Sandhi Splitting

Automatic Sandhi Splitting Method for Telugu

Introduction

Word Segmentation :

Word segmentation is the problem of dividing a string of written language into its component words which involves finding the respective word boundaries.

In English and many other modern languages finding the word boundaries is a bit easy because of the presence of white spaces or punctuation between words.

However, it is not straightforward in many Asian languages such as Chinese, Japanese or in Indian Languages and especially for agglutinative languages like Telugu as they do not delimit words by white-space.

Introduction

Sandhi :

Sandhi is a process in which two or more words unite to form a compound word by undergoing some modification in the resulting word.

The modification is seen at the position of interaction of the constituent words as they are influenced by adjacency.

Linguistically if two word's are uniting, it is seen in the last syllable of the left-hand side word and first syllable of the right-hand side word.

Most of the Times sandhi is

Sandhi = Noun + Verb (rAmudannAdu = rAmudu + annAdu)

Pronoun + Verb (bAludannAdu = bAludu + annAdu)

Approach

1. Extracting base words and compound words and breaking them into their constituent syllables.
2. Building both Top-down (TD) and Bottom-up (BU) finite state transducers of those syllables, where each state corresponds to a syllable for faster searching.
3. For each compound word, traverse through the transducers both TD and BU and find the various possible syllables that has undergone sandhi. Scoring the possible syllables w.r.t. the compound word syllable and finally giving the possible outcomes along with a best outcome.

Breaking Words into Syllables

Syllable :

A Syllable is made up of consonants and vowels.

In most of the cases Syllables ends with vowels.

The Vowels that we used for splitting string into Syllables is

a, e, i, o, u, A, E, I, O, U.

Ex:

Verb : voccAdu (came)

Syllables : vo + ccA + du

Noun : rAmudu (Ram)

Syllables : rA + mu + du

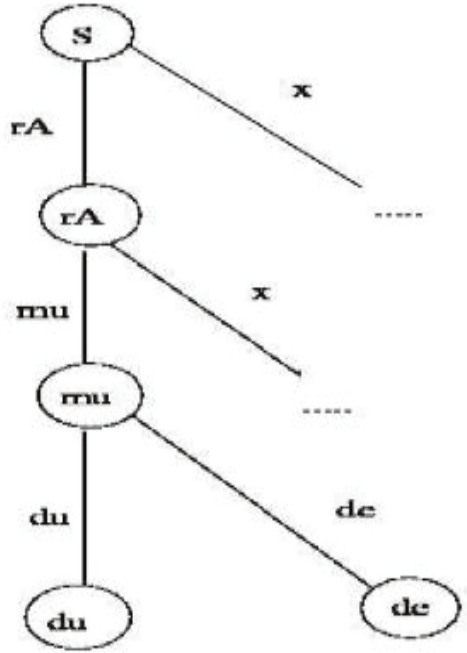
Building Finite State Machine

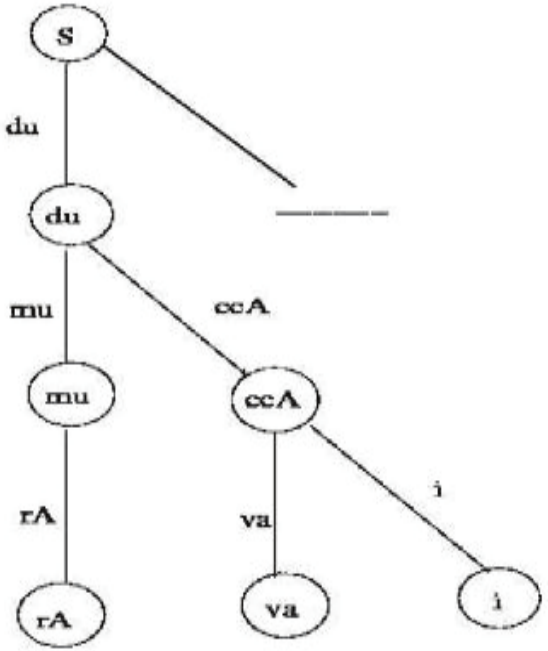
From the syllable-based base words we build the finite state machines, both top-down and bottom-up , where each state corresponds to a syllable.

The mathematical model for the automata is $(S, \Sigma, \delta, s_0, F)$ where,

- S is the set of States (Syllables).
- Σ is the input compound word delimited at syllable endings (a set of syllables).
- δ is the state-transition function $\delta : S * \Sigma \rightarrow S$.
- s_0 is the initial state.

Building Finite State Machine





Sandhi Splitting

Splitting is the process to find the possible base words that constitute the given compound word. This is achieved by traversing through both the automata (Top Down and Bottom Up).

For example, to split a compound word, rAmudoccAdu (Ram came), into its consituent words the traversing through the automatas is shown in above figures.

A few possible end states for Top-down and Bottom-up are also shown in the figures.

we will be left with few possible outcomes for the constituent words. we use a Probability based scoring method to score the possible outcomes.

Scoring Method

From the list of possibles, we compute the joint probability between each of the TD-list syllables and BU-list syllables using the below scoring function :

Let X , Y denote the TD and BU lists.

Therefore, the joint-probability is given by :

$P(X = x, Y = y) \Rightarrow P(X = x|Y = y) * P(Y = y) \Rightarrow P(X= x) * P(Y = y)$

(Since X & Y are independent to each other)

where,

- $x \in$ Top-Down list.
- $y \in$ Bottom-Up list.

Transition Method

In case of Pronoun and verb Interaction we can use Scoring Approach, but in case of nouns they are unlimited.

To split Nouns we need to compute transitions from compound syllable to constituent word syllables.

Ex : Transition do → du + vo

Compound Word : rAmudoccAdu (Ram Came)

Syllables : rA + mu + do + ccA + du

Verb : voccAdu (came)

Syllables : vo + ccA + du

Noun : rAmudu (Ram)

Syllables : rA + mu + du

Conclusion

It can split compound words which occur with different possible combinations of constituent words that are seen in the corpus.

Reference

Automatic Sandhi Splitting Method for Telugu, an Indian Language by Phani Chaitanya Vempatya and Satish Chandra Prasad Nagalla.