International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 8 Issue 03 Series. I || March 2019 || PP 73-77

Study on Morphological Analyzer and Generator for Malayalam

Remya Sivan
Atria institute of technology

Abstract: *Morphology is the branch of linguistics that deals with formation of words. Morphological Analyzer is the program which takes word as input and gives root word or morpheme and grammatical structure of the word as output. And this process is known as Morphological Analysis. Morphological generator is the reverse of analyser ie Take particular morpheme or root word and grammatical information as input and generate particular word as output. Morphological analyser and generator used in many natural language processing applications such as Machine translation, pos tagging, spell check. In this paper we discuss about various existing approaches for MAG (morphological analyser and generator) and also compare these approaches w.r.t Malayalam languages*

Keywords: *corpus based, FSA, FST, stemmer, DAWG, Paradigm based approach*

Date of Submission: 17-02-2019 Date of acceptance: 03-03-2019

I. Introduction

The two approaches to morphology,the branch of linguistics that deals with formation of words are Analysis and Synthesis .Analysis is the process of find out the root words or morphemes from the given word. The program which does this process is known as morphological Analyzer. Synthesis is formulating word from given root word and grammatical structure. And the program is known as morphological generator.Morphemes, which are resulting from morphological analysis, are meaningful words. There are two types of morphemes. Free morphemes and bond morphemes. Free morphemes are the words which are not possible to break down again and againex cat. A bound morpheme is the one which is not having independent existence ,comes along with free morphemes ex: un,es,s. Two classes of morphemes are root and affixes. Root morphemes are the meaningful word to which affixes are added ex: run. Affix morphemes are morphemes which can be added to roots to change their meaning ex: ex: ing. The resulted word will be running

There are two types of morphology Inflection and Derivational morphology. Inflectional morphology is the study of the process of word formation where new words are forming without changing the meaning of root word only grammatical structure of root word changes. For ex,root word is ‘run’ and affix is ‘ing’ the resultant word is running. Derivational morphology is the study of the process of word formation where new words are forming with new meaning and new grammatical structure. For ex The root word is’ go ‘and affix is ‘at ‘the resultant word is goat.[10][11]

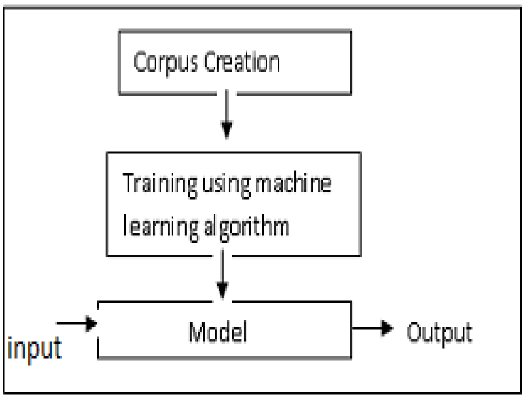
The rest of the paper is organizes as follows in section 2 we present several approaches existing in morphological analysis. In section 3 we provide a comparative study among the existing morphological analysis approaches. In section 4 we are doing survey on Malayalammorphological analysers. Finally, section 5 has the concluding remarks.

II. Approaches to Morphological Analysis

In this section we are discussingdifferent approaches to morphological analysis–corpus based, rule based,stammer based, paradigm based ,finite state automata ,finite state transducers based, two level morphology ,DAWG approach and hybrid method[9]

1. Corpus based

This method is implemented using machine learning algorithms .here exist a corpus, a large collection of annotated words. The machine learning algorithms generate the rules required for morphological analysis from this corpus .the model generated in this way can be used to identify the roots and grammatical structure of a input word [1].The disadvantage of this approach is the time consumption in corpus creation process and performance of the analyser purelydepends on the features and size of the corpus.



2. Rule based

Rule based approach is based on set of rules and dictionary that contains root and morphemes. Morphological analyser will try to match the input word to the rule which is already defined. If the match happens, system will return the root and the grammatical structure of the input word from the dictionary .If the input word is not present in dictionary, the system fails.[2] The main disadvantage of Rule based approach is every rule is depends on the previous rule. So one rule fails it will affect the entire rule that follow it. Another disadvantage is the same word will match to different set of rules and this will lead to ambiguity. Performance of the system purely based on set of rules and Dictionary content [6].

3. Stemmer based approach

Stemming is the process of reduction of the inflected or derived words into stem or root. The program who does this process is known as stemmer .There are four different approaches for stemming [14][3]

•Table lookup: Make a table of all word forms and their stems. Use B-tree or Hash table to lookup the table for the input.

•Successor Variety: Determine the word and morpheme boundaries based on the distribution of letters. The successor variety of a given string is the number of different characters follows the current character in a text. Once the successor variety of a given word is determined, use this information to segment the word .There are many segmentation methods, cut off method, peak and plateau and entropy method etc.

•N-gram: After identifying unique diagrams of given word pair, calculates similarity measure using Dice's coefficient S/A+B.

•Affix removal: remove the prefix/suffix from the word form and return stem. The two affix removal techniques are simple removal and longest match. In longest match method make use of two dictionaries, root dictionary and suffix dictionary and set of rules. Root dictionary has all possible root words and suffix dictionary has all possible suffixes. According to the rule,select the longest possible string from the word and remove it. Repeat this process until no characters are there to remove .Then comparison of the word will be done with root dictionary. Porter algorithm is one of the widely used suffix stripping algorithm

4. Paradigm based approach

Words from language are categorized into word classes like noun, pronoun, verb, adverb, adjective, pronouns, and prepositions. Each word class will be classified into various paradigms.For any word class the all the words which are having same inflection patterns are grouped into same paradigm. So in the same word class it can be resulted number of paradigms based on inflectional patterns. Each paradigm has root word and all possible inflectional forms of the word. The number of inflectional words should be same for all the paradigms in a particular word class.[15]

This approach makes use of root table .Root table has all the roots with paradigm number. Paradigm based approach well suited for agglunative languages. Most of the Indian languages make use of this approach. Disadvantage of paradigm based approach is result based on the content of paradigm table. Some words can be included under different paradigms since their meaning changes according to the context.

5. Finite state automata

A finite automaton is a formal system. It remembers only a finite amount of information. Information represented by its state. State changes in response to inputs. Rules that tell how the state changes in response to inputs are called transitions .Finite automata defined as a 5 tuple system A=(Q,q0,Σ,δ,F) .Q is a set of finite states including start state final states and normal states.q0 is the start state . Σ is the set of finite no of inputs. δ is the transition function that shows when automata moves from one state to next state in response to input ,F is the finite set of final states.Given a sequence of inputs, start in the start state and follow the transition from each symbol in turn. Input is acceptedif you wind up in a final (accepting) state after all inputs have been read.FSA's modeled from inflectional rules. Finite state automata best known as recognizers because they accept set of input strings.[17]

6. Finite State transducers

Finite state transducers are advanced version of finite state automata with two tapes one input tape and one output tape. It reads the input from input tape and writes the output to output tape and generate the relationships between the input string and output string and vice versa. Because of their bidirectional property FST can be used as analyser and generator. FST can be defined as 6 tuple System A=(Q,q0,Σ,τ,δ,F) .Q is a set of finite states including start state final states and normal states.q0 is the start state . Σ is the set of finite no of inputs. τ is the finite set of output symbols .δ is the transition function that shows when automata moves from one state to next state in response to input.F is the finite set of final states [17]

7. Two level morphology

It is a general computational model for word form recognition and production by Kimmo Koskenniemi.Two level morphology based dictionary and set of two level rules.Two important components of two level morphology are finite state automata component and dictionary component. Roots and affixes are listed in dictionary and Rule changes are encoded in automata[8]. Automata component of two level automata consist of several two headed finite automata. Lexical and surface levels of two level morphology are connected by these several finite state automata. Because of these parallel automata there is no intermediate stage between lexical and surface level. Dictionary component lists the morphemes and possible sequence of morphemes within the word [16].

8. Directed acyclic word graph (DAWG) approach

Directed acyclic word graph (DAWG) is a very good data structure for lexicon representation.DAWG can be used for both morphological analyser and generator. It is language independent; it does not utilize any morphological rules. There are two different types of DAWG a. Deterministic and Non deterministic [18] .In deterministic DAWG there is no more than one outgoing link from the node with same symbol. But here need of stop character required to mark the end of the string. In nondeterministic DAWG there is a possibility to have more than one outgoing link from the node with the same symbol. If we plan to make analyser and generator nondeterministic DAWG is the best choice. Deterministic DAWG we can use for the systems which perform as either analyser or generator with better response time.

9. Hybrid Approach

Hybrid approach is a combination of both paradigm and suffix stripping method. It is very good approach for developing morphological analyser for morphologically rich and agglutinative languages. Accuracy of the system depends upon the morphological dictionary and suffix list used [5].

Comparison study of different approaches

S.no	Approach	Advantage	Disadvantage
1	Corpus based	Produced improved results if the language have well organized corpus	Corpus creation is time consuming .Performance of the system depends on the feature and size of the corpus
2	Rule based	Produced improved results if the system has well organized dictionary and rules	Every rule is depends on the previous rule. So one rule fails it will affect the entire rule that follow it.Same word will match to different set of rules this will lead to ambiguity..
3	Stemmer based	Fast and Can handle removal of double letters Ex Getting--→get it allows a particular query term to match documents containing any of the morphological variants of the term	Stemming can't relate words which have different forms based on grammatical constructs. Ex :Root form of word Better is Good But stemmer would fail to resolve it
4	Paradigm based approach	Suitable for agglutinative language Improved result with well-defined paradigm table	Result based on the content of paradigm table. same wordmay possess many features
5	Finite state automata	Most significant tool for computational linguistics.	FSA can be used as either MA or MG not together

		It is a system modelling technique There are number of ways to implement FSA It support mass data processingAbility to process multiword tokens in the analyser	Sometimes no of states can be unmanageable
6	Finite state transducers	Same machine can be used for MA and MG	Since it creates many transitions it is hard to implement
7	Two level morphology	Does not depend on any rule compiler or finite state algorithm Rules are declarative so it allows analysis and generation within the same grammar There is no intermediate representation or results Rule orderings did not matter	There is no clean way for a rule to make reference to non-graphemic properties
8	Directed acyclic word graph (DAWG) approach	Language independent Use single data structure for both analysis and synthesis Faster than two level morphological analyser	large size of structure record for each key can’t be determined uniquely

MAG for Malayalam

- Morphological Analyzer for Malayalam Using Machine Learning

V.P. Abeera1, S. Aparna1, R.U. Rekha1, M. Anand Kumar1, V. Dhanalakshmi1,K.P. Soman1, and S. Rajendran2 Amrita Vishwa Vidyapeetham, Ettimadai, Coimbatore developed a corpus based morphological analyser for Malayalam using machine learning technique SVM.this system require one corpora with linguisticallyinformation. These morphological information are automatically extracted from the annotated corpora.

- Rule Based Morphological Analyzer for Malayalam Nouns: Computational Analysis of Malayalam Linguistics Jancy Joseph, Dr.BabuAnto developed a morphological analyser for Malayalam Nouns using Suffix-Stripping Method in a Rule-Cum-Dictionary Based Approach. They have included testing corpus of 500 nouns with 40 inflections of each noun. The performance of the system could be improved by number of nouns in the testing corpora and consider more no of rules. Input word checked with the word dictionary to identify the stem word given and return the noun. If it does not match with the dictionary strip off suffix from right side and classify suffix to its correct class.re initialize word without suffix and repeat the process until Dictionary finds the root word
- A suffix stripping based morph analyser for malayalam language:Rajeev R, Rajendran N, and Elizabeth Sherly proposed a morphological analyser for malayalm using suffix stripping metyhod with sandhi rules.
- Malayalam Noun and Verb Morphological Analyzer: A Simple Approach Nimal J Valath1, NarsheedhaBeegum developed a morphological analyser for Malayalam using a rule based-pattern matching hybrid approach with the help of a pre-tagged language lexicon. They have used combined approach of paradigm and suffix stripping method.
- Implementation of Malayalam morphological analyser based on hybrid method:vinod pm and bhadranvk language technology centre CDAC thiruvananthapuram developed a morphological analyser based on hybrid method combination of paradigm and suffix stripping method.the performance of the system based on morphological dictionary and suffix list used. Their dictionary has 54240 entries. They have used ltoolbox for morphological analysis.Ltoolbox make use of FST for lexical processing.
- Morphological analyser for Malayalam verbs:SaranyaSk proposed a hybrid approach of paradigm and suffix stripping method for verb analysing .Sandhi rules has been used to identify the correct stem. There are 3200 verbs in their dictionary. They have created 28 paradigms based on past tense marker.
- Morpheme boundary identification using letter successor variety by Indu Joseph Thoppil and Elizabeth Sherly. This analyser based on Letter successor variety which is based on stastical co-occurrence measures and contextually similar words

III. Conclusion

Morphological analysis is the integral part of natural language processing.morphological analyser is the prerequisites for pos tagging and machine translation system. In this paper I have studied different techniques for MAG and tried to understand existing MA or MAG for Malayalam.

References

[1]. Morphological Analyzer for Malayalam Using Machine LearningV.P. Abeera1, S. Aparna1, R.U. Rekha1, M. Anand Kumar1, V. Dhanalakshmi1,K.P. Soman1, and S. Rajendran2 Amrita Vishwa Vidyapeetham, Ettimadai,

[2]. Rule Based Morphological Analyzer for Malayalam Nouns: Computational Analysis of Malayalam Linguistics Jancy Joseph, Dr.BabuAnto

[3]. A suffix stripping based morph analyser for malayalam language:Rajeev R, Rajendran N, and Elizabeth Sherly

[4]. Malayalam Noun and Verb Morphological Analyzer: A Simple Approach Nimal J Valath1, NarsheedhaBeegum

[5]. Implementation of Malayalam morphological analyser based on hybrid method:vinod pm and bhadranvk language technology centre CDAC

[6]. Morphological analyzer for Malayalam verbs:SaranyaSk

[7]. Development of Prototype Morphological Analyzerforthe South Indian Language of KannadaT.N. Vikram and Shalini R. UrsInternational School of Information Management, University of Mysore, Manasagangotri,Mysore-570006, Karnataka, India {shalini,vikram}@isim.ac.in

[8]. A two-level morphological analysis of English Lauri Karttunen and Kent Wittenburg

[9]. Speech and Language ProcessingAn Introduction to Natural Language Processing, Computational Linguisticsand Speech RecognitionDaniel Jurafsky and James H. Martin

[10]. Study of stemming algorithmsSavithaKodimalaUniversity of Nevada, Las Vegas

[11]. What is Morphology? MarAronoffKirstenFudeman

[12]. A Straightforward Approach toMorphological Analysis andSynthesisK. Sgarbas, N. Fakotakis, G. Kokkinakis

[13]. Morpheme boundary identification using letter successor variety by indu joseph thoppil and Elizebathsherly

[14]. Automatic Language-Specific Stemming in Information Retrieval by john A. Goldsmith,Derrick Higgins,Svetlana Soglasnova

[15]. A paradigm-based morphological analyser by fredkarlson department of general linguistics,university of helinski

[16]. Two-level model for morphological analysis by koskenniemi,department of general linguistics,university of helinski

[17]. Finite State Morphology Alexander Fraser & Liane Guillou {fraser,liane}@cis.uni-muenchen.de CIS, Ludwig-Maximilians-UniversitätMünchen Computational Morphology and Electronic Dictionaries SoSe 2016 2016-05-09

[18]. A Straightforward Approach to Morphological Analysis and Synthesis by Kyriakos N. Sgarbas, Nikos D. Fakotakis, George K. Kokkinakis

Remya Sivan" Study on Morphological Analyzer and Generator for Malayalam" International Journal Of Engineering Science Invention (Ijesi), Vol. 08, No. 03, 2019, Pp 73-77