IJCSE International Journal of Computer Sciences and Engineering Open Access
Study Paper Volume-2, Issue-5 E-ISSN: 2347-2693

The Role of Morphological Analyzer and Generator for Tamil language in Machine Translation Systems

Ananthi Sheshasaayee [1] and Angela Deepa.V.R [2*]

[1,2*] *Department of Computer Science & Application, Quaid-E- Millath Government College for Women (Autonomous), Chennai 600 002, India*

www.ijcseonline.org

Received: 08/04/2014 Revised: 22 /04/2014 Accepted: 15/05/ 2014 Published: 31 /05/2014

Abstract— Natural language processing aims to design and build software that will analyze, understand and generate languages that humans use naturally. Machine translation is a very important application in Natural language processing (NLP).Currently, Statistical machine translation plays a predominant role in machine translation of larger vocabulary tasks. Achieving this goal is not an easy task especially when it comes to languages like Tamil which are agglutinative in nature. Deep analysis is needed at the word level to confine the correct meaning of the word from its morphemes and categories.The computational implementation of analysing natural language is done by Morphological analyzer.This paper is a ground work for better understanding of various approaches that are used to develop morphological analyzer and generator of Tamil languages.

Keywords— Natural language processing(NLP), Morphology, Morpheme, Support vector machine(SVM), Automata theory. Agglutinative languages

I. INTRODUCTION

Machine translation is the task of translating the text in source language to target language automatically. Though research has been made for past three decades building a machine translation is still an open problem relating to the language grammatical coherence, analyzing the word forms, creating bilingual dictionaries, language generation etc. Morphology is one of the levels in the science of languages, which deals with the structure and formation of words.[9].The linguistics rules plays a major role in the structure of the language and the morphological rules helps in the formation of the word in a sentence. For highly agglutinative and morphological rich language like Tamil developing the linguistic tools and the machine translation system is a challenging task. Therefore the languages are preprocessed to handle the morphology richness of the language. Morphological analyzer plays a predominant role in preprocessing the morphological language to find the lemma and its morphological information to build better unswerving machine translation system.

II. STATE OF ART

A. AMRITA Morph Analyzer and generator for Tamil- A Rule Based Approach (2010): Dr. A.G. Menon, S.Saravanan, R. Loganathan and Dr. K. Soman, Amrita University, Coimbatore : Their developed a rule based Morphological Analyzer and generator for Tamil using finite state transducer called AMAG [8]. The system consists of list of 50000 nouns, around 3000 verbs and a relatively smaller list of adjectives. The efficiency of the system is based on lexicon and orthographic rules from a two level morphological system.The better performance is proved by comparing the proposed AMAG with the existing Tamil morph analyzer and generator called ATCHARAM

B. *A Novel Algorithm for Tamil Morphological Generator (2010): M.Anand Kumar, V.Dhanalakshmi and Dr.K.P Soman, CEN, Amrita University, Coimbatore:* Development of a morphological generator for Tamil based on suffix stripping algorithm [6] .This system is simple,efficient and independent algorithm whih can handle compound words,transitive,intransitive and also proper nouns.The system consists of two modules, in which the first module handles the lemma/root part and the second module handles the Morpho-lexical information. The system required the following information: morpho-lexical information file, suffix table, paradigm classification rules and stemming rules.

C. *An Improvised Morphological Analyzer cum Generator for Tamil (2010): This work is proposed byParameswari K, CALTS, University of Hyderabad:* It deals with the improvised database implemented on Apertium for morphological analysis and generation [3][5]. The improvised MAG uses the Finite State Transducers algorithm for one-pass analysis and generation, and the Word and Paradigm based database.The developed system performance is measured and compared for its speed and accuracy with the other available Tamil Morphological analyzers which were developed in CALTS and AU-KBC research Centre, Anna University.The final experimental result showed that the proposed MAG performs better and accurate than the other.

D. A Sequence Labelling Approach to Morphological Analyzer for Tamil (2010): Anand Kumar M,Dhanalakshmi

Corresponding Author: *Angela Deepa.V.R*

© 2014, IJCSE All Rights Reserved

V, Soman K.P and Rajendran S of AMRITA Vishwa Vidyapeetham, Coimbatore: This morphological analyzer for Tamil language is based on sequence labelling approach [7]. In the proposed work morphological analyzer problem is solved using machine learning approach which is redefined as classification problem. This is a corpus based approach,which consists of 130,000 verb words and 70,000 noun words respectively. The training and testing is performed with support vector machine algorithms.The system is tested with 40000 verbs and 30000 nouns taken from Amrita POS Tagged corpus. The performance of the system was compared with other related systems developed using the same corpus and the results showed that SVM based approach outperform other.

E. FSA-based morphological generator for Tamil (2010): A finite state automata based morphological generator is developed by Menaka S, Vijay Sundar Ram and Sobha Lalitha Devi [4]. The finite-state based morphological generator is well-suited for highly agglutinative and inflectional languages like Tamil. For evaluating the nouns and verbs with the correct and wrong inputs two separate systems are experimented to evaluate their nature to bring out accurate results.

F. Rajendran's Morphological Analyzer for Tamil: The first step towards a preparation of morphological analyzer for Tamil was initiated by 'Anusaraka' group of researchers under the guidance of Dr Rajendran [4], Tamil University, Tanjavoor. '*Anusaraka*' is machine translation project intended for translation between Indian languages. The developed morphological analyzer for Tamil was used for translating Tamil language into Hindi at the word level.

G. Ganesan's Morphological Analyzer for Tamil[2]: Ganesan developed a morphological analyzer for Tamil to analyze CIIL corpus. He exploits phonological and morphophonemic rules as well as morphotactic constraints of Tamil in building morphological analyzer. Recently he has built an improved and efficient morphological parser.

H. Kapilan's Morphological Analyzer for Tamil Verbal Forms[2]: Another attempt was made by Kapilan for he prepared a morphological analyzer for verbal forms in Tamil.

I. Deivasundaram's Morphological analyzer: Deivasundaram has prepared a morphological analyzer for Tamil for his Tamil Word Processor. He made use of phonological and morphophonemic rules and morphotnatic constraints for developing his parser.

J. AUKBC Morphological Parser for Tamil[2]: AUKBC NLP team under the supervision of Dr Rajendran developed a Morphological parser for Tamil. The API Processor of AUKBC makes use of the finite state machinery like PCKimmo. It parses, but does not generate.

K. Vishnavi's Morphological Generator for Tamil[2]: Vaishnavi researched for her M.Phil. Dissertation on morphological generator for Tamil. The Vaishanvi's morphological generator implements the item and process model of linguistic description. The generator works by the synthesis method of PCKimmo.

L. Ramasamy's Morphological Generator for Tamil[2]: Ramasamy has developed a morphological generator for Tamil for his MPhil dissertation.

M. Winston Cruz's Parsing and Generation of Tamil Verbs[2]: Winston Cruz makes use of GSmorph method for parsing Tamil verbs. GSmorph too does morphotactics by indexing. The algorithm simply looks up two files to see if the indices match or not. The processor generates as many forms as it parses and uses only two files.

N. Vishnavi's Morphological Analyzer for Tamil[2]: Vaishnavi again researched for her Ph.D. dissertation on the preparation of Morphological Analyzer for Tamil. She proposes a hybrid model for Tamil. It finds its theoretical basis in a blend of IA and IP models of morphology. It constitutes an in-built lexicon and involves a decomposition of words in terms of morphemes within the model to realize surface well-formed word forms. The functioning can be described as defining a transformation depending on the morphemic nature of the word stem. The analysis involves a scanning of the string from the right to left periphery scanning each suffix at a time stripping it, and reconstructing the rest of the word with the aid of phonological and morphophonemic rules exemplified in each instance. This goes on till the string is exhausted. For the sake of comparison she implements AMPLE and KIMMO models. She also evaluates TAGTAMIL, API Analyzer and GSMorph.She concludes that Hybrid model is more efficient than the rest of the models.

O. Dhurai Pandi's Morphological Generator and Parsing Engine for Tamil Verb Forms[2]: It is a full-fledged morphological generator and a parsing engine on verb patterns in modern Tamil.

P. RCILTS-T's Morphological analyzer for Tamil[2]: Resource Centre for Indian Language Technological Solutions has prepared a morphological analyzer for Tamil. It is named as 'Atcharam'. 'Atcharam'takes a derived word as input and separate into root word and associated morphemes. It uses a dictionary of 20000 root words based on fifteen categories. It has two modules - noun and verb analyzer based on 125 rules. It uses heuristic rules to deal with ambiguities. It can handle verb and noun inflections.

Q. RCILTS-T's Morphological generator for Tamil[2]: Resource Centre for Indian Language Technological Solutions also developed a morphological generator for Tamil. It is named as 'Atchayam'.'Atchayam' generates words when Tamil morphs are given as input. It has two

© 2014, IJCSE All Rights Reserved

major modules – noun and verb generators. The noun section handles suffixes like plural markers, oblique form, case markers and postpositions. The verb section takes tense and PNG makers, relative and verbal participle suffixes, and auxiliary verbs. It uses sandhi rules and125 morphological rules. It handles adjectives and adverbs. It has word and sentence generator interfaces.

III. TAMIL LANGUAGE

Tamil belongs to the southern branch of the Dravidian languages, which is rich in literary tradition. It belongs to the a family of around 26 languages native to the Indian subcontinent. The language is categorized as into three periods, Old Tamil (300 BCE – 700 CE), Middle Tamil (700–1600) and Modern Tamil (1600–present).Tamil language is basically a verb-final and a relatively free word order language.

IV. CHALLENGES IN TAMIL LANGUAGE

Tamil is one of the longest surviving classical languages in the world. It is spoken by more than 66 million people all over the world. Tamil is a morphological rich content language and quite complex since it inflect to person, gender and number markings and also combines with the auxiliaries that indicate aspect, mood, causation, attitude etc.Noun root inflects with plurals, oblique, case, postpositions and clitics. Therefore a single noun form can inflect to more than five hundred word forms along with postpositions. A single verb root can inflect to more than two thousand word forms including the auxiliaries. Eventually the identified roots are to be tagged at the word level for further language processing. But due the complexity of the verb forms, capturing it in a machine analyzable and generatable form is a challenging task.

V. MORPHOLOGICAL ANALYSIS

Tamil is one of the Dravidian languages and as such has an agglutinative grammar. It requires deep analysis at the word level to capture the correct meaning of the word from its morphemes and categories. Generally in Tamil language inflections to the root word are post propositional eventually it takes few thousand inflected form of words. It takes both lexical and inflectional morphology.[8] In Lexical morphology the derivational and compounding morphemes are added to the root , which changes the word meaning and its class. Similarly inflectional morphology deals with the change of the form of a word by adding the inflectional morphemes to the root.

Morphological analysis is the process of segmenting a given word into a sequence of morphemes. A morphological analyzer is a computational tool to analyze word forms into their roots and functional elements. The morphological generator is the reverse process of an analyzer i.e. a well formed word is generated with given root and functional elements. The morphological analyzer and the morphological generator are two essential basic tools for building any language processing application. The design and implementation of morphological analyzer and generator for Tamil language provides research tools for various applications in NLP. A morphological analyzer will return its root/stem word along with its grammatical information depending upon its word category. Like, for nouns it will provide gender, number, and case information and for verbs, it will provide tense, aspects, and modularity. Different methods have been evolved for the implementation of a morph analyzer.[12][13].

VI. MORPHOLOGICAL ANALYSER AND GENERATOR APPROACHES:

There are many language dependent and independent approaches used for developing morphological analyzer and generator [11]. Recent development in Indian language NLP shows that many morphological analyzer and generator are created successfully using these approaches. A brief description of most commonly used approaches is as follow:

A. Corpus Based Approach: In case of corpus based approach, a large sized well generated corpus is required for training. Training of the corpus through any machine learning algorithm collects the statistical information and other necessary features from the corpus. The collected information is used as a MAG model. The performance and the reliability of the system will depends on the feature and size of the corpus. The disadvantage is that corpus creation is a time consuming process. This approach is suitable for languages having well organized corpus.

B. Paradigm Based Approach: The paradigm based approach is well suited for languages which are highly agglutinative in nature.This variant of the scheme has been used widely in NLP. For a particular language, each word category like nouns, verbs, adjectives, adverbs and postpositions will be classified into certain types of paradigms.MAG model is developed based on their morphophonemic behaviour and a Paradigm based morphological compiler program .In this approach a linguist or the language expert is asked to provide different tables of word forms covering the words in a language. Based on this information and the feature structure with every word form a MAG can be build. It is stated that morphological analyzers are developed for almost all Indian languages using paradigm based approach.

C. Finite State Automata (FSA) Based Approach:

Finite state machine or finite state automation FSA (or finite automation) uses regular expressions and is used to accept or reject a string in a given language . In general, FSA is an abstract device used for recognizing simple syntactic structures or patterns. The behaviour of a system composing of state, transitions and actions. When FSA start working, it

© 2014, IJCSE All Rights Reserved

will be in the initial stage and if the automation is in any one of final state it accept its input and stops working.

D. Two- Level Morphology Based Approach:
In 1983, Kimmo Koskenniemi, a Finnish computer scientist developed a general computational model for word-form recognition and generation called Two- level morphology . It is a framework for computational morphological analysis.It incorporates an explicit new formalism for describing morphological and morphophonological phenomena. The advantage of two- level morphology is that the model does not depend on a rule compiler, composition or any other finite-state algorithm. The "two-level" morphological approach consists of two levels called lexical and surface form and a word is represented as a direct, letter-forletter correspondence between these forms. The Two-level morphology approach is based on the following three ideas:

• Parallel Rules are symbol-to-symbol constraints and are conceptually and computationally simpler.which avoids rule interactions., not sequentially like rewrite rules.
• The constraints can refer to the lexical context, to the surface context, or to both contexts at the same time.

E. Finite State Transducers (FST) Based Approach:
FST is a modified version of FSA by accepting the principles of two level morphology. A finite state transducer essentially is a finite state automaton that works on two more tapes most common way to think about transducers is as a kind of "translating machine" which works by reading from one tape and writing onto the other. FST's can be used for both analysis and generation (they are bidirectional) and it act as two level morphology. By combining the lexicon, orthographic rules and spelling variations in the FST, we can build a morphological analyzer and generator at once.

F. Stemmer Based Approach:
Stemmer based approach is a program oriented approach which specify all possible affixes with replacement rules. Potter algorithm is one of the most widely used stemmer algorithm and it is freely available. Stemmer uses a set of rules containing list of stems and replacement rules to stripping of affixes. The advantage of stemmer algorithm is that it is very suitable to highly agglutinative languages like Dravidian languages for creating MAG.

G. Suffix Stripping Based Approach: Dravidian language like Tamil which are highly agglutinative in nature has Words usually formed by adding suffixes to the root word serially. Due to this property a MAG can be successfully build using suffix stripping approach.[10]. Once the suffix is identified, the stem of the whole word can be obtained by removing that suffix and applying proper orthographic (sandhi) rules. A set of dictionaries like stem dictionary, suffix dictionary and also using morphotactics and sandhi rules, a suffix stripping algorithm successfully implements MAG.

H. Directed Acrylic Word Graph Based Approach:
Directed Acrylic Word Graph (DAWG) is a very efficient data structure that can be used for developing both morphological analyzer and generator. DAWG is language independent and does not depend on any morphological rules or any other special linguistic information.It is a lexicon representation and fast string matching, with a great variety of application.Using this approach, the University of Partas Greece developed MAG for Greek language at first time. There after the method is applied for other languages including Indian languages.

VII.CONCLUSION

In this paper work, we have explained the features of the Morphological analyzer and generator.Additionally almost all approaches were applied in Tamil language.Since the language is morphological rich and are agglutinative in nature building an accurate system for the Tamil language is a challenging task.

REFERENCES

[1] Raji Sukumar. A, Dr. Babu Anto , "Morphological Synthesizer a Linguistic Tool for Malayalam verbs",International journal Computational Linguistics and Natural Language Processing.Volume-02,Issue-.53,ISSN2279-0756 Page no(346-349),May 2013
[2] Antony, P. J., and K. P. Soman. "Computational morphology and natural language parsing for Indian languages: a literature survey." International journal Computer Science Engineering Technology" Volume-03,Issue-04, ISSN : 2229-3345 ,Page no(136-146),April 2012
[3] Parameshwari K, "An Implementation of APERTIUM Morphological Analyzer and Generator for Tamil", Language in India. 11:5 Special Volume:Problems of Parsing in Indian Languages, May 2011.
[4] Menaka, S., Vijay Sundar Ram, and Sobha Lalitha Devi. "Morphological Generator for Tamil." Proceedings of the Knowledge Sharing event on Morphological Analysers and Generators, LDC-IL, Mysore, India ,Page no (82-96),March 22-23, 2010
[5] Parameswari K, "An Improvized Morphological Analyzer cum Generator for Tamil: A case of implementing the open source platform APERTIUM", Knowledge Sharing Event 1 – CIIL, Mysore, March 2010.
[6] M.Anand Kumar, V.Dhanalakshmi and Dr. K P Soman, "A Novel Algorithm for Tamil Morphological Generator", 8th International Conference on Natural Language processing (ICON),IIT Kharagpur, December 8-11,2010.
[7] Anand Kumar M, Dhanalakshmi V, Soman K.P and Rajendran S, "A Sequence Labeling Approach to Morphological Analyzer forTamil Language", (IJCSE) International Journal on Computer Science

 © 2014, IJCSE All Rights Reserved

and Engineering Volume 02,Issue No. 06, Page no(2201-2208),2010.

[8] Dr. A.G. Menon, S. Saravanan, R. Loganathan Dr. K. Soman,, "Amrita Morph Analyzer and Generator for Tamil: A Rule Based Approach",In proceeding of Tamil Internet Conference, Coimbatore, India.Page no(239-243),2009.

[9] Jayan, Jisha P., R. R. Rajeev, and S. Rajendran. "Morphological Analyser for Malayalam-A Comparison of Different Approaches." International journal of computer science and Information technologies(ICSIT),Volume02,Issue-02, Page no(155-160),2009

[10] Saranya S.K, 'Morphological Analyzer for Malayalam Verbs, A Project report. www.pdfcreat.com/computer-malayalam-0.html 2008.

[11] Choudhary and Narayan Kumar, 'Developing a Computational Framework for the Verb Morphology of Great Andamanese', Doctoral dissertation, Centrefor Linguistics, Jawaharlal Nehru University, New Delhi, India 2006.

[12] Hughes,John' Automatically Acquiring a Classification of Words', Doctoral dissertation , School of Computing,University of Leeds ,2006.

[13] Pesetsky, D.'Russian Morphology and Lexical Theory', Manuscript, MIT11.1979, Siegel, Dorothy Carla,Topics in English morphology, Doctoral dissertation,Massachusetts Institute of Technology.1974

AUTHORS PROFILE

Dr. Ananthi Sheshasaayee
Associate .Prof. & Head,
PG & Research Dept. of Comp.Sci
Quaid-e-Millath Govt.College for Women
Chennai,India.

Angela Deepa.V.R
Research Scholar
PG & Research Dept. of Comp.Sci
Quaid-e-Millath Govt.College for Women
Chennai,India.

 © 2014, IJCSE All Rights Reserved