Tamil Morphological Analyser

Vijay Sundar Ram R, Menaka S and
Sobha Lalitha Devi

AU-KBC Research Centre. MIT Campus
of Anna University, Chennai-44

Presentation Outline

- Introduction
- Previous Works
- A brief description on Tamil morphology
- Our Approach
- Evaluation and Discussion

Introduction

- Morphological analysis of a word is the process of segmenting the word into component morphemes and assigning the correct morphosyntactic information.

- For a given word, a morphological analyser (MA) will return its root word and the word class along with the other grammatical information depending upon its word class.

- MA returns all possible parse for a given word, without considering the context.

Introduction (Contd…)

- MA is a very essential for languages having rich inflectional and derivational morphology such as morphologically rich languages such as

  - Dravidian languages (Tamil, Telugu, Malayalam and Kannada),
  - Finno-Ugric languages (Finnish, Estonian, Hungarian, Turkish),
  - Indo-Aryan languages (Hindi, Bengali, Marathi, Gujarati).

Previous Works

- There are several approaches attempted for MA.
- The two-level morphology approach by Kimmo Koskenniemi is the early attempts.,
  - two-level representation,
    - surface level is to describe word forms as they occur in written text
    - lexical level to encode 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 Automata.
- This approach is used for recognizing and generating word forms.

Previous Works (Contd…)

- A rule based, heuristic analyser for Finnish nominal and verb forms was developed by Jappinen (Jappinen 1983).

- A word-grammar based morphological analyser for agglutinative languages was introduced by Agirve (Itziar 2000), here they have worked on Basque, a highly agglutinative language.

- Arabic Finite State Transducer for morphological analysis using Xerox Finite State Transducer (XFST) was built by Beesley.

Previous Works (Contd…)

- Finite State Automata based MA was developed in Tamil.
- In Bengali, unsupervised methodology is used in developing a MA (Sajib Dasgupta, 2007) and two-level morphology approach was used to handle Bengali compound words.
- There are rule based MA for Sanskrit (Girish Nath Jha 2007) and Oriya (Mohanty 2004).

Tamil Morphology

- Tamil belongs to the South Dravidian family of languages.
- It is a verb-final and a relatively free word order language.
- It has a rich inflectional and derivational morphology.

Tamil Morphology (Contd…)

- Agglutination is one of features of this language.
- When suffixes attach to the root several orthographic changes take place.
- The order in which suffixes attach to a root form determine the morphosyntax of the language

Our Approach

- Morphological analysis of Tamil, morphologically rich language using Finite State Automata (FSA) and the paradigm approach.

- FSA are the proven technology for efficient and speedy processing.

Our Approach (Contd…)

- FSA using all possible suffixes is built.
- Categorize the root word lexicon based on paradigm approach to optimize the number of orthographic rules.
- Morphosyntax rules for correct analysis for the given word.
- The analysis of the word is done suffix by suffix.

Limitations of two-level morphology

1, Developing Finite State transducers that encode very complex two-level rules is not easy.

2, morphological categories are not directly encoded as a part of the lexical form.

3, lexical representation tends to be arbitrary.

4, various diacritical features inserted into the lexical strings to insure proper analysis makes Kimmo-style awkward or impractical for generation (Beesley 1996).

Finite State Automata (FSA)

- FSA is an abstract device used for recognizing simple syntactic structures or patterns.
- Depicted by directed graph, called State Diagram and in a tabular form as State Table.
- From a mathematical perspective it is regarded as a function, mapping a set of string to the set {Accept, Reject}.
- Based on the transition given by the FSA, it is classified as Non-deterministic FSA (NDFSA) and deterministic FSA (DFSA).

Modeling of Suffix based FSA

- FSA is modeled using all possible suffixes ie., all the allmorphs.
- FSA is built by considering the suffixes from left to right of the word, ie moving from end of the word towards the root word.
- Whenever the transition is triggered by the suffix, that suffix is stripped from the word and required orthographic corrections are done.

Modeling of Suffix based FSA (Contd…)

Sample of the State Table

Current State	Next State	Symbol
0	0	ai
0	0	utaiya
0	1	kal
0	1	ai
0	1	utaiya

Orthographic Rules in FSA

- Orthographic rules are the spelling rules used to model the changes that occur in a word, usually when morphemes are combined (Jurafsky 2000).
- The characters that are deleted from the root word or the suffix, when a suffix (allomorph) is affixed, it is stored after the suffix in the state table.
- Example

0	0	atu	a

makanuTaiyatu = makan + uTaiya + atu.

Root Information in FSA.

- After the orthographic correction characters the category of the root is added in the state table.

0 1 ńkaL m N13

marańkaL = maram + kaL

Lexicon

- The root words into different groups, where every word in each group will have similar orthographic changes (sandhi changes), when a suffix is added to it.

- We have categorized noun into 36 paradigms and verbs into 34 paradigms.

- The lexicon has 44055 root words.

Morphosyntax Rules

- A set of rules that explains which classes of morphemes can follow other classes of morphemes inside a word.
- Root -> plural marker -> case marker -> clitc.
- This set of rules filter out the correct parsing of the word from the FSA.
- Here we have 286 rules.

Handling of Compound Words

Steps involved in Handling of Compound Words

Step 1: Parsing the suffixes from the last suffix to the first suffix in the word, and checks for the root word in the given category in the FSA.

Step 2: If the root word is not matched then step 3

Step 3: The root word is split based on syllables and checked with the root dictionary

Step 4: Once a word is matched, the remaining part of the word is splitted similarly and compared with the root dictionary.

Step 5: If the complete root word, is matched into different root words in the dictionary, this multiple words as root with suffix information is given as analysis.

Step 6: If the complete root is not matched even after splitting into multiple words, the analysis is given as unknown word.

Handling of Agglutinated verbs

Verb which is inflected, agglutinated with the pronoun,

vaŃtavan -> va: + Ńt + a + avan

come+root past RP pronoun

Agglutination of inflected verb and verb illai (negation)

- the verb illai agglutinate with the infinite verb forming one word, such as

varavillai -> va: + a + illai

come+root inf negative verb

Evaluation

- We have evaluated the system with two sets of web data,
  - first set: general domain
  - second set: tourism domain.
  - 50K words from each domain.

Evaluation (Contd…)

Types	General Domain	Tourism Domain
Total number of Words	50,000	50,000
Analysed words	46620	45085
Error due to Missing morphosyntax rules and state table entries	223	344
Error due to agglutination	485	531
Error due to missing root word	1345	1987
Input Error	1327	2053
Correctness of analysis	93.24%	90.17%

Evaluation (Contd…)

- The tourism documents have more compound words and the agglutination of words is more.
- There are more number of named entities such as person name, place name, area specific words.
- The sentences commonly end with a:kum, a copula verb.

Evaluation (Contd…)

Similarly there are more compound nouns, such as

maNme:TukaLuTaiya -> maN+me:Tu+ kaL + uTaiya

sand dune pl genetive

Compound root suffix

Thank You !!!