Morphological Analyzer for Malayalam: A Literature Survey

Aswani Shaji
Dept.Of Computer Science and Engineering
College of Engineering Poonjar

Sindhu L
Dept.Of Computer Science and Engineering
College of Engineering Poonjar

ABSTRACT
Natural Language processing deals with analysing, generating and understanding the languages that human generally use. Morphological Analyzer is an important part in Natural Language Processing. Morphological Analyzer always returns the morpheme and its associated grammatical structure. This paper describes about the different techniques in morphological analyzer and different implementations of morphological analyzer in Malayalam.

General Terms
Morphological Analysis, Natural Language Processing, Malayalam.

Keywords
Suffix Stripping, Paradigm, Finite State Automata, Two Level Morphology, Finite State Transducer, Directed Acyclic Graph, Corpus Based Approach.

1. INTRODUCTION
Morphological analysis[1] is the study of the structure and formation of words. The basic unit is called as morpheme. Morpheme is smallest units which have meaning. There are two classes for morphemes which is stem and other is affixes. Stem is always the meaning bearing word and affixes are the pieces which add extra meaning to the stem. Morphological structure is just one way of grouping languages.

Usually there are three classification and they are isolating languages (eg.Chinese), Agglutinative languages (eg. Turkish) and Inflecting languages (eg. Latin).

The Agglutinative language like Malayalam is rich in inflections, which require complex procedures to extract its inflections and grammatical information. Inflection[2] means the modification that we are done to a word depending upon the tense, mood, voice etc.

Morphological analyzer indentifies the stem and affixes of the word provided.

In Malayalam most of the lexical items, like nouns, verbs, etc are inflected heavily, giving information such as person, number, tense, and mood respectively. These inflections can be nested in many cases. The nesting increases the difficulty in identifying the morphological features.

2. TECHINQUES USED FOR MORPHOLOGICAL ANALYSING
There are different approaches for morphological analysis. Some of the different morphological analysing techniques are explained below:

2.1. Finite State Automata
Finite state automata [3] is a device that always accepts or rejects a string. It uses regular expressions as its input. A string is said to accepted if it reaches the final state of FSA otherwise it is said to be rejected Regular expressions are powerful tools for text searching. FSA can be used to represent morphological lexicon and recognition

2.2. Two Level Morphology
The distinction that linguists can make between morphotactics and morphophonemics is the basis for this model.

It describes phonological alternations in finite-state terms. It is having fully parallel rules instead of usual cascaded rules, rules could be thought of as statements that directly constrain the surface realization of lexical strings.Logical constraints are easier to understand but they don't have opaque interactions. This constraint-based model is independ on, composition or any other finite-state algorithm or on a rule compiler and called it as two-level morphology. This methodology is based on three ideas[5]:

- Rules are applied in parallel and not sequential and they are symbol to symbol constraints
- The constraints can refer to the lexical context either to the surface context or to both at the same time.
- In tandem , morphological analysis and lexical lookup are performed

2.3. Finite State Transducers
FST is a version of FSA[3]. It is used to computationally represent lexicon. It can be done by accepting the principle of two level morphology.

The two level morphology represents a word as a correspondence between lexical level and surface level. An FST is a two tape automaton. We combine lexicon, orthographic rules and spelling variations in the FST in order to build a morphological analyzer .Tamil morphological analyser make use of this approach side by side with paradigm.

2.4 Corpus based Approach
Corpus is a collection of text in a particular language. 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.

2.5 DAWG (Directed Acylic Word Graph)
DAWG is having different application, which can be efficiently used for lexicon representation along with string matching. This structure can store finite strings in a compact way, also it can takes the advantage of common affixes in the string. This method has been successfully implemented for Greek language by University of Partas Greece. Non-deterministic DAWG data structure can be used for both morphological analysis and generation and if we only need one function then it is better to use deterministic DAWG with

a better response time. This method is effective for Indian Languages too. This approach does not utilize any morphological rules or any other special linguistic data.

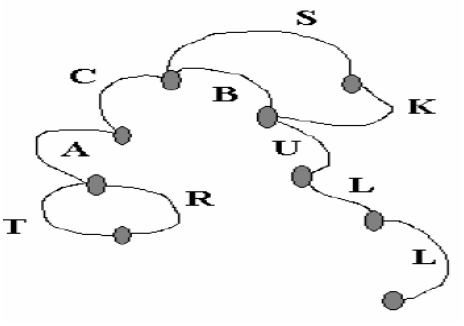

Fig 1: Directed Acyclic Word Graph

2.6 Paradigm Approach

A paradigm provides all the forms of a given stem along with its feature structure. This approach is more appropriate for inflectionally rich languages.

The linguist or the language expert provides different tables of word forms which cover the words in a language. These each table covers a set of roots, which along with the affixes generate the word forms. Most of the Indian Language make use of this method. Based on paradigms a program is developed which generates add delete string for analyzing. Paradigm approach rely on findings that the different types of word paradigms are based on their morphological behaviour.

Words are classified like nouns, verbs, adverbs, etc. Each category will be divided into certain types of paradigms based on their morphophonemic behaviour. For example noun maraM (tree) belongs to a paradigm class is different form adhyaapika (teacher) which belongs to a different paradigm class as they behave differently morphophonemically.

2.7 Suffix Stripping

In languages like Malayalam, which is highly agglutinative, a word is formed by adding suffixes to root or the stem. This method is used for words analysis. In Malayalam the words do not posses prefixes and circumfixes. But morphologically highly complex words exist , which are formed by continuously adding suffixes to the stem. Suffix stripping method is used for words in languages with more suffix

In the paper by Dinesh Kumar ,Et.al[9],the morphological analyzer accept the input text and it is transliterated to an intermediate representation which is stored as a file. This file used while traversing the FSA. then each sentence is fed to tokenizer and each token is checked in the dictionary. After identifying the root the it go for searching the affix based on the morphtactics.

In paper by V P Abeera,Et.al[10],is a two step process. Initially it is morphological data creation for Malayalam language and the next stage is implementation of morphological analyzer. In implementation phase include three sub phases. In pre-processing stage surface form is converted into sequence of units which is considered as the input. In segmentation of morpheme phase the words input is segmented according to morpheme boundary and then trained model predicts the label of each segments. Here SVMTool is used for morphological analysing and as it is using machine attached to the single language. If a suffix is identified then we can find the exact stem by removing the single and applying sandhi rules.

The general format is

Word stem + suffixes

The main grammatical categories in Malayalam are Noun and Verb. Stem is either can be verb or noun .

This suffix stripping method requires a stem dictionary, in order to identify a valid stem, a suffix dictionary which contains all the possible suffixes, morphotactics rules and morphophonemic rules or the sandhi rules.

Nouns can have case markers as their suffixes. Normally verbs inflect for tense, aspect and mood. Thus the verbal forms are stripped into suffixes which denote different aspects, moods and tenses.

3. RELATED WORKS

In a system developed by Nimal J Valath, Et.al [6], they developed morphological analyzer for nouns and verbs using combined approach of paradigm and suffix stripping method. They initially transliterated Malayalam to English, which help to find the occurrence of affixes easily. This available input is then checked whether it is present in the dictionary of Malayalam words .Following each suffix is extracted from the given word. Then obtained root word is checked in the dictionary. If it is the retransliteration of the word from English to Malayalam is done. But this system is developed for only verbs and nouns which is considered to be its drawback.

In another system, developed by Rajeev R R,Et.al [7],they used the following methodologies: they make use of syntactic constraint ordering ,finite state automata for more appropriate syntactic parsing, the morphological parsing is done with finite state transducers followed by noun and verb analysis following Sandhi identification.

In a system, developed Vinod P M,Et.al[8],make use of Lttoolbox for, morphological analysis ,generation ,lexical processing etc. A dictionary contains blocks like alphabet definition, definition of symbols, paradigm etc. Paradigm approach contains groups which are having similar inflection pattern. A Malayalam morphological clustering has noun paradigm and verb paradigm. In suffix Stripping module, they separate the word and the suffix from the surface form, here they make use of suffix stripping algorithm.
learning during training ,SVM modules are learned from training corpus.

In a memory based system which is developed in Dutch[11],algorithms can learn mapping if a sufficient number of instances of these mappings presented to them. Here each word form is considered and each of them is analyzed and created task instance using windowing method. Using windowing transforms each word form is divided into as many instance it has as its letters and find the root word. In case of having similar labelling they are joined to give ambiguous class.

In [12] deals with the major issues we face during the development of Malayalam Morphological Analysis. Some of them are Multiple Suffix problem, Handling Abbreviations and proper nouns, Identification of chunks, encoding issues etc

In[13] deals with a comparison of the morphological analyzers developed in Malayalam using rule based approach and probabilistic approach. This paper says that the most accurate method among two is rule based approach.

Also the paper by Meera Subhash .Et.al [14] deals with extracting the root words and remove inflections using rule based approach using morphophonemic rules

In a system developed by Jisha P Jayan Et.al [15] make use of suffix stripping method for morphological analysis.

In a system developed by Anand kumar M, Et.al [16] make use of machine learning in Tamil language which is similar to [10] .

Table 1: Comparison of Different Morphological Techniques

Morphological Techniques	Related Work	Method	Advantage	Limitation
Finite State Automata	i. Rajeev R R,Et.al [7] ii. Dinesh Kumar ,Et.al[9]	A string is said to be accepted if it reaches the final state of FSA else it is rejected	i.Language modelling ii. Mass data processing	i. Not a good method for morphological analysis
Two level Morphology	No work done	Model for generation and recognition of word forms	i. Linear representation ii. Sequential ordered rewrite rules are used	i.Designed to work with linear orthographic input only
Finite State Transducers	i. Rajeev R R,Et.al [7] ii. Dinesh Kumar ,Et.al[9]	Advanced form of Finite State Automata	i. It is not recursive in nature ii. It is used for word recognition	i.Difficult to implement ii.Less calculation ability
Corpus Based Approach	i.V P Abeera,Et.al[10], ii. Anand kumar M , Et.al[16]	Corpus is used for morphological Analysis	i. Use of Corpus provides improved result.	i.Results depends on the content of corpus used
Directed Acyclic Word Graph	No work done	It is a structure used for lexicon representation along with the string matching	i.Language Independent ii.Do not use any morphological rules	i. Can move only in the direction specified
Paradigm Approach	i.Nimal J Valath, Et.al [6] ii.Vinod P M,Et.al[8]	Paradigm defines words along with its feature	i. Use of Paradigm also provide improved result	i. Results depends on the content of paradigm ii.Same word may possess different feature
Suffix Stripping	i.Nimal J Valath, Et.al [6] ii.Vinod P M,Et.al[8] ii. Jisha P Jayan Et.al [15]	Removing suffix from word to get the stem word	i. Simpler to maintain	i. Provide poor result in certain exception ii.Result is limited to the lexical categories in case of exception

4. CONCLUSION

Natural Language processing deals with processing the languages humans use. In Natural Language processing morphological analysis play a vital role. Morphological analysis deals with analysing individual words.

This paper discussed the different morphological analyzers and the different techniques used for morphological analyzer. Most of the works are done in noun and verbs. . As most of the works are done in noun and verbs we can able to extend the work in pronoun, adverb, adjective, etc

As Malayalam is an agglutinative language which is rich in inflections and it is more complex procedure to find out the stem and its affixes from a given word. So far now a fully fledged morphological analyzer is not developed in Malayalam

So we can conclude that the most appropriate method is rule based method for Malayalam morphological analyzer.

5. REFERENCES

[1] Speech and Language Processing-An Introduction to Natural Language processing, Computational Linguistics and Speech Recognition , Jurafsky, Daniel and Martin, James H (2002)

[2] http://en.wikipedia.org/wiki/Inflection

[3] http://galaxy.eti.pg.gda.pl/katedry/kiw/pracownicy/Jan.Daciuk/personal/thesis/node12.html

[4] Saranya S K ,Morphological Analysis of Malayalam Verbs

[5] http://www.mit.edu/~6.863/fall2012/lectures/lecture3bw.pdf

[6] Nimal J Valath, Narsheedha Beegum, Malayalam Noun and Verb Morphological Analyzer:A Simple Approach, International Journal of Software & Hardware Research in Engineering,ISSN No:2347-4890 Volume 2 Issue 8,August 2014

[7] Rajeev R,R, Rajendran N And Elizabeth Sherly,A Suffix Stripping Based Morph Analyser For Malayalam Language

[8] Vinod P M,Jayan V,Bhadran V K,Implementation of Malayalam Morphological Analzer Based on Hybird Approach, Language Technology Centre CDAC

[9] Dinesh Kumar,Gurpreet Singh Josan,Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey, International Journal of Computer Applications (0975 – 8887)Volume 6– No.5, September 2010

[10] V P Abeera,S Aparna,R U Rekha,M Anand Kumar,V Dhanalakshmi, K P SomanMorphological Analyzer for Malayalam Using Machine Learning

[11] Antal Van Den Bosh,Walter Daelemans ,Memory Based Morphological Analysis

[12] Vinod P M, Jayan V and Bhadran V K, Issues in Development of Malayalam Morphological Analyzer", in 978-1-4673-2272-0/12 IEEE, 2012.

[13] Rinju O.R, Rajeev R. R, Reghu Raj P.C., Elizabeth Sherly, Morphological Analyzer for Malayalam: Probabilistic Method Vs Rule Based Method

[14] Meera Subhash, Wilscy. M, S.A Shanavas,A Rule Based Approach For Root Word Identification In Malayalam Language

[15] Jisha P.Jayan,Rajeev R R,Dr. S Rajendran,Morphological Analyser and Morphological Generator for Malayalam - Tamil Machine Translation

[16] Anand kumar M, Dhanalakshmi V, Rajendran S, Soman K P,A Novel Approach to Morphological Analysis for Tamil Language.