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

Sobha Lalitha Devi, Marimuthu K, Vijay Sundar Ram R, Bakiyavathi T and Amudha K

AU-KBC Research Centre,

Chrompet, Chennai.

Abstract

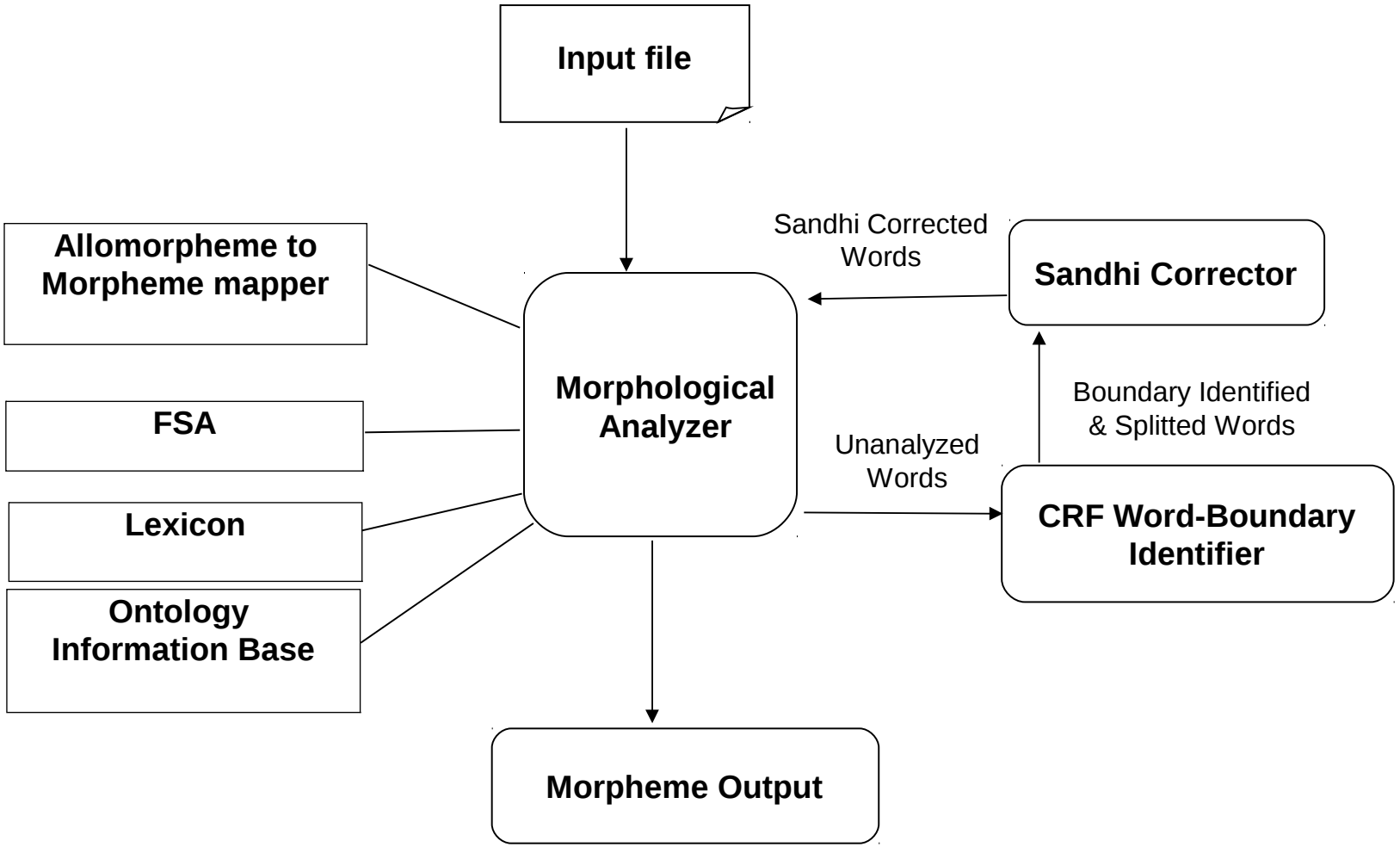
- We present an efficient morphological analysis system for Tamil
- The individual morphemes are extracted using Finite State Automaton (FSA).
- Morpheme analysis is done by modeling the regular inflectional pattern exhibited by Tamil as an FSA.
- Handled the compound and agglutinated words using CRF word-boundary identifier.
- On linguistic evaluation the system achieved an encouraging accuracy of 86.17%

Introduction

Tamil is a highly inflectional Dravidian language.

- A verb-final language with a relatively free word order.
- Higher degree agglutination is a common phenomenon.
- Features of morphological analysis output –
  - vital information required for various NLP applications
  - Machine translation, Information retrieval systems, Anaphora resolution ……

System Architecture



Linguistic Resources

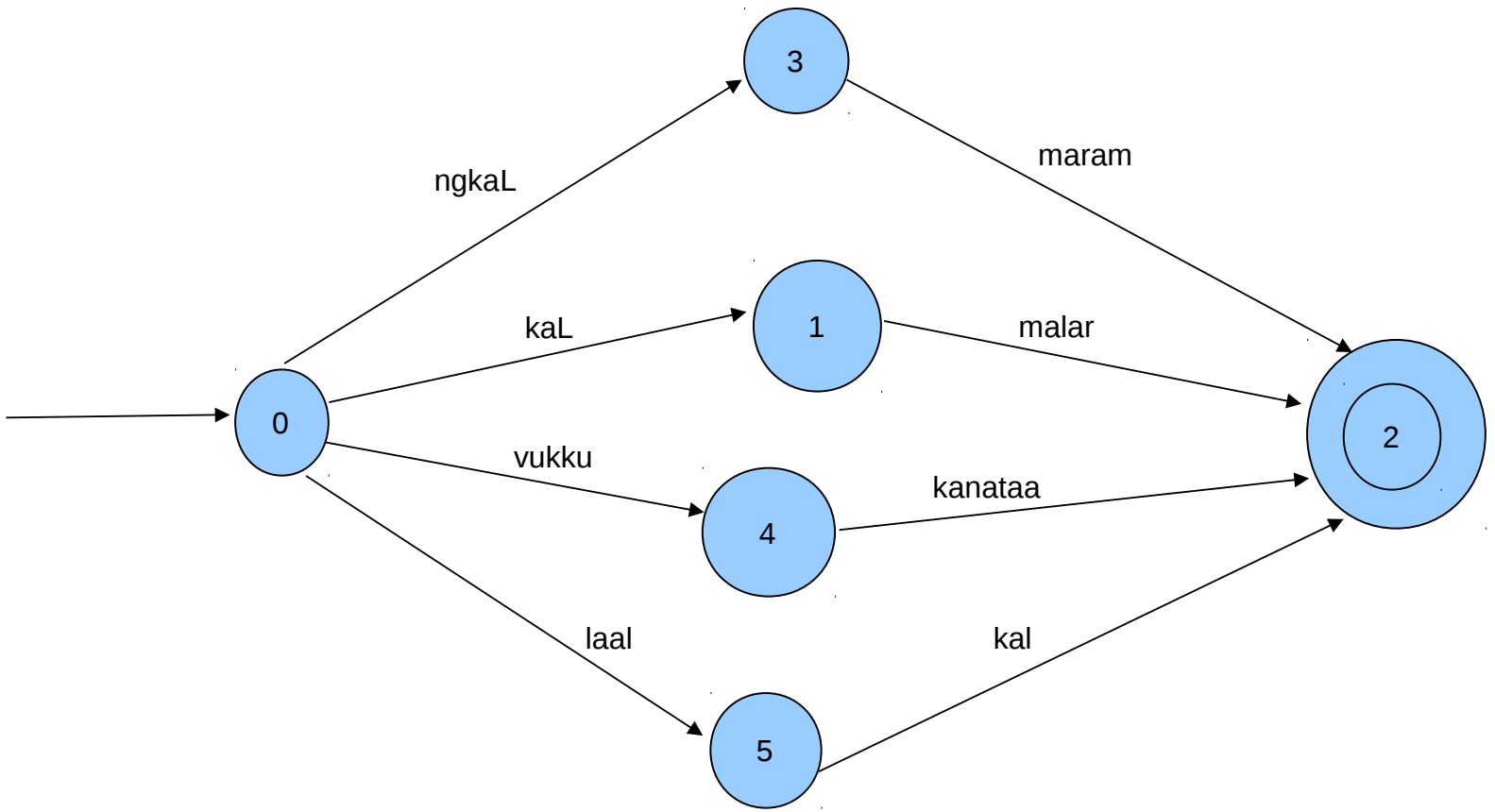
Linguistic Resources used for the development of Tamil morphological analyzer include:

- *Lexicon*
- *Inflectional Rules*
- *Ontology information base*
- *Allomorpheme to Morpheme mapper*

CRF Word-Boundary Identifier

- Used for identifying and splitting the compound and agglutinated words at the constituent words' boundaries using CRFs.
- The splitted words are given to a Sandhi corrector.
- Sandhi corrector performs required Sandhi corrections to the words that are splitted.
- After splitting, these words passed to the morphological analyzer

Sample FSA



Analysis Algorithm

1: For each input word (W)
2: do morphological analysis
2.1 if (W) exists in lexicon (Lex)
return (W) + grammatical categories
2.2 else check suffixes (S1 ... Sn) in FSA
2.2.1 if FSA accepts (S1 ... Sn)
return (W) + grammatical categories
2.2.2 else (UW) = return (W)
3: For each unanalyzed word (UW)
3.1 Identify word-boundaries of constituent words (CW)
3.1.1 For each (CW)
3.1.1.1 do Sandhi correction
3.1.1.2 go to step 2.1
3.2 For (CWs) with Incorrect word-boundaries
3.2.1 return (CWs') original word

Linguistic Evaluation

Table 1. Affix and Non-Affix Performance Results

Partitions	*Precision(%)*		*Recall(%)*		*F-measure(%)*	
	Non-Affix	*Affix*	*Non-Affix*	*Affix*	*Non-Affix*	*Affix*
Sample 1	85.83	83.41	97.06	85.79	91.10	84.59
Sample 2	83.83	82.86	85.97	87.85	84.89	85.28
Sample 3	85.51	86.68	88.70	90.40	87.08	88.50
Sample 4	92.35	81.79	89.17	87.31	90.73	84.46
Sample 5	91.68	82.69	90.43	86.85	91.05	84.72
Total	**87.84**	**83.49**	**90.27**	**87.64**	**89.04**	**85.51**

Linguistic Evaluation

Table 2.Performance Results - Standard Metrics

Partitions	*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**

Error Analysis
[Causes for failure of Word Analysis]

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

Conclusion

- We presented a morphological analyzer for Tamil which achieved a high precision & recall values.
- Agglutination and Compound words are handled using CRF based word-boundary identifier
- Our approach can be extended to any morphologically rich and agglutinative language provided the resources such as
  - paradigm-classified lexicon
  - morphotactics of the language to model the FSA are made available.

Thank You !