Morpheme Extraction in Tamil using Finite State Machines
(FIRE-2013 - Morpheme Extraction Task)

Sobha Lalitha Devi, Marimuthu K, Vijay Sundar Ram R, Bakiyavathi T, Amudha K
AU-KBC Research Centre,
MIT Campus of Anna University,
Chromepet, Chennai, India

Abstract

In this work we present an efficient morphological analysis system for Tamil where we extract the individual morphemes using Finite State Automaton (FSA). The aim is to perform morpheme analysis by modeling the regular inflectional pattern exhibited by Tamil as an FSA. We handle the compound and agglutinated words using CRF based word boundary identifier. On linguistic evaluation the system achieved an encouraging accuracy of 86.17%

Keywords: Tamil Morphological Analyzer, Morpheme Extraction Task, Dravidian Languages, Finite State Automaton

1. Introduction

Tamil is a morphologically rich and highly inflectional language. It is inflectional in nature because several morphophonemic changes happen as the words are formed by adding suffixes to the root word. It is a verb-final language with a relatively free word order. Higher degree agglutination is also a common phenomenon.

The morphotactics defines the order with which the morphemes are arranged in a word and the morphophonemics describe the changes that happen when suffixes attach themselves to the root words.

In any given lexical category the process in which the morphemes or inflections occur as affixes to the root words is called Affixation. Specifically the morphemes which occur only as suffixes to the root words during word formation is called as Suffixation. There is no circumfix in Tamil.

The features of morphological analysis output serve as vital information in various modules of NLP systems such as machine translation, information retrieval systems, anaphora resolution engines, and question-answering systems. Hence it is mandatory to have a robust and highly accurate morphological analyzer as it directly affects the performance of the systems.

Here we discuss the development of an efficient finite state morphological analyzer for middle, contemporary and modern Tamil.

2. Literature Survey

Morphological analysis using two-level morphology by Kimmo Koskenniemi is the early attempt where he tested the formalism for Finnish language (Koskenniemi 1983). In the two-level representation the surface level describes the word forms and the lexical level encodes the lexical units such as stem and suffixes. The two-level rules define a mapping between the two levels and they are represented in a Finite State Automaton (FSA). For the languages such as Arabic, Dutch, English, French, German, Italian, Japanese, Portuguese and Swedish morphological analyzers have been developed using the two-level morphological model (Schulze 1994).

Generally rule based approaches are used for building a morphological analyzer (Rajendran.S et al., 2001). S. Viswanathan et al (2003) and Vijay Sundar Ram et al (2010) have developed Tamil morphological analyzer using FSA and paradigm-based approaches.

Unsupervised learning methodology is used to develop a morphological analyzer for Bengali language (Sajib Dasgupta, 2007) and the formalism of two-level morphology was used to handle Bengali compound words.

Eryigit and Adali (2004) propose a suffix stripping approach for Turkish language. The rule based and agglutinative nature of Turkish morphology makes the language to be modeled using FSMs and also it does not need a lexicon. But for the morphological analysis of Tamil the lexicon serves the primary purpose of reflecting the inflectional pattern as the language has a well defined paradigm-based inflection. This makes the task of modeling the inflectional and orthographic rules for the whole Tamil morphology using FSMs an easy one.

3. Morphological Analysis

In the following subsections we describe noun, pronoun and verb analysis.

3.1 Noun Analysis

The analysis of nouns should identify the root word, its suffixes along with gender, number and case information. Ontology information is utilized for determining and assigning the gender category. Nouns occupy a prominent part in Tamil language and it takes derivational and inflectional suffixes. It inflects to eight case suffixes which are listed in table 1.

Table 1. Case markers and suffixes in Tamil

Case	Suffixes
Nominative (NOM)	no suffix
Accusative (ACC)	“ஐ ” (ai)
Ablative (ABL)	“இலிருந்து ” (ilirunthu),“இடமிருந்து ” (itamirunthu)
Dative (DAT)	“கு ” (ku)
Locative (LOC)	“இல் ” (il)
Instrumental (INS)	“ஆல் ” (aal)
Inclusive (INC)	“உம் ” (um)
Genitive (GEN)	“இன் ” (in), “அது ” (athu), “உடைய ” (utaya)
Sociative (SOC)	“ஓடு ” (Otu), "உடன்"(*utan)*

Affixation, specifically suffixation, to a noun happens in a particular order in Tamil. It will always happen in one of the two ways:

- case suffix attaches to the noun root. Postpositions follow the case suffixes which in turn are followed by clitics (“ஏ”(ee), “ஓ” (O)).
- Plural suffix attaches to the noun root which is followed by case suffixes and postpositions which in turn is followed by clitics.

The morphosyntax of noun inflections will be represented as,

- Noun Root + [Case] + [Postpositions] + [Clitics]
  E.g. “மலரைத்தானே”(malaraiththaanee)
  malar +ai +thaan +ee
  flower,N ACC PSP Clitic
  (Only the flower)

- Noun Root + [Number] + [Case] + [Postpositions] + [Clitics]
  E.g. “மலர்களைத்தானே”(malarkaLaiththaanee)
  malar +kaL +ai +thaan +ee
  flower,N Plural ACC PSP Clitic
  (Only the flowers)

3.2 Pronoun Analysis

The analysis of pronouns should identify the root word, its suffixes along with singular/plural and case information. Ontology information is utilized for determining and assigning the gender category. Pronouns in Tamil confine themselves to a given set of words. They behave exactly like nouns with a few exceptions. Specific to these exceptions, sometimes the word itself will indicate the number information. E.g. ungaL, niingaL, thangaL etc. Alternatively the words like “ivai”, “avai” and “evai” indicate plurality by itself and also they optionally take another number suffix “kaL” to again indicate the plurality. But this kind of double plurality does not have any special significance. The word “evai” may function as a question word.

3.3 Verb Analysis

The analysis of verbs should identify the root verb and its suffixes, Tense, Aspect and Modal [TAM] along with the Person, Number and Gender [PNG] information. Ontology information is utilized for assigning the gender category. The verbs in Tamil can be classified into two types: Finite and Non-Finite verbs.

3.3.1 Finite Verbs

The tense suffix attaches to the verb root which is followed by PNG suffix and in turn can optionally be followed by clitics or postpositions or case suffixes. The tense can be Present/Past/Future in the affirmative case while the negative form does not take any tense. The Morphosyntax of Finite Verbs can be represented as

- Verb Root + [Tense] + [PNG] + [Clitics] + [Postpositions]
  E.g. "படித்தானே"(patiththaanee)
  pati + thth + aan + ee
  study,V PAST 3SM clitic
  (he only studied)

- Verb Root + [Tense] + [PNG] + [Case] + [Postpositions]
  E.g. “படித்தவனும்தான்” (patiththavanumthaan)
  pati +thth +avan +um +thaan
  study,V PAST 3SM Incl +psp
  (including him who studied)

3.3.2 Non-Finite Verbs

The verb root directly takes one of the following suffixes - Infinitive or Verbal Participle, Conditional, Concessive, Hortative, Optative which in turn may be followed by Emphatic or Interrogative or Supplemental markers. The morphosyntax of non-finite verb can be given as,

Verb Root + [NEG] + [INF/VBP/CONC/COND/HORT/OPT] + [EMP] + [INT/SUPP]
E.g. varaathathaal
vaa +aatha +athaal
come,V NEG COND
(because of not coming)

The gender may fall in one of the following categories: Male, Female, Male/Female, and Neuter. The common aspect includes: Perfective, Completive and Habitual whereas the mood can either be imperative or conditional.

4. Tamil Morphological Analyzer

The morphological analyzer processes each word from right to left fashion. If a suffix is accepted by FSA at the initial state then it progresses to the next state and proceeds subsequently till it reaches the final state where it finds the root word paradigm. If the morphemes are accepted at all states by the FSA then the word is considered as a valid word and the analysis is returned.

Independent and irrespective of the context, the primary task of morphological analyzer is to return all possible analysis results for each correct word assuming that root word for a given word exists in lexicon. Thus a single word can have analysis results in all possible lexical categories like noun, verb, adjective, adverb, preposition etc.

4.1 System Architecture

The system takes files where each line consists of a single word which is to be morpheme analyzed.

Each line in output file consists of the word and its morphemes. The word and the morphemes are tab separated while the morphemes are space separated. If a word has multiple analysis results then they are also included in the same line and they are separated by semicolons from the previous analysis results.

The morpheme analysis is done at two levels. At the first level, the words are morpheme analyzed by the morphological analyzer. The words that are returned as unanalyzed words in this level are given to the CRF word-boundary identifier. The words which fall in compound or agglutinated words category will be boundary identified.

CRF based word-boundary identifier is used for identifying and splitting the compound and agglutinated words at the constituent words' boundaries. The word boundary identifier makes use of CRF trained model for identifying the word boundaries. The splitted words are given to a Sandhi corrector. The function of the Sandhi corrector is to perform required Sandhi corrections to the words that are splitted by the word-boundary identifier.

After this step, these words are Sandhi corrected by linguistic rule-based Sandhi corrector. These words are then passed to the morphological analyzer for morpheme analysis. The words that are still unanalyzed after this level are returned as unanalyzed words itself.

For analyzing individual morphemes, the system makes use of the following resources: Tamil lexicon, FSA modeled from inflectional rules, Ontology information base, allomorpheme to morpheme mapping information base.

The overall system architecture is depicted in figure 1.

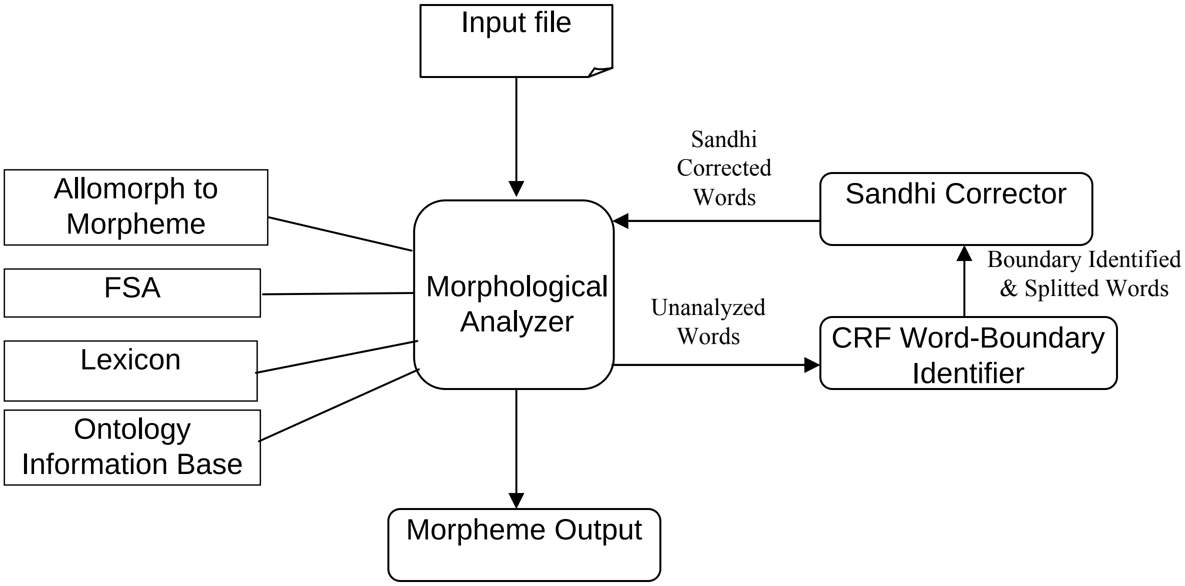

Figure.1. System Architecture

In the following sub sections we describe the linguistic resources that are used for developing the morpheme extraction system**.**

4.2 Linguistic Resources

The linguistic resources that are required for the development of Tamil morphological analyzer include a paradigm-classified lexicon, inflectional rules for all paradigms of lexical categories, ontology information base, AT&T FSM library.

4.2.1 Lexicon

Based on the root word ending and the inflectional pattern that the root word exhibits, the words in the lexicon are classified into various paradigms among their lexical categories. In Tamil we have 36 noun paradigms and 34 verb paradigms. We have used a lexicon having 137945 root words.

4.2.2 Inflectional Rules

The inflectional pattern depends on the root word endings and it is consistent for all the words in a particular paradigm. These inflectional rules form the crux of the morphological analyzer where it represents the needed morphophonemic changes for each of the paradigms of all lexical categories. The possible inflectional rules are prepared for all the paradigms of lexical categories and transformed into Non-Deterministic FSA(NDFSA) which is subsequently minimized to Deterministic FSA(DFSA) using AT&T FSM library.

A sample finite state automaton for some word categories is given in figure 2.

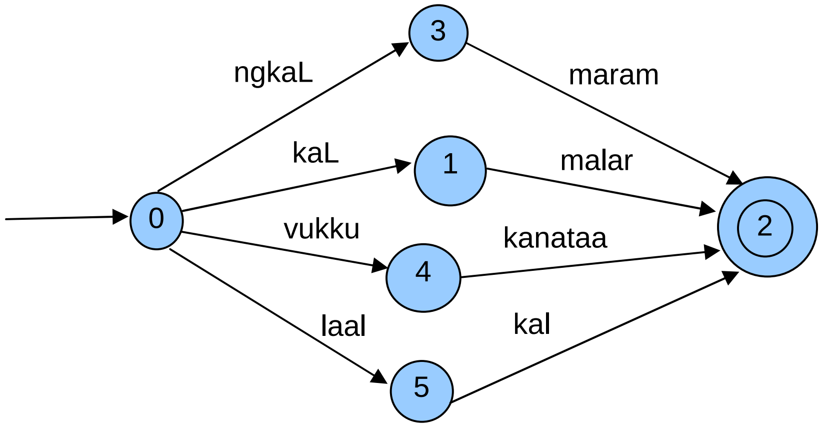

Figure.2. Sample Finite State Automaton

State table for the FSA in figure 2 is represented in table 2.

Table 2. State Table for the FSA in Figure 2

Current State	Next State	Symbol
0	1	kaL
0	3	ngkaL
0	4	vukku
0	5	laal
3	2	maram
1	2	malar
4	2	kanataa
5	2	kal

For the word “marangkaL”, the system starts from the initial state (here state 0) and based on the words’ suffix the machine progresses to the next state(here state 3) and then input is accepted at the next state which is also the end state (here state 2).

5. Linguistic Evaluation of Morphological Analyzer

The task organizers performed linguistic evaluation of the submitted system where samples of 1000 words have been used from the gold standard test set to extract pairs of words having common morphemes for comparison. Five random samples have been used for the evaluation and the Precision, Recall and F-measure values are calculated. Average of these values yielded the overall Precision, Recall and F-measure values which are tabulated in table 3.

Table 3. Performance Results

Experiment	***Precision (%)***	***Recall (%)***	***F-measure (%)***
Sample 1	83.89	87.92	85.86
Sample 2	83.06	87.40	85.18
Sample 3	86.45	90.06	88.22
Sample 4	83.92	87.80	85.82
Sample 5	84.11	87.57	85.80
Total	84.29	88.15	86.17

5.1 Error Analysis [Causes for failure of Word Analysis]

The unanalyzed words may fall in one of the following categories. The words in each category follow a certain writing style which caused the recognition failure. Further investigation is required to overcome these limitations. The possible causes for morpheme analysis failure are listed below.

1. Absence of inflectional rules
2. Uncommon transliterations
3. English acronyms
4. Errors in input words
5. Spoken language words

6. Discussion

For analyzing the unanalyzed English acronyms it is possible to maintain a list of common acronyms and perform a list lookup. Spoken language forms of words in Tamil are only a handful. This can be handled by performing a dictionary mapping between spoken and their equivalent written forms. Minimum edit distance based approaches can be applied to words which fall in the category of input errors. This approach may aid us to retrieve correct words which can eventually be morpheme analyzed.

7. Conclusion and Future Work

We presented a morphological analyzer for Tamil which got high precision and recall. Our approach can be extended to any morphologically rich and agglutinative Indian language provided the resources such as paradigm-classified lexicon, morphotactics of the language to model the FSA are made available. Some of the problems for the failure of word analysis can be solved using standard techniques as mentioned in discussion section which may subsequently increase the accuracy.

References

Antworth, E. L., 1990. *PC-KIMMO: A Two-level Processor for Morphological Analysis,* occasional Publications in Academic Computing. Summer Institute of Linguistics, Dallas, Texas.

Eryiğit, Gülşen and Adalı,Eşref. 2004.*An Affix Stripping Morphological Analyzer for Turkish,* In IASTED International Multi-Conference on Artificial Intelligence and Applications. Innsbruck, Austria, pages 299–304.

Jurafsky, Daniel and Martin, James H. 2002*. Speech and Language Processing: An Introduction to Natural Language processing, Computational Linguistics and Speech Recognition.*

Koskenniemi, Kimmo. 1983. *Two-level Morphology: a general computational model for word-form recognition and production,* University of Helsinki, Helsinki.

M Anand kumar, V Dhanalakshmi, K P Soman and S Rajendran. 2009. *A Novel Approach For Tamil Morphological Analyzer*, Proceedings of Tamil Internet Conference, Cologne, Germany, Page no:23-35.

S. Viswanathan, S. Ramesh Kumar, B. Kumara Shanmugham, S. Arulmozi and K. Vijay Shanker. 2003. *A Tamil Morphological Analyzer*, ICON-2003, pp. 31-39

Taku Kudo. 2005. CRF++, an open source toolkit for CRF, *http://crfpp.sourceforge.net*

Vijay Sundar Ram R, Menaka S and Sobha Lalitha Devi (2010), "Tamil Morphological Analyser", in "Morphological Analysers and Generators", (ed.) Mona Parakh, LDC-IL, Mysore, pp. 1 –18.