes obtained from the manual analysis are grouped according to their initial phonemes. Group A contains all the suffixes starting with /-a/, Group AA will contain all the suffixes starting with /-aa/ and so on. A detailed analysis of the morphophonemic change occurring when the final syllable of the root word concatenate with initial phoneme of the suffix is carried out. They are generalized and morphophonemic rules are derived to identify the root form of the word [13].

3.5. Morphophonemic Rule Implementation for Root Word Identification

A word in the corpus is analysed from right to left. The first encountered suffix part is removed and if any link morph such as /in/, /u/ is present, they are also removed. The remaining part of the word is taken for further analysis. If the last syllable is /y/ or /v/, it is sufficient to chop them for obtaining the root form of the word. If the last syllable is /ththa/ convert this syllable to /-aM~/. In this way, a list of rules is there to generate root form of words. Most of the words are confined to these rules. Only limited exceptions are identified. Some of them are listed in Table 2. In these words, some are having final syllables similar to suffix ending. They are stored separately as exceptional words and considered them as root words.

Table 2. Example for list of exceptional words

Examples for list of exceptional word
skumaaR~
bil~
kaan~
Ooroo
Koccu

The suffixes agglutinated with noun and verbs are classified as noun suffix, verb suffix and dual functional suffix (suffixes which agglutinate with both). The NS, VS and DS as shown in Figure 1. contain these suffixes. These are the accessory files to the Root Word Identifier system. The words from the corpus are fed to the Root Word Identifier system and morphophonemic rules as discussed in section 3.5 are applied to the word to obtain the root words. The system iteratively identify the suffixes and word. At the end of the task, the root word is identified. The complete details of the words from the Malayalam corpus are stored in Computational Lexicon. Block diagram of the system is shown in Figure 1.

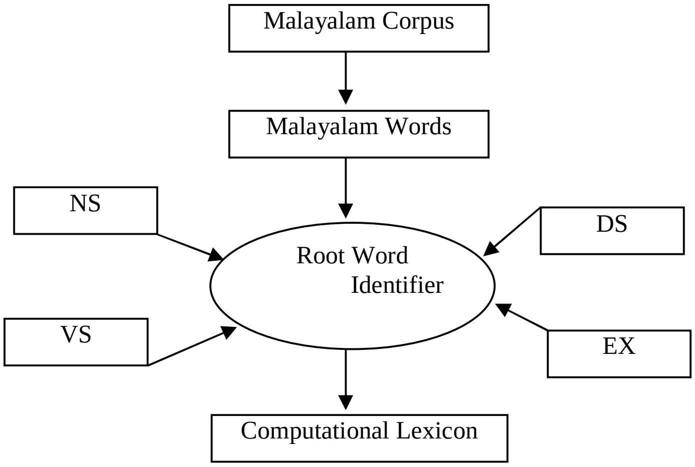

Figure 1. Block Diagram showing the input and output to Root Word Identifier

3.6. Algorithm for Root Word Identifier

```
begin
  step1: scan the word from right hand side
  step2: identify the suffix S and compare S with VS
       if present
         store the word in verb category
       else
         step3: compare S in NS,DS and EX
              if present
                step4: Store the word according to the corresponding grammatical category
                       then, Remove S and apply morphophonemic rules
         step5: repeat steps from 1 to 4 until the root word is encountered
    step6: Store the root word in computational lexicon with grammatical category
end
```

4. RESULTS AND DISCUSSIONS

The work is implemented with Practical Extraction and Report Language (PERL) [14] in Linux environment using Unicode supportive font.

Given a corpus as input, the Root Word Identifier system can generate a computational lexicon having the linguistic details about the words in the corpus such as the root form of the word, grammatical category of root word, inflected forms of the root word, all the suffixes agglutinated with the word, name of each suffix, words obtained after removing each suffix and their grammatical category.

An example which shows how the word *kuttikaLutekuute* meaning ‘with the children' is processed by the Root Word Identifier

[word from Malayalam corpus]
- + + +

[Phonetic notation]
kuttikaLutekuute- kutti + kaL~ + ute + kuute

[Description]
word- child+ plural suffix + genitive case suffix + suffix showing interior movement

[Linguistic information]
kuttikaLutekuute-Noun word
kuttikaLute - Noun word
kuttikaL~ - Noun word
kutti - Noun word – Root word

An objective analysis of the performance of the system is done using the statistical measures - Recall, Precision and F-measure.

1. Precision (P) = $\frac{tp}{tp+fp}$

2. Recall (R) = $\frac{tp}{tp+fn}$

3. F-measure= 2*P*R/(P+R)

where tp is the true positives, fp is the false positives and fn is the false negatives [15].

The typographical forms and foreign language words, which have less relevance in this work, are removed from the test corpus and the remaining 23,045 words were used to test the system. Root Word Identifier program will identify the words in the corpus and gives the root form of the word with its grammatical category such as noun, verb or dual functional. The system showed Precision, Recall and F-measure values as 95.42%, 95.05% and 95.22% when identifying root words. Compound words and reduplicated words are stored as a single root word. Only a small percentage of words were unprocessed and were processed wrongly. The system is having many advantages when compared with existing morphological analysers. Only a monolingual Malayalam corpus is required to develop a computational lexicon. So no need to have any lexicon or machine readable dictionary or manual assistance for collecting words.

5. CONCLUSION

As a part of developing a computational lexicon for Malayalam there arises a need for developing a Root Word Identifier which identifies the root form of the word. A detailed morphological analysis is carried out to study the underlying structure of Malayalam words. Morphophonemic rules are derived to obtain the root form of the word automatically with its grammatical category and inflected forms. This rule based approach with a larger corpus can contribute much to the development of a full-fledged computational lexicon.

REFERENCES

[1] Raja Raja Varma. A.R, (2000) “Keralapanineeyam”, D.C Books, Kottayam-12. India.

[2] Menaka, S. Vijay Sundar Ram and Sobha Lalitha Devi, (2010) “Morphological Generator for Tamil”, In Morphological analysers and Generators, (ed.) Mona Parakh, LDC-IL, Mysore, Pp 82-96.

[3] Saranya S.K, (2008) “Morphological analyser for Malayalam verbs”, Unpublished M.Tech Thesis, Amrita School of Engineering, Coimbatore.

[4] Vinod P.M, Jayan, and Sulochana K.G, (2011) “Malayalam Morphological Analyser: A hybrid approach with Apertium Lttoolbox” Proceedings of ICON -2011: 9th international Conference on Natural Language Processing, Macmillan publications, India. Pp 219-224.

[5] Forcada M. L, B. Bonev, Ortiz S. Rojas et. al., (2010) “Documentation of the Open-Source Shallow-Transfer Machine Translation platform Apertium”, Available on-line at: http://xixona.dlsi.ua.es/~fran/apertium2documentation.pdf.

[6] http://www.vocabulary.com/

[7] Dash N.S, (2005) “Corpus Linguistics and Language Technology”, Mittal Publications, NewDelhi

[8] Biber Douglas, Conrad Susan & Reppen Randi, (1998), “Corpus Linguistics Investigating language structure and use”, Cambridge University

[9] Seshagiri Prabhu, M., “Vyakaranamitram,” (4thed.), Kerala Sahithya Academy, Thrissure, 1983

[10] Saranya S.K., “Morphological analyzer for Malayalam verbs”, Unpublished M.Tech Thesis, Amrita School of Engineering, Coimbatore, 2008

[11] Shanavas, S.A., “Structure of a Computational Lexicon of Malayalam”, Unpublished PhD Thesis, Jawaharlal Nehru University, New Delhi, 1996

[12] R.E. Asher, T.C.Kumari, (1997) “Malayalam”, Routledge London and New York.

[13] Meera Subhash, Wilscy M, Shanavas S.A, (2011) “A statistical identification of English loan words in Malayalam documents”, Proceedings of ICON -2011: 9th international Conference on Natural Language Processing, Macmillan publications, India, Pp 91-95.

[14] Hammond.M, (2003) “Programming for Linguist: Perl for Language researchers”, Blackwell Publishing, UK.

[15] Manning D, Schutze Hinrich, (1999) “Foundations of statistical natural language processing”, MIT Press, Cambridge, London.

Authors

Meera Subhash M.Sc PGDCA MCA Mphil
Researcher Scholar (Computational Linguistics)
Department of Computer Science
Kerala University, Thiruvanathapuram

Dr. Wilscy. M
Head of the Department
Department of Computer Science
Kerala University, Thiruvanathapuram

Dr. S.A Shanavas
Ass. Professor and Hon. Director (TRCML)
Department of Linguistics
Kerala University, Thiruvanathapuram