Malayalam Stemmer

Vijay Sundar Ram R, Pattabhi R K Rao T
and Sobha Lalitha Devi
AU-KBC Research Centre, Chennai

Introduction

- Stemming is the process of getting the stem for a given word by the removal of suffixes affixed to the root word by derivational and inflectional process.
- Used in information retrieval task as a recall-enhancing device.
- The stemming differs from lemmatization, as the stem generated may not necessarily be a lemma (syntactic root word).

Introduction (Contd…)

- For the word *marattil* (tree+loc) in Malayalam,
- When stemmed,
  - the removal of the location case suffix *-il*,
  - the stem is *maratt*, (an oblique)
- Here *maram* is the root word.

Previous Works

- **Julie Beth Lovins (1968)**
  - One of the oldest published works on stemmers
  - rule based stemmer
  - a single pass, context–sensitive, longest match stemmer
  - removes a maximum of one suffix from a word
- **Porter's stemming algorithm (1980)**
  - used widely in different IR systems for English
  - has 60 suffixes, two recoding rules and a single type of context-sensitive rule to determine whether a suffix should be removed
  - uses a minimal length based on the number of consonant-vowel-consonant strings remaining after removal of a suffix
- **Statistical stemmer for Spanish**
  - Buckley et al. (1995)
  - simple stemmer by examining lexicographically similar words to discover common suffixes.

Previous Works

- **Statistical Stemmer – Goldsmith (2000)**
  - suffix discovery from language sample by
    - employing automorphology
    - a minimum-description-length-based algorithm
    - highly computationally intensive

- **Statistical Stemmer - Oard et al (2001)**
  - Suffix discovery from text collection
    - end n-grams frequencies of the strings were counted (where n = 1, 2, 3, 4) for the first 500,000 words of the text collection
    - the frequency of the most common subsuming n-gram suffix was subtracted from the frequency of the corresponding (n-1)-gram

Previous Works

- **Xu and Croft (1998)**
  - analyzing the co-occurrence of words
  - use a variant of expected mutual information to measure the significance of the association of words
  - developed for Spanish
- **Roeck and Al-Fares (2000)**
  - developed for Arabic
  - use dice coefficient to measure string distance
  - cluster the result to generate equivalence classes of words
- **Rogati et al. (2003)**
  - developed for Arabic
  - use a machine learning approach

Previous Works

- **Ramanathan and Rao (2003)**
  - developed for Hindi
  - uses rule based approach
  - use a handcrafted suffix list
  - suffixes are eliminated from word endings based on some rules
- **YASS (2007)**
  - Majumder et al.,
  - developed for Bengali
  - use a clustering-based approach to discover equivalence classes of root words
  - a set of string distance measures are defined, and the lexicon for a given text collection is clustered using the distance measures to identify these equivalence classes.

Our Approach

- Constructed a stemmer based on the principle of iteration, as the suffixes are added to the stem in a order, which is governed by the morphotactic rules.

- This strict rule based word formation helps in building a Finite State Automata (FSA) of suffixes.

Our Approach (Contd…)

- FSA is built using all possible suffixes, where the next state is determined using the morphotactic rules of the language.

- The orthographic variation during the affixation of the suffixes is also handled in the FSA.

Finite State Automata (FSA)

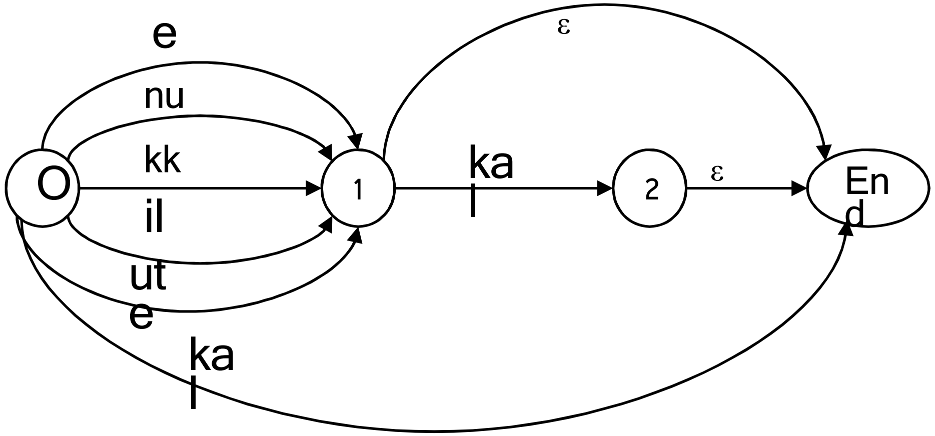
- Finite State Automata is a model of behavior composed of a finite number of states and transitions between these states.
- Recognizing simple syntactic structures or patterns.
- An automaton is normally depicted by directed graph, called State Diagram and it is also represented in a tabular form as State Table.

Modeling of Suffix based FSA

- FSA is modeled using all possible suffixes ie all allomorphs.
  - where allomorphs are defined as a morpheme that is manifested as one or more morphs in different environment.
  - Eg. *u, i* are the allomorphs of the past tense marker in Malayalam.
- Here the FSA is built by considering the suffixes from left to right of the word .

Modeling of Suffix based FSA

Sample State Diagram



Sample State Table

Current State	Next State	Transition Symbol
0	1	nu
0	1	kk
0	1	il
0	1	ute
0	1	e
0	3	kal
1	2	kal
1	3	e
2	3	e
3	endstate	

Oblique stem to root - Using Sandhi Analyzer

- Most of the applications such as information extraction, machine translation, named entity recognition require the root form of the given word
- Use a sandhi analyzer to generate root form of the word from the oblique form
- The sandhi analyzer consists of a set of sandhi rules
- This analyser performs the orthographic changes required to produce the root word.

Oblique stem to root (Contd…)

For example

- *marattil*
  - – the stemmer gives
    - *maratt* (oblique stem).
  - – The sandhi analyser produces
    - *maram* (Root)

Evaluation

- A set of words collected from online Malayalam newspaper, *Mathrubhumi*
- The input words are classified into three classes
  - Nouns with case markers
  - Nouns with Plural marker and case makers
  - Verbs
- We obtain an average accuracy of 94.76% from the stemmer
- The sandhi analyzer generates correct root forms from the oblique form with an accuracy of 95.83%, if correct oblique forms are given as input
- Whereas the accuracy of the sandhi analyser with incorrect oblique forms as inputs is 90.5%

Evaluation

- On analysis of test data, we found that many of the words are formed by the agglutination of more than one word
  - For example
    - *Avana:yirunnu*
      - *avan+aiyirunnu*
        pronoun+ copula

        ‘It was he’
- For such the stemmer failed to give correct oblique form
- Such words require to be properly segmented before giving those as input to Stemmer
- A word segmentation module is required

Evaluation

- Evaluated with a set of words collected from online Malayalam newspaper, *Mathrubhumi*.

Type Of Words	No. of Words	Correct Oblique Forms Generated		Correct Root Forms Generated after using Sandhi Analyser			
				With Error Stems		Without Error Stem	
Word + Case Marker	1000	956	95.6 %	914	91.4 %	918	96.02%
Word + Plural + case marker	1000	962	96.2 %	918	91.8 %	923	95.95%
Word + Tense + Auxiliary	1000	919	91.9 %	883	88.3 %	883	96.08%
Total	3000	2843	94.76 %	2715	90.5 %	2724	95.83%

Summary

- A stemmer for Malayalam, a morphologically rich language using Finite State Automata, as the word formation is strictly based on the morphotactic rules.
- Performs with an accuracy of 94.76 %.
- Oblique stem are converted to root using a sandhi analyser.

Thank You !!