International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Morphological Analyzer for Malayalam: A Review on Various Approaches

Anagha .N[1], Vidya .M[2]

[1]Student, Department of Computer Science &Engineering, Vidya academy of Science & Technology, Thalakottukara, Thrissur, Kerala, India

[2]Assistant Professor, Department of Computer Science &Engineering,
Vidya academy of Science & Technology, Thalakottukara, Thrissur, Kerala, India

Abstract: ***A morphological analyzer is the automated implementation of human ability to analyze a language which always returns a morpheme with the suffix associated with it. Since Malayalam is an agglutinative language with large number of inflections, an efficient morphological analyzer is required which uses the best possible method. There are different approaches for analyzing a language morphologically. This paper describes such methods like paradigm approach, suffix stripping, finite state automata, corpus based method etc and their limitations and advantage over one another.***

Keywords: Morpheme, paradigm, suffix stripping, finite state automata, corpus

1. Introduction

For any language to be analyzed properly, it is necessary to understand that language by the machine. In fact it is the biggest challenge in natural language processing. Morphology is the identification, analysis and description of structures of a given language's morphemes and morphological analysis is the process of studying the structure and formation of words. Morphological analyzer segments the word into morphemes. A morpheme is the simplest meaning bearing word in a language. A word can be divided into two classes stem and affixes. In Malayalam language the affixes simply means the suffixes. Stem is usually the part with a proper meaning and suffix adds different aspects of a word. Generally languages are classified into three classes. They are isolating(Chinese), agglutinative (Dravidian) and inflectional(Latin).

2. History of Morphological Analysis

The history of Natural language processing into four phases with distinctive concerns and styles [1].The first phase of work in NLP is lasting from the late 1940s to the late 1960s.It was driven by Machine Translation, the second phase is from the late 60s to the late 70s which is flavored by Artificial intelligence. The third phase which is to the late 80s deals with grammatico logical, while fourth which lasts to the end of century focused on lexical and corpus data. Among these stages, only the last one focuses on the morphological aspects and many methodologies were used to implement the morphological analyzer during this stage.

3. Related Works so Far

Many works have been done in morphological analysis in natural language processing so far. In a paper named An affix stripping morphological analyzer for Turkish [6] which has been published in 2004 a new methodology is proposed for doing the analysis of Turkish words with an affix stripping approach and without using any lexicon. The rule-based and agglutinative structure of the language allows Turkish to be modeled with finite state machines (FSMs). In contrast to the previous works, in this study, Finite state machines are formed by using the morphophonemic rules in reverse order. Corpus Linguistics is another approach that aims at investigating and analyzing large collection of text samples. For ages this approach has been used in a number of research areas. It generally includes a large collection of machine readable data of actual language including literature and non-literature text samples.

Theoretically, corpus is (C)apable (O)f (R)epresenting (P)otentially (U)nlimited (S)elections of texts. Inflectional morphological analyzer for Sanskrit[5],suggests a Sanskrit morphological analyzer that identifies and analyzes inflected noun-forms and verb-forms in any given sandhi-free text. An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modeling Finite State Networks [7] describes Morphological ambiguity is a major concern for syntactic parsers, POS taggers and other NLP tools. For example, the greater the number of morphological analyses given for a lexical entry, the longer a parser takes in analyzing a sentence and the greater the number of parses it produces. Xerox Arabic Finite State Morphology and Buckwalter Arabic Morphological Analyzer are two of the best known, well documented, morphological analyzers for Modern Standard Arabic (MSA). In a work for a Rule based Morphological Analyzer for Classical Tamil Text [9], the analyzer identifies root and suffixes of a word and assigns its grammatical categories. Some of these approaches are used for Morphological Analyzers.

More accurate results are generated by using the rule based approaches. The rule based approach used for morphological analysis which are based on a set of rules and dictionary that contains root words and morphemes. A Novel Approach for English to Dravidian Language Translation System [10] developed a statistical machine translation system for English to South Dravidian languages like Malayalam and Kannada

by incorporating syntactic and morphological information. A bilingual corpus was used to extract data for translating from one language to another.

4. Approaches on Malayalam Morphological analyzer

Morphological analyzer is the automated system of a human ability to analyze and understand language. Malayalam is a language in the Dravidian family and which shows the characteristics of agglutinative language class . There has been various methodologies to analyze this language morphologically.[3]Though they were not yet able to produce maximum efficiency, neither those methods can lead to a fully fledged morphological analyzer. The different approaches are:

A. Paradigm based approach

- Paradigm approach can be implemented by using the Apertium lttoolbox A root word can have different forms and a paradigm defines those various forms of a given stem. In this method a word is provided with the paradigm it follows.
- For a morphologically rich language like Malayalam, this is a very good approach. The paradigms cannot be chosen arbitrarily, they should be certain specific groups which are necessary for stating the syntax of the language.
- A paradigm defines all different forms of a word. The certain paradigm groups are created based on the inflectional categories. So before implementation, one has to sort out all possible inflected forms of a lexical item.
- The linguist must provide the possibly longest tables of word forms. Each of these tables covers a set of root words in a language.
- Generally in Malayalam words are classified into nouns, pronouns, noun locatives, verb, adverb, adjectives and postpositions and these categories are grouped into certain paradigm types depending on their morphophonemic behavior.
- Sample inflections for the different categories are : maraM-marangaL-maratte-marattooT-marangale…var-vannu-varunnu-varuM-vannirunnu-vannuvenkil sundara-sundaramaaya-sundaramaayat,avan-avane-avanmaare-avanaal,muulaM-muulamaaNu-muulamuLLat-muulavuM…
- The paradigm number and inflection list for each paradigm class can be extended to any length. It is very important that the inflections list must contain same number of inflections.

In the work done based on the paradigm approach, the authors listed the following data [3] in Table I .

Table 1: List of Paradigms

	Noun	Pronoun	Noun Locative	Verb	Adverb	Adjectives	Postpositions
No of paradigms	30	28	6	42	7	14	5
No of inflections in each paradigm	1085	1085	40	842	12	20	40

A. Suffix Stripping Based Approach

Suffix stripping [3] is another powerful method used for morphological analysis. This approach makes use of a root word dictionary, for valid stem identification and a suffix dictionary which contain almost all possible inflections of nouns/verbs of Malayalam language .This approach also uses a trained set of sandhi rules which are generated based on paradigm classes.

A Malayalam word is a combination of stem and suffixes. The advantage of suffix stripping method is that even if the input is not found in the main database, we can analyze the inflected word and find its root word and suffix separately. Once the suffix is found in the suffix list, that suffix is stripped off and a corresponding sandhi rule is applied to find the stem. This approach not only depends on a single lookup table, but two dictionaries with stored rules.

Nouns are linguistic categories which can take cases with them and also it provides 'PNG'informations i.e. Gender number and person. But verb is considered as a grammatical category which takes tense aspect and modularity, which is denoted as 'TAM".

Samples of such markers are:
Adjectve markers-karutha,cheriya,chuvanna,pazhaya etc
Verb markers-kku,nnu,ntu,unnu etc
Postposition markers-poole,,kaal,kontu,kurichu,veenTi etc.

The suffix is a key term for an agglutinative language like Malayalam since there is no prefix or circumfix in the language. The properties of Malayalam language is used to form the morphotactics and morphophonemic rules. Since Malayalam has a tendency to combine one or more word/words with a root word, it is highly complex to strip suffixes. When a word is undergoing a stripping process, Malayalam language requires a lot of morphophonemic changes in the word formation in each step of the process corresponding sandhi rules have to be applied. The speed of the process is good compared to the other existing methods.

B. Hybrid Method

Hybrid method is a combination of both paradigm based approach and suffix stripping method. This approach combines the advantages of both paradigm and suffix stripping methods and minimizes the limitations. The method combines categories whose morphophonemic behavior is similar. The algorithm identifies the suffix first and then the root word by applying sandhi rules. There will be a collection of all possible suffixes that can be found attached to the stem. This is used for the suffix identification comparing the suffix with the list of all possible suffixes, the rule of longest

matching suffix is used. The inflection list is checked first here. This feature is the advantage and disadvantage of this approach. If the word is in the inflection list, then the searching process will be faster than in the root word dictionary. Sometimes it is not efficient. In hybrid method, the input words which have to be analyzed are first checked in the inflection list. If the input is found in the list with same features, then it identifies the valid stem and suffix.

D. Finite State Automata

A finite-state automaton is a device that can be in one of a finite number of states. In certain conditions, it can switch to another state. This is called a transition. When the automaton starts working (when it is switched on), it can be in one of its initial states [11]. There is also another important subset of states of the automaton: the final states. If the automaton is in a final state when it stops working, it is said to accept its input. The input is a sequence of symbols. A string is said to be accepted if it reaches the final state of FSA else it is rejected. The advantage of this method is that it models language and supports mass data processing,but it is not a good method for morphological analysis.

E. Two level Morphology

This method describes phonological alternations in terms of finite state automata [2].It makes use of fully parallel rules instead of usual cascaded rules. The rules here are considered to be complete statements. This method is mainly depends on three key features. First is the mode of application of rules. The rules are applied parallel,not in sequential order. They are symbol to symbol constraints. Second is about the constraints. It can refer to lexical context either to the surface context or to both simultaneously. Third is that here both morphological analysis and lexical lookup are performed. No works done based on this method till date.

F. Finite State Transducers

Transducers are automata that have transitions labeled with two symbols. One of the symbols represents input, the other - output. Transducers translate (or transduce) strings. It is actually a modified version of FSA. Here FST which is two tape automation combine lexicon, orthographic rules and spelling variations in FST to develop a morphological analyzer.

G. Corpus Based Approach

Corpus [8] is a collection of text in a particular language. In order to make the corpora more useful for doing linguistic research, they are often subjected to a process known as annotation. An example of annotating a corpus is part-of-speech tagging, or *POS-tagging*, in which information about each word's part of speech (verb, noun, adjective, etc.) is added to the corpus in the form of *tags*. In morphological analysis this raw corpus is provided as input and the generates segments of words of the input provided. This segments obtained is similar to the morphological segments. This is combined approach of Corpus based as well as Paradigm.

5. Comparison of Approaches

Equalize the length of your columns on the last page. If you are using *Word*, proceed as follows: Insert/Break/Continuous. Below is the comparison table of different approaches. In all approaches we can see the advantages and disadvantages [2].From this table,it is clear that an efficient morphological analyzer requires the combination of different approaches.

Table 2: Comparison of Different Approaches

Approaches for morphological analysis	*Advantages*	*Disadvantages*
Paradigm approach	The use of paradigm Approach provides more efficient results	i.efficiency rely on the content of paradigm ii.a single word can posses different aspects and features
Suffix stripping approach	Easy to deal with since inflection list is not larger as root word dictionary	I .Behaves too badly in certain exceptions. ii. result is limited to the lexical categories.
Hybrid approach	Gives good results if in suffix is found in inflection list	Poor system if a root word is used
Finite state automata	i.Language modelling ii. Mass data processing	i. Not a good method for morphological analysis
Two level morphology	I .Linear representation ii.sequential ordered rules are used	Suitable for linear orthographic input only
Finite state transducers	i.it is used for word identification ii . it isnot recursive in behaviour	i.Hard to implement ii.precision in result is low
Corpus based approach	Produces improved result	Result depends on the corpus content

6. Conclusion

In Natural Language processing morphological analysis play a vital role. Malayalam is a language which shows heavy amount of agglutination. In all approaches discussed the accuracy depends upon the suffix list or the inflections listed. If a developer failed to list a possible suffix in Malayalam language, the accuracy of the system gets affected. If we are using the suffix stripping method, the accuracy lies on the splitting part. Also a hit occurs only when splitted suffix is found in the pre-tagged suffix list. All approaches have such issues independently. From the above survey, it is clear that a single approach is not sufficient to develop an efficient morphological analyzer in Malayalam. And also all the works done till now mainly concentrated on the noun and verb classes. There are even more categories for verb such as mood and aspect. Adjectives, pronoun, postpositions etc are another area to explore using these methods.

References

[1] A. Natural language Processing:a historical review,Karen Sparck Jones Computer Laboratory,University of Cambridge , sparckjones@ cl.cam.ac.uk,http://www.cl.cam.ac.uk,October 2001

[2] Morphological Analyzer for Malayalam: A Literature Survey Aswani Shaji Sindhu L Dept.Of Computer Science and Engineering College of Engineering Poonjar. International Journal of Computer Applications (0975 -8887) Volume 107 –No.14, December 2014.
[3] Morphological Analyzer for Malayalam A Comparison of Different Approaches by Jisha P Jayan,Rajeev R R ,S. Rajenndran,IJCSIT International Journal of Computer Science and Information Technology,vol 2,No 2,December 2009,pp 155-160.
[4] Morphological Analyzer for Malayalam: Probabilistic Method Vs Rule Based Method Rinju O.R.,, Rajeev R. R, Reghu Raj P.C., Elizabeth Sherly International Journal of Computational Linguistics and Natural Language Processing Vol 2 Issue 10 October 2013 ISSN 2279 0756
[5] Inflectional Morphology Analyzer for Sanskrit, Girish Nath Jha, Muktanand Agrawal, Subash, Sudhir K. Mishra, Diwakar Mani, Diwakar Mishra,Manji Bhadra, Surjit K. Singh. Special Center for Sanskrit Studies, Jawaharlal Nehru University, New Delhi, 110067
[6] An affix stripping morphological analyzer for Turkish Glen Eryiit and Eref Adal Dep. of Computer Engineering, Istanbul Technical University 34469 Ayazaa, Istanbul, Turkey proceeding of the IASTED international conference. Artificial intelligence and application, Feb 16 18,2004,Inbruck,Austria
[7] An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks Mohammed A. Attia School of Informatics, The University of Manchester mohammed. attia@postgrad.manchester.ac.uk
[8] http://en.wikipedia.org/wiki/Text_corpus
[9] Morphological Analyzer for Classical Tamil Texts: A Rulebased approachR.Akilan and Prof. E.R.Naganathan(Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore) Programmer, Central Institute of Classical Tamil, Chennai. akilan.rp@gmail.com
[10] Anovel approach for english to south Dravidian language statistical machine translation system by Unnikrishnan p,Antony p j and Soman K P computational engineering and networking department, amrita vishwa vidyapeetham,coimbatore, tamil nadu, india (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2749-2759
[11] http://galaxy.eti.pg.gda.pl/katedry/kiw/pracownicy/Jan.Daciuk/personal/thesis/node12.html.