MORPHOLOGICAL ANALYSER FOR MALAYALAM LANGUAGE USING FINITE AUTOMATA

Liji S K
Department of Computer Science
Sullamussalam Science College,
Areekode, Kerala, India
liji.s.k@gmail.com

Abstract- Natural language processing (NLP) is an important field of Computational Linguistics and Artificial Intelligence. It deals with analyzing and understanding the languages that people generally use. Morphological Analyzer is an important part of Natural Language Processing. It is essential for question-answering, information retrieval, machine translation and spell checking, speech recognition etc. This paper proposed the implementation of a morphological analyzer for Malayalam Language using Finite Automata. The system returns the morpheme and its associated grammatical structure of Malayalam documents.

Keywords-Natural Language Processing (NLP), Morphological analyzer, Finite State Automata (FSA)

INTRODUCTION

Morphological analyzers are essential for any type of Natural Language processing works. Most of the Indian languages are agglutinative in nature and inflection varies from language to language. In Malayalam most of the words, such as nouns, verbs, adjectives, and adverbs are inflected heavily, giving information such as a person, number, tense and mood respectively. These inflections may be nested in many cases. The nesting increases the difficulty in identifying the morphological features. It has a productive morphology that allows the creation of complex words which are highly ambiguous. Malayalam has a productive morphology that allows the creation of complex words which are highly ambiguous. Due to the complexity, the development of a Morphological analyzer for Malayalam is a tedious and time-consuming task. Depending upon its word category, the morphological analyzer returns its root word along with its grammatical structure. For nouns, it will provide gender, number, and case information and for verbs, it will be tense aspects, and modularity.

LITERATURE REVIEW

Morphological Analysis is the first step in any natural language processing system. In India, there are many languages spoken throughout the country. The processing of those languages needs Morphological Analyzers for each language. Morphological Analyzer for other Indian languages, such as Hindi, Kannada etc. is already available. For Malayalam, researchers are still going on to develop a complete and efficient Morphological Analyzer.

Sunil R et al [1] proposed a strategy for Morphological Analysis and Synthesis of verbs in Malayalam for English Malayalam machine translation. They used a rule-based classification mechanism. A Word synthesizer module is used for the classification and formation of verbs.

Vinod PM et al [2] developed a Malayalam Morphological Analyser based on a hybrid approach. They combined the methodologies of both paradigm and suffix stripping approaches. The main objective of their study is to help the language students as well as common people.

Jisha P.Jayan et al.[3] use a bilingual dictionary for Malayalam and Tamil which consist of the root/ stem of the Finite-State Transducers based morphological analyzers.

Nisar Habash et al [4] proposed a morphological Analyzer for Egyptian Arabic. The system extends existing resources, the Egyptian colloquial Arabic lexicon and follows part of speech tagging. It also normalizes multiple orthographic variants to a conventional orthography.

In a paper, Rinju et al [5] deal with a comparison of the morphological analyzers developed in Malayalam using a rule-based approach and probabilistic approach. This paper says that the most accurate method among the two is the rule-based approach.

A system developed by Nimal J Valath Et.al [6], developed a morphological analyzer for nouns and verbs using a combined approach of paradigm and suffix stripping method.

The morphological analyzer for many of the South Indian languages of the Dravidian family, like Kannada, Malayalam, Tamil, are found in the literature. Ramaswamy, et al. [7] propose a rule-based morphological analysis with Finite-State Transducers for Kannada.

TECHNIQUES FOR MORPHOLOGICAL ANALYZER

Various NLP research groups have developed different methods and algorithms for morphological analysis. Some of the algorithms are language-dependent and some of them are language independent. A brief survey of various methods involved in Morphological Analysis includes the following.

1.Finite State Automata (FSA)

A finite-state automaton is a mathematical model for an abstract digital computer, consist of states and arcs called transitions. Each FSA has exactly only one initial state and one or more final states. The transitions are the connections between the states. Conventionally, the states are represented as circles and the transitions between them are represented as labelled to arcs; and an arrow is used to indicate the initial state and double circles are used to indicate the final state. The finite-state automata are best understood as recognizers because they accept a finite set of input strings

2. Two Level Morphology

The-two level model is based on a lexicon system and a set of two-level rules. The two-level model is fully bidirectional both conceptual and process-based. It can also be interpreted as a morphological model of the performance processes of word-form recognition and production.

3. Finite State Transducer (FST)

A Finite-State transducer is an enhanced finite-state machine. FST transform one string to another by a process. A finite-state transducer accordingly implements a relation between two formal languages: an upper-side and lower-side regular language, it literally 'transduces' strings from one language into the other. The process of transformation is called transduction.

4.Corpus-Based Approach

Corpus or text corpus is a large and structured set of texts. They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory. Corpus-based methods are called "supervised" when they learn from previously sense annotated data, and therefore they usually require a large amount of human intervention to annotated the training data.

5. Paradigm Based Approach

Paradigm is the complete set of related word forms associated with a given lexeme. A familiar example of paradigms is the inflections of nouns. Accordingly, the word forms of a lexeme may be arranged by classifying them according to shared inflectional categories such as tense, aspect, mood, number, gender or case.

SYSTEM IMPLEMENTATION.

Here we proposed a Malayalam morphological analyzer using finite-state automation. A Finite State Machine or Finite State Automation (FSA) is a model of behaviour composed of state, transitions and actions. A finite-state automaton is a device that can be in one of a finite number of states. If the automation is in a final state, when it stops working, it is said to accept its input. The input is a sequence of symbols.

FSA is used to accept or reject a string in a given language. It uses regular expressions. When the automaton is switched on it will be in the initial stage and start working. In the final state, it will accept or reject the given string. In between the initial state and finite state, there is a transition process of switching over to another state. Regular expressions are powerful tools for text searching. FSA can be used to represent morphological lexicon and recognition.

A sample automata for accepting the words □□□□□□, □□□□□□□□□□ is shown in figure 1

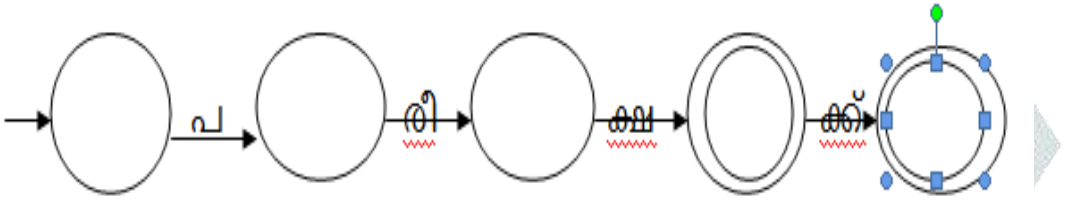

Figure1: Automata for accepting the words പരീക്ഷ, പരീക്ഷക്ക് .

The above automata will only halt when it accepts the words പരീക്ഷ, പരീക്ഷക്ക് by entering to its final states.

CONCLUSION AND FUTURE ENHANCEMENT.

Natural Language processing deals with the processing of native languages. Morphological analysis plays an important role in NLP applications. The morphological analysis deals with analyzing individual words.

This paper presents a Finite State Machine based morphological analyzer for the Malayalam language. The system works only in noun and verb form; we can extend the work in pronouns, adverbs, adjectives, etc. More new Finite Automations implementations are possible.

REFERENCES

[1] Sunil R, Nimtha Manohar, Jayan V, K G Sulochana. "Morphological Analysis and Synthesis of Verbs in Malayalam". ResearchGate, 2012.

[2] Vinod PM, Jayan V, Bhadran VK, " Malayalam Morphological Analyser based on hybrid approach ". Proceedings of the Twenty-Fourth Conference on Computational Linguistics and Speech Processing (ROCLING). 2012.

[3] 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

[4] Nizar Habash and Ramy Eskander and Abdelati Hawwari. "A Morphological Analyzer for Egyptian Arabic". Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON2012), pages 1–9, Montréal, Canada, June 7, 2012. c 2012 Association for Computational Linguistics.

[5] Rinju O.R, Rajeev R. R, Reghu Raj P.C., Elizabeth Sherly. "Morphological Analyzer for Malayalam: Probabilistic Method Vs Rule Based Method". ResearchGate. 2013.

[6] Finite State Morphology by Kenneth R. Beesley and Lauri Karttunen, CSLI Publications.

[7] Ramasamy Veerappan, Antony P J, S Saravanan and Dr. Soman K P. Article: A Rule based Kannada Morphological Analyzer and Generator using Finite State Transducer. International Journal of Computer Applications 27(10):45-52, August 2011.

[8] Jurafsky.D & Martin J.H. (2000) Speech and language processing: an introduction to
natural language processing, computational linguistics and speech recognition. Upper Saddle
River, NJ: Prentice Hall.

[9] Manning, C.D. & Schutze, H. (1999). Foundations of statistical natural language processing. Camridge, MA: MIT Press