மாநாட்டுக் கட்டுரைகள்
CONFERENCE PAPERS

11

உலகத் தமிழ் இணைய மாநாடு

INTERNATIONAL TAMIL INTERNET CONFERENCE 2012

Morphological Analyzer for Classical Tamil Texts: A Rule-based approach for Case Marker

R.Akilan (akilan.rp@gmail.com)
Programmer, Central Institute of Classical Tamil, Chennai.
Prof E.R.Naganathan (ernindia@gmail.com)
HOD, Department of Computer science and Engineering, Hindustan University, Chennai.
Dr. G.Palanirajan (matrixpalani@gmail.com)
Associate, Central Institute of Classical Tamil, Chennai.

Abstract

This paper describes the works to build a Morphological Analyzer for Classical Tamil using Rule-based approach. Morphology is the study of internal structure of the word. Morphological analysis is a process of segmenting words into morphemes and a process of analyzing the word formation. Morphological analyzer is a tool for any type of Natural Language Processing work. It is a computer program which takes words as input and produces its grammatical structure as output. It identifies and segments the words and assigns the grammatical information. Capturing the agglutinative structure of Tamil words by an automatic system is a challenging job. This paper is going to reveal a rule-based approach for case marker.

Introduction

Natural Language Processing (NLP) is a computerized approach to analyze the text based on a set of theories and set of technologies. And, being a very active area of research and development, the basic objective of Natural Language Processing is to facilitate human-machine interaction through the means of natural human language.

Morphological analysis of a word is the process of segmenting the word into component morphemes and assigning the correct morphosyntactic information. For a given word, a morphological analyzer (MA) will return its word and the word class along with the other grammatical information depending upon its word class. MA returns all possible parse for a given word, without considering the context. MA is a very essential for languages having rich inflectional and derivational morphology such as morphologically rich languages like Dravidian languages.

Morphological Analyzer is a vital tool in NLP applications. In morphological rich languages, as there are multiple affixation, the finer grammatical information which helps in building efficient NLP applications, can be obtained only from Morphological Analyzer. Morphological Analyzer is required in most of the applications such as information extraction, QA system, machine translation and spell checker. There are several approaches attempted for Morphology for Tamil, We present a methodology for morphological analysis of Tamil, a morphologically rich, in this paper. We present a rule-based method for Morphology for Classical Tamil, particularly case marker.

Tamil morphology

Tamil belongs to the Dravidian family of languages. It is one of the Classical Languages. It is a verb-final language and has a relatively free word order; it is an inflectional language. Agglutination is another feature of the language. Tamil morphology is characterized as agglutinative or concatenative, i.e., Words are formed by successfully adding suffixes to the root word in series. When suffixes attach to the root several morphophonemic changes take place. The orders in which suffixes attach to a root form determine the morphosyntax of the language and the various changes that take place when a suffix attaches are called the morphophonemics.

Challenges in Morphological Analyzer for Classical Tamil

Tamil is a classical language which belongs to the Dravidian language family. Tamil literature has existed for over two-thousand years. The morphological structure of Classical Tamil is quite complex since it inflects to person, gender, and umber markings and also combines with auxiliaries that indicate aspect, mood, causation, attitude etc in verbs. A single erb root can inflect for more than two-thousand word forms including auxiliaries. Noun root inflects with plural, oblique, ase, postpositions and clitics. A single noun root can inflect for more than five hundred word forms including ostpositions. The root and morphemes have to be identified and tagged for further language processing at word level. he structure of verbal complex is unique and capturing this complexity in a machine analyzable and generatable format s a challenging job. The formation of the verbal complex involves arrangement of the verbal units and the interpretation f their combinatory meaning. Phonology also plays its part in the formation of verbal complex in terms of orphophonemic or *sandhi* rules which account for the shape changes due to inflection.

Methodology

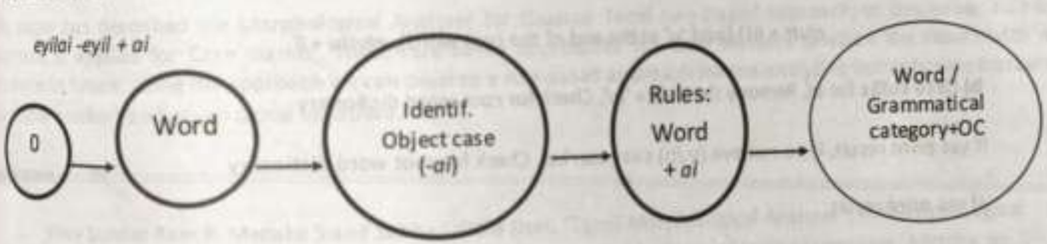
Finite State Automata (FSA)

FSA is a model of behavior composed of a finite number of states and transitions between these states. FSA is an abstract device used for recognizing simple syntactic structures or patterns. An automata is normally depicted by directed graph, called State Diagram and it is also represented in a tabular form as State Table. An FSA, as a string processing device, accepts strings as input and decides if the structure is correct, that is, it either accepts or rejects the string. From a mathematical perspective it is regarded as a function, mapping a set of string to the set {Accept, Reject}.

Case Marker

Case system links a noun phrase and other parts of the sentence through inflection markers, or a word which may be called as a adposition including preposition and postposition

1. Object case (*-ai*)



2. Instrumental case (-o*ṭu*)

koṭiyoṭu-koṭi + oṭu

3. Associative case (*ku, -kku, -akku, -ukku*)

4. Dative case (iṉ)

karumpiṟku-karumpu + iṉ + ku, cāttaṟku-cāttaṉ + ku
maturaikku-maturai + (k) ku, tamakku-tam + akku
avaṉukku-avaṉ + ukku

5. Genitive case (*-atu*)

eḷḷatukuppai- eḷ + atu + kuppai, paṭaiyatukuḷām - paṭai + atu + kuḷām

6. Locative case (*-kaṇ*)

poruṭkaṇ - poruḷ + kaṇ, malarkkaṇ - malar + kaṇ

7. Nominative case (0)

Rules for Case Markers

Rule 1: Case suffix

The morphosyntax of case suffix may be summarized as

Root + {*ai*} { *āl*} { *iṉ*} {il} { *ku* } { *kku*} {*akku*} {*ukku*} {*atu*} {*kaṉ*}

a) After segment the case marker, if remaining word end with consonant later add '-u'.

The following example illustrate the inflection of a case suffix of '*iṉ*'

aṉpiṉ -> check the root word dictionary

aṉp + { *iṉ*} (if 'no' remove the suffix *iṉ)*

aṉp + { *iṉ*} (add 'u' at the end of the root word), *aṉpu* + *iṉ*

The following example illustrate the inflection of a case suffix of '*il*'

eḻuttil -> check the root word dictionary

eḻutt + {il} (if 'no' remove the suffix *il),*

eḻutt + {il} (add 'u' at the end of the root word), *eḻuttu* + *il*

b) Case suffix for *ai*, Remove the suffix '*ai*', Check for root word dictionary

If yes print result, If no remove (y /n) case marker, Check for root word dictionary

If yes print result

Ex. tōḻiyai -> *tōḻiy ai* {remove the suffix *ai }*

tōḻiy {check dictionary / no word}

tōḻi y {remove y the last phoneme} = *tōḻi+y+ai* {Result}

c) To split a suffix from the word. The word consists of double letters between words and suffix namely ṉ, ḷ, l, ṉ which are segmented separately.

(*kaṉ, kaḷ, col, taṉ*)

Ex. *kaṉṉukku -> kaṉ+ṉ+ukku, kaḷḷukku -> kaḷ+ḷ+ukku, collukku -> col+l+ukku*

taṉṉai -> taṉ +ṉ+ai

Analysis

The morphological analysis identifies root and suffixes of a word. Generally rule-based approaches are used for morphological analysis which are based on a set of rules and dictionary that contains root words and morphemes. In rule-based approach, a particular word is given as an input to the morphological analyzer and if that corresponding morpheme or root word is missing in the dictionary then the rule-based system fails. Here each rule depended on the previous rule. So if one rule fails, it affects the entire rule that follows. In the course of testing of the rule, certain inconsistencies and lapses in recognizing certain word, First have been found nineteen thousand and nine hundred Classical Tamil root word corpus has been taken for analysis of that corpus is applied the case markers rules. The careful appraisal and study of the words is conducted to identify and overcome the lapses by incorporating certain amount of data into the root word to enhance the coverage and the overall performance of the morphological tools. The following problems are also well noted

1. Some words end with y (ய்) and v (வ்) which is a part of the word. Do not operate their *Sandhi* rules. For examples *añcāy, kāy, pāy, vāy* and *tev.*
2. In some words which require the doubling of the end consonant before add the suffix. For example
   kaṉṉai (kaṉ+ṉ+ai), maṉṉai (maṉ+ṉ+ai)
   tammai (tam+m+ai), emmai (em+m+ai)
   collai (col+l+ai), pallai (pal+l+ai)
   neyyai (ney+y+ai), meyyai (mey+y+ai)
   poṉṉai (poṉ+ṉ+ai), viṉṉai (viṉ+ṉ+ai)
3. Moreover -u (*ṭu* and *ṟu*) adding rule is a role between word and suffix. Here it is doubling the final consonant after removing the case marker. If the check before the dictionary For example
   nāṭṭai (nāṭu+–ṭṭ+ai), vīṭṭai (vīṭu+–ṭṭ+ai), āṟṟai (āṟu+–ṟṟ+ai), kūṟṟai (kūṟu+–ṟṟ+ai)
4. -u adding rule plays an important role between word and suffix. Here doubling of a variant consonant occurs after removing the case marker. For example

mārpai (mārpu+–rp+ai), cālpai (cālpu+–lp+ai)

Conclusion

This paper has described the Morphological Analyzer for Classical Tamil rule-based approach; in this paper rule-based approach is applied for Case marker. These rule-based approaches for case markers produce the result with more accuracy. In future, using the approach we can develop a rule-based approach for the analyzing not only case markers but also other markers and grammatical variations.

Reference

1. Vijay Sundar Ram R, Menaka S and Sobha Lalitha Devi, "Tamil Morphological Analyser", in "Morphological
Vijay Sundar Ram R, Menaka S and Sobha Lalitha Devi, Central Institute of Indian Languages, Mysore, pp. 1–18.
Analysers and Generators", (ed.) Mona Parakh, LDC-IL, Central Institute of Indian Languages, Mysore, pp. 1 –18.
2010
2. Anand Kumar M, Dhanalakshmi V, Soman K.P, "A Sequence Labeling Approach to Morphological Analyzer for
Anand Kumar M, Dhanalakshmi V, Soman K.P, "A Sequence Labeling Approach to Morphological Analyzer for
Tamil Language", International Journal on Computer Science and Engineering Vol. 02, No. 06, 2010
3. Beesley, Kenneth R. Arabic, "Finite-State Morphological Analysis and generation", Proceedings of the 16th
International Conference on Computational Linguistics, Vol.1.Copenhagen, Denmark. Page No: 89-94, 1996.
4. Viswanathan, S., Ramesh Kumar, S., Kumara Shanmugam, B., Arulmozi, S. and Vijay Shanker, K. "A Tamil
Morphological Analyzer", Proceedings of the International Conference On Natural language processing ICON
2003, Central Institute of Indian Languages, Mysore, India. pp. 31–39, 2003.
5. Parameshwari K, "An Implementation of APERTIUM Morphological Analyzer and Generator for Tamil", an e
journal of Language in India (www.languageinindia.com), May 2011 Special Volume: Problems of Parsing in
Indian Languages, May 2011

CONFERENCE BOOK BROUGHT TO YOU BY

NEW HORIZON MEDIA
Enlightening minds
Enriching lives

An imprint of
NEW HORIZON MEDIA PRIVATE LIMITED
177/103, Ambals Building, Lloyds Road
Royapettah, Chennai - 14. INDIA
Phone : 91 + 44 + 42009603
www.nhm.in, Email : support@nhm.in

உத்தமம்
INFITT
உலகத் தமிழ்த் தகவல் தொழில்நுட்ப மன்றம்
International Forum for Information Technology in Tamil

www.infitt.org

CO-SPONSORED BY

SRM
UNIVERSITY