BEST: Journal of Recent Trends in Economics, Ancient history & Linguistic Research (BEST: JRTEAHL)
Vol. 3, Issue 1, Jun 2017, 15-22

SMART MORPHOLOGICAL ANALYZER FOR TAMIL

M. RAJASEKAR[1], N. RAJASEKHARAN NAIR[2] & A. UDHAYAKUMAR[3]

[1]Research Scholar, Department of Languages, Hindustan University, Chennai, Tamil Nadu, India

[2,3]Professor, Department of MCA, Hindustan University, Chennai, Tamil Nadu, India

ABSTRACT

In the emerging field of Natural language processing, the morphological analyzer is one of the most basic tool. It analyses the word forms in a given sentence and identify the root word and shows its various morphological forms. In this research work I have proposed a factored method to analyze the words as morphologically in a given sentence.

KEYWORDS: Morphology, Morphemes, Natural Language Processing, Support Vector Machines

INTRODUCTION

The term morphological analysis is the wider area in the current research and development. The same process will be done in chemical engineering, bio-medical engineering etc., In spite of its individuality; Dravidian languages have no morphological analyzer in public domain. The absence of linguistic tool like morphological analyzer motivated me to develop a new linguistic tool for Tamil language.

Morphology

In any language the grammar is widely divided into two parts, morphology and syntax. In 1859, August Schleicher seeded the concept of morphology in linguistics. Morphology mainly deals with rules, syntax and words. Computational morphology is nothing but development of techniques and concept for computational morphological analysis of word forms. In systematic manner anyone can get the morphological information of the given word in the sentence.

Morphological Analyzer

In the first step of a morphological analysis dividing a word into lemma and morpho-lexical information. In the real life, a word is obviously a sequence of characters in a particular language delimited by spaces, punctuation marks etc., A word can be classified as two types: simple and compound. A simple word means it will be a single root word or stem together with prefixes or suffixes. A compound word can be divided into two or more independent simple word. Constituents of a simple word are defined as morphemes [1]. An analyzer of words in a given sentence, it simply look the delimiters in the sentence. It identify the word, then divide as the simple word or a compound word. If it is a compound word then break into its constituent simple words. Then proceed into analyze. Example for a Morphological analysis is given below.

- மரங்கள் = மரம் (பெயர்ச்சொல்) + முன்னிலை+ கள் (பன்மை)
- படித்தான் = படி(பெயர்ச்சொல்) +த் + முன்னிலை + ஆன் (ஒருமைஆண்பால்)

Morphological Generator

A morphological generator the reverse of an analyzer. The input should be a root word, its features. Then the generator generate the morphological word. Example for morphological generator is given below:

- மரம் (பெயர்ச்சொல்) + முன்னிலை+ கள் (பன்மை) ===>மரங்கள்
- படி(பெயர்ச்சொல்) +முன்னிலை + ஆன் (ஒருமைஆண்பால்) = ==>படித்தான்

Role of Morphological analyzer in NLP

Morphological analyzer plays a vital role in Natural Language Processing. Some applications use the morphological analyzer.

- Machine Translation
- Grammar checker
- Spell checker
- Information Retrieval and Extraction
- Knowledge representation

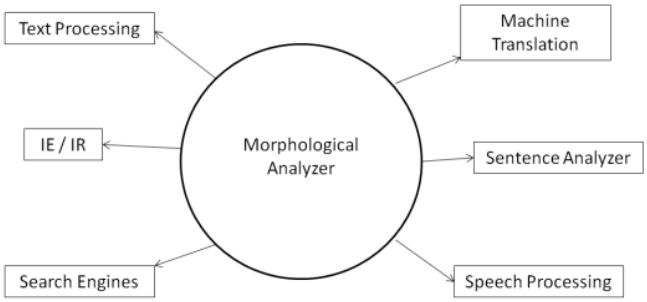

Figure.1: Morphological Analyzer

LITERATURE REVIEW

The powerful tool in the Natural Language Processing like, Morphological Analyzer plays major role in the Tamil language. It is the important process to derive a word in Tamil. Now we will see some of the research methods already exhibited by the special researchers in Natural language processing for Tamil.

A. *Morphological Analyzer for Classical Tamil Texts: A Rules based approach* by Mr. R. Akilan, with the guidance of Professor E.R. Naganathan [2] have proposed a analyzer for classical Tamil language. In this work, they have framed a rules to analyze the words and characters. This rules will be modified based on the requirement in future.

B. *Tamil Morphological Analyzer*, by Mr. S. Rameshkumar, and Mr. S. Viswanathan [3], In this research work they have developed an API to retrieve the root word from a inflected word.

C. *An Effective rules based system for Morphological Parsing of Tamil Language*, by Mr. Karthick, Mr. Praveen, Dr. V. Gopalakrishnan [4], NIT, Trichirapalli, They have designed a system to give the parsed result of a particular word. They have analyzed the given word in surface level and lexical level.

The Analysis of these research work it is very important to make a useful morphological analyzer in Tamil, and it should be effective than these models. All of the above models are used only rules to analyze the segments. In my research work I have implemented the SVM Support Vector Machine algorithms to give best results while analyzing the given input.

OBJECTIVES

There are major objectives to develop this morphological analyzer.

- It will be helpful for the machine translation in Tamil
- To provide linguistic help in machine learning
- To provide the detailed information of syntax and morphemes in Tamil
- Main object is, it is the part of my research work. Machine translation for a native language Tamil

Morphological Analysis System for Tamil

In any machine translation system morphological analysis is the second process. First one the POS Tagging. Figure.2. shows the overall structure of the Morphological analyzer in Tamil.

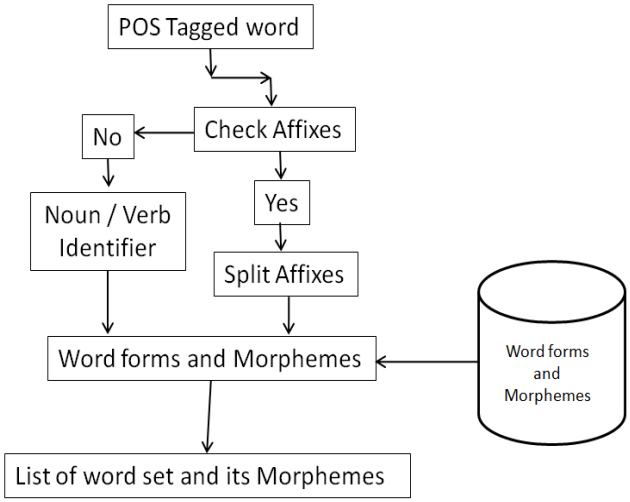

Figure 2: Structure of Morphological Analyzer

Steps

- The POS Tagged word will be undergone for the affix checking.
- If affixes are found, it will go for split of affixes from the root word.
- If affixes not found, it will go for noun / verb identification.
- After that the word set will go for morphological identification from the database.
- Then each and every words in the list will be tagged with its appropriate morphemes.

MACHINE LEARNING APPROACH FOR MORPHOLOGICAL ANALYZER

In the morphological analysis process the machine identifies root and affixes of a particular word. Commonly, rules based machine learning approach is used for morphological analysis. In this type of approach the machine has set of rules and bilingual dictionary with root and morphemes. In this rules based approach each rule depends on the previous rule. So, it is very difficult, if one rule fails, then next rule also will be failed. It affects after the entire rule which is already failed.

Machine Learning

- Machine learning is a subset of Artificial Intelligence, discussed with the design of algorithms that can be implemented in the examples. Machine learning can be classified as two types:
- Supervised Machine Learning
- Unsupervised Machine Learning

In the supervised learning the input and output samples are used. But in the unsupervised learning only input samples can be used.

Morphological Analyzer based on machine learning does not require any morphological rules. It only needs morphologically segmented corpora [5]. In the supervised classification, a significant process is sequence labeling. In a sequence method we can assign the label for each word from a given sentence. This sequential labeling approach can be classified as two types: [4].

- Joint Segmentation
- Raw Labeling

The joint segmentation will get a labeling for a whole segment. In the raw labeling method each input will get a labeling. In our morphological analyzer a sequence means a word, and element means a character in that word. We can express that as,

W (Noun/Verb) = R(Noun/Verb) + I (Noun/Verb)

In this expression, W means a word, R means a root word of that given word, I means the morphological inflections of that word. The inflections I can be $i_1+i_2+i_3…..+i_n$. n – number of morphemes of the word.

Let X can be a finite set of input characters and Y can be a set of output characters. Then, $x_n \epsilon$ X. similarly, $y_n \epsilon$ Y.Finally, the labeling segmentation are,

Table1. Permutations of the given word

Inputs	x_1	x_2	x_3	x_4	….	…..	x_n
Outputs	y_1	y_2	y_3	y_4	….	…..	y_n

The permutations of a given word to the morphological analyzer will be generated as above.

MORPHOLOGICAL ANALYZING USING SVM TOOL

In the year of 1992, Vladimir N, Bernhard E. Boser, Isabelle M. Guyon and. Vapnikare suggested a way to create nonlinear classifiers by applying Support vector technique [6] . At first is used for binary classification technique. Later it

is extended to multi-class classification. Now, my work morphological analyzer problem should be viewed as classification problem. This classification problem can be recitified by supervised machine learning algorithms [2].

It is the technical algorithm for binary classification problem.

Let $\{(x_1y_1), \ldots(xny_n)\}$ to be a N data set to be classified in set of rules. Where each element x_1 is a vector in R^N and $y_1 \in \{-1,1\}$ is the class label.

Support Vector Machine

SVM tool is the open source tool to generate sequential taggers on SVM. Actually SVM Tool was found for POS tagging. The SVM Tool has three components.

- Model Learner
- Tagger
- Evaluator

The SVM Model learner is used to learn the Tag details of the corpus with its corresponding data. The Tagger has to tag a sequence of models with the data. The evaluator produces the tagging results.

Morphology in Tamil Language

Morphological analysis in Tamil is very rich. Words in Tamil language are formed with lexical roots with one or more affixes. The lexical roots and other affixes are called morphemes. The first lexical root may or may not be followed by functional or grammatical morphemes. For example and word *மாம்பழங்கள்* 'MampazhangaL' in Tamil, will be divided into *மாம்பழம்* 'Mampazham' and *கள்* 'kaL'. It is bound to the lexical root to add plurality to the lexical root.

Linguistic syntax in Tamil

Tamil has linguistically strong syntax. It has the standard word order (SOV). It is also a free-word order language. In a word order of a sentence, a noun can appear in any permutation of its morphemes before the final verb, but it gives the same meaning of that sentence. Of course Tamil is a null subject language. It have SOV pattern not at all it has the SOV pattern. In Tamil language only it is possible to have only a verb, **சென்றது**, 'seNRathu' ('have gone') or only a subject and object without a verb, **இதுஎனதுமரம்**, 'ithueNathumaram' (This is my tree).

The Rules to form a Word in Tamil

The rules determine the type of the output associated with the rule. In Tamil, the grammatical category may change or may not change. The following are the rules in Tamil to form a word from English.

- Noun ➔ Noun
- Noun ➔ Verb
- Noun ➔ Adjective
- Verb ➔ Noun

- Verb ➔ Adjective
- Verb ➔ Verb
- Adjective ➔ Noun
- Adverb ➔ Verb

Morphological Analyzer

The morphological analyzer will analyze the tagged word then produce the morphological terms of that word. It will be helpful to find the exact meaning of the given word in target language (English). In this process the system can verify the ambiguity. With the output of this morphological analyzer the system can reduce the ambiguity in finding the exact meaning of the given word. The morphological generator will generate different terms for a noun and a verb also.

The morphological terms are:

Noun Terms

Table 2: Nouns Morphemes

Nominative (நியமிக்கும்உரிமை)	புலி	புலிகள்
Accusative (2-ம்வேற்றுமை)	புலியை	புலிகளை
Benfefative (3-ம்வேற்றுமை)	புலிக்காக	புலிகளுக்காக
Dative (நான்காம்வேற்றுமை)	புலிக்கு	புலிகளுக்கு
Ablative (ஐந்தாம்வேற்றுமை)	புலியிலிருந்து	புலிகளிலிருந்து
Genitive (ஆறாம்வேற்றுமை)	புலியினது	புலிகளிது
Instrumental (காரணமாகஇருக்கிற)	புலியால்	புலிகளால்
Sociative (ஐக்கியப்படுத்தும்) ஓடு	புலியோடு	புலிகளோடு
Sociative (ஐக்கியப்படுத்தும்) உடன்	புலியுடன்	புலிகளுடன்
Locative (தேடு)	புலியில்	புலிகளில்

Morphological forms for the given Noun are:

- கண்ணகியை
- கண்ணகியால்
- கண்ணகிக்கு
- கண்ணகியின்
- கண்ணகியிடம்
- கண்ணகியினுடைய
- கண்ணகியினது
- கண்ணகியினுடையது
- கண்ணகியோடு

- கண்ணகியினிடத்தில்
- .கண்ணகியிடமிருந்து

Verb Terms

Table 3: Verb Morphemes

PNG	PRESENT – PNG	PAST – PNG	FUTURE – PNG
1S	எழுதுகிறேன்	எழுதினேன்	எழுதுவேன்
1P	எழுதுகிறோம்	எழுதினோம்	எழுதுவோம்
2S	எழுதுகின்றாய்	எழுதினாய்	எழுதுவாய்
2SE	எழுதுகின்றீர்	எழுதினீர்	எழுதுவீர்
2PE	எழுதுகின்றீர்கள்	எழுதினீர்கள்	எழுதுவீர்கள்
3SM	எழுதுகின்றான்	எழுதினான்	எழுதுவான்
3SF	எழுதுகின்றாள்	எழுதினாள்	எழுதுவாள்
3SE	எழுதுகின்றார்	எழுதுவார்	எழுதினார்
3SN	எழுதுகின்றது	எழுதியது	எழுதுவது
3PE	எழுதுகின்றார்கள்	எழுதினார்கள்	எழுதுவாரகள்
3PN	எழுதுகின்றானா	எழுதினானா	எழுதுவானா

RESULTS

The result of the morphological analyzer is showed below. The result is evaluated with three states. Character level, Word level and sentence level.

- *Word level evaluation*

The output of the system is evaluated by manually. In this sample evaluation we have collected 2480 words. And it contains 26,767 characters.

- *Character level evaluation*

The character level evaluation also done. In Tamil a single character also have some powerful meaning. The results and evaluation chart also given below.

Table 4: Evaluation

Type	Noun		Verb	
	Word	Char	Word	Char
Tested	1382	14916	1098	11851
Correct	1276	14001	978	10912
Accuracy	92.3	93.8	89	92

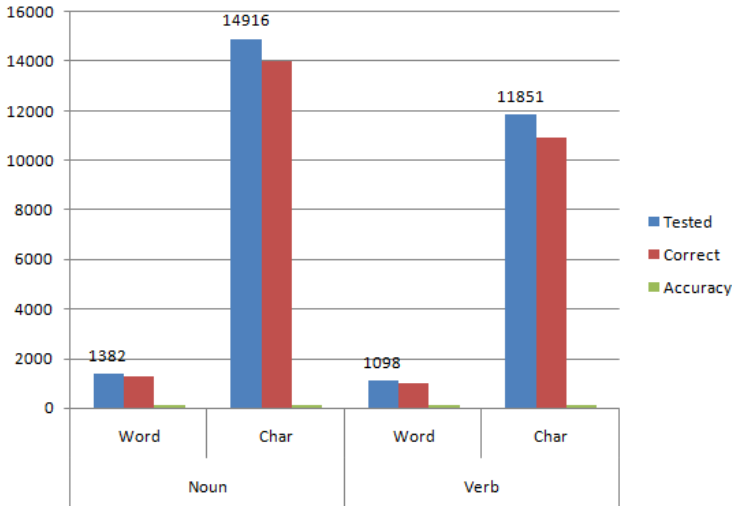

Figure 3: Evaluation

CONCLUSIONS

The proposed morphological analyzer system is under progress. It is one of the processes in the whole research work. The output of the system compared by manual terms in morphological analysis. We are planning to give more accuracy and comparing with other morphological tools.

REFERENCES

1. Anandan. P, RanjaniParthasarathy, Geetha T.V.2002. Morphological Analyzer for Tamil,
2. ICON 2002,RCILTS-Tamil, Anna University, India
3. Morphological Analyzer for Classical Tamil Texts: A Rule based approach, R. Akilan* and
4. Prof. E.R. Naganathan, IJISET, Vol. 1 Issue 5, July 2014.
5. Tamil Morphological Analyser, AU-KBC Centre, S. Ramesh Kumar, S. Viswanathan
6. An Effective rules based system for Morphological Parsing of Tamil Language, by Mr.
7. Karthick, Mr. Praveen, Dr. V. Gopalakrishnan, NIT, Trichirapalli,
8. Hal Daume (2006), http://nlpers.blogsp-ot.com/2006/11/-getting-started-\
9. insequencelabeling.html.
10. Jes´us Gim´enez and Llu´ıs M`arquez,2006,SVMTool:Technical manual v1.3,August 2006.