Morph and POS Tagger for Tamil: A Rule based Approach

M. Ganesan
CAS in Linguistics
Annamalai University
ganesan_au@yahoo.com

Tamil

- an agglutinative language
- morphologically rich & complex
- A word can have 11 morphs
- A verb can conjugate to 1600 word forms

WORD

- MORPH : A minimum meaningful unit in a language
  1. Free morph
  2. bound morph

WORD : A sequence of characters between two successive spaces

A word may be

Root

Prefix + root

Stem + suffix(es)

Stem + stem + suffix(es)

Prefix + stem + suffix(es)

POS Tagging

- Form Vs Function
- Purpose
- Basic POS Tags: 7
  - » NN Noun
  - » PN Pronoun
  - » FV Finite Verb
  - » NV Non finite Verb
  - » AJ Adjective
  - » AV Adverb
  - » IN Indeclinable

Tag sets

- Elaborate (in hundreds)
- Medium
- Small

- Agglutinative languages need elaborated one

Level of Tagging

word level

morph level

- Advantages of morph level tagging
  - to identify the syntactic role of the word
  - in parser development
  - word sense disambiguation

Methods of Tagging

- manual
- automatic
- hybrid

Approaches in Automatic Tagging

- Rule based
- Stochastic
- Neural

AUTOMATIC TAGGING - Rule based

It should do:

- identifying a word
- segmenting the word for its components
- labeling the components for their lexical / grammatical values
- specifying the grammatical value of the word in a given sentence

AUTOMATIC TAGGING

Problems:

- Words are always not simple; compound & conjoined
- Internal and external sandhi
- Inconsistency in writings
  - more than one spelling
  - with / without sandhi operation
  - without / with spacing in conjoined words
- Words having different grl. categories
- Affixes having more grl. functions

Post - Editing

- have possible, multiple tags
- eliminate wrong tags by context
  - automatic
  - manual
- tag the untagged manually
- tools for post-editing

Advantages of Automatic Tagger

- consistent
- more economy
- time saving

Offshoots

- Morphological Analyzer
- Parser
- Spell checker
- Grammar checker
- Lemma extractor
- Paradigm generator