Sandhi Splitting of Marathi Compound Words

Joshi Shripad S.

Department of CSE, JNT University, Hyderabad, Andhra Pradesh, India
E-mail : joshi.jntu@gmail.com

Abstract – **Sandhi splitter is an important module for Natural Language (NL) system for Marathi in which words can be combined together to form a larger string of words. The research in Natural language processing is being carried out in variety of areas like speech processing, text analysis, text processing, text mining etc. Among all there is a need of analysis of word in the given language. The formation of word may be the result of combination of two or more words. Separation of the words in meaningful sub-words is sandhi splitting (Sandhi-Vichcheda). In this paper we are presenting the rules and the rule based algorithm for sandhi splitting of Marathi compound words.**

Keywords – Compound Marathi Words, Rule Based Algorithm, Sandhi-Vicheda,

I. INTRODUCTION

Natural Language Processing (NLP) requires many pre-processing stages to analyze, understand and generate Natural Language(NL). Each of the stages may form a subtask which itself can be used in various applications. Sandhi-splitting is one such subtask that forms a pre-requisit for complete analysis of input text in NL and used in various NLP applications like Machine Translation Systems(MTS), tagging of large text corpora, spell checker, building a NL text search engine etc.

The word sandhi refers to a wide variety of phonological changes at morpheme or word boundary in which two letters combine and they have certain changes. External and internal are the two processes of sandhi. Internal process governs the combination of suffix with root or stem in declension, conjugation and derivation. External sandhi is a linguistic phenomenon which refers to a set of sound changes that occur at word boundaries.

External sandhi formation can be orthographically reflected in some languages (e.g. Marathi). When two words are combined, the uniting letters i.e. the final letter of the first word and the initial letter of the second undergo a change, thus एक (one) + ईश्वर (lord) = एकेश्वर (one lord), the अ andई coalescing into ए. These changes of letters are denominated Sandhi or combination by Sanskrit grammarian.

The laws of Sandhi belong either to the union of the vowels or to the union of the consonants, the former being denominated *Ach Sandhi (the combination of the vowels)* and the latter the *Hal Sandhi* (the combination of the consonants)[1].

In the paper, we have proposed a rule based algorithm for sandhi splitter (Ach sandhi only) of Marathi words.

1.1 The Marathi Language

Marathi is an Indo-Aryan language. It is the official language of Maharashtra and Goa and is one of the 23 official languages of India. It is the 19th most spoken language in the world. With its about 90 million speakers (70 million native speakers plus about 20 million second language speakers) it is comparable in rank with languages like Korean or Vietnamese.

Marathi has the fourth largest number of native speakers in India Marathi has some of the oldest literature of all modern Indo-European, Indic languages, dating from about 1000 AD. The major dialects of Marathi are called Standard Marathi and Warhadi Marathi. Standard Marathi is the official language of the State of Maharashtra. Marathi is written in the Devanagri script and draws much of its vocabulary from Sanskrit. The language derives its grammar and syntax from Sanskrit and is therefore one of the Indo-Aryan languages. Marathi is normally spoken using a combination of 49 sounds, 13 vowels, 36 consonants.

Marathi morphology makes use of agglutinative, inflectional, and analytic forms. In natural language processing, languages with rich morphology pose problems of quite a different kind than isolating languages. In the case of agglutinative languages, the main obstacle lies in the large number of word forms that can be obtained from a single root. Marathi uses many morphological processes to join words together, forming complex words. These processes are traditionally referred to as sandhi (from Sanskrit, "combination"). For example, भोजन + आलय gives the word भोजनालय.

II. RELATED WORK AND BACKGROUND

The work related to sandhi processing has been done by French scholar Gerard Huet[2]. An online program named “ The Sanskrit Reader Companian” has been built for segmenting and tagging simple Sanskrit phrases. It does sandhi-vichcheda for segmenting by simple string segmentation applying string de-concatenation techniques and gives multiple results in many cases.

The Technology Development for Indian Languages (TDIL) program of the Ministry of Information Technology (MIT), Govt. of India, in one of its project named ‘Computer Assisted Sanskrit Teaching & Learning Environment’ (CASTLE)[3] funded for Jawaharlal Nehru University, New Delhi claims to have developed a sandhi- viccheda system which takes a word as input and returns the constituent words in the DOS environment. But this work is also not available for download anywhere on the TDIL website.

Amba Kulkarni, in her Anusaaraka project[4] at Rashtriya Sanskrit Vidyapeetha, Tirupati has developed a sandhi analyzer system. Its methodology is that using the sandhi rules, the programme splits the given word into two words and then checks whether the two words are recognized by a morphological analyzer. If any of the words is not recognized, the sandhi split function is called recursively.

Sanskrit sandhi analyzer[5] developed by Sachin kumar deals with a vowel sandhi. the sandhi rules were formalized for reverse automation. Various linguistic resources for sandhi recognition and analysis were also developed and adapted with certain limitations.

Telugu Spell Checker[6] developed by Uma Maheshwar Rao et. al. is incorporated with external sandhi splitter algorithm which uses the approach Generate-Analyze-Constrain-Evaluate.

The rule Based Algorithm for Sandhi-Vicheda Of Compound Hindi Words developed by Priyanka Gupta,Vishal Goyal[7] have reported an accuracy of 60-80% depending upon the number of rules to be implemented.

Sandhi splitter or sandhi analyzer for Marathi as a separate module has not been reported yet.

III. IMPLEMENTATION

We have implemented the Rule-Based algorithm to takes compound words and splitting the words into meaningful words and identifying the rule used according to the Marathi grammar Sandhi-Vichcheda rules.

Algorithm:

Step 1: Take the Marathi word

Word=’ सूर्यास्त’

Step 2: Dividing into syllables

Syl=[’स’,’◌ू’,’र’,’◌्’,य’,’◌ा’,’स’,’◌्’, ,‘त’]

Step 3: Adding one syllable to form a string and check in database.

Step 4: If the string matched with the word in database then first word is the string then use the following rules to add the vowel to the second word and goto step 5 else goto Step 3.

1st Rule: a“अ” With a “अ”:

सूर्यास्त	सूर्य+अस्त
मही	म+अही
बीजगणित	बीज+अगणित
कटाक्ष	कट+अक्ष

2nd Rule: a“अ” With a “आ”:

भोजनालय	भोजन+आलय
पुस्तकालय	पुस्तक+आलय
देवालय	देव+आलय
धनादेश	धन+आदेश

3rd Rule: A“◌ा” With a “अ”:

विद्यार्थी	विद्या+अर्थी
अपेक्षाभंग	अपेक्षा+अभंग

4th Rule: A“◌ा” With a “आ”:

शिवालय	शिवा+आलय

विद्यालय	विद्या+आलय
महिलाश्रम	महिला+आश्रम

5th Rule: A“ि” With a “उ”:

अत्युक्ती	अति+उक्ती
अत्युत्तम	अति+उत्तम

6th Rule: A“ी” With a “ई”:

महीश	मही+ईश
रजनीश	रजनी+ईश
अमलेश्वर	अमल+ईश्वर

7th Rule: A“ु” With a “उ”:

गुरुपदेश	गुरु+उपदेश

8th Rule: A“ु” With a “उ”:

भूद्धार	भू+उद्धार

9th Rule: A“अ” With a “ इ”:

ईश्वरेच्छा	ईश्वर+इच्छा

10th Rule: A“अ” With a “ई”:

गुणेश	गुण+ईश
अंगदेश	अंगद+ईश

11th Rule: A“ा” With a “ई”:

उमेश	उमा+ईश

12th Rule: A“अ” With a “उ”:

चंद्रोदय	चंद्र+उदय
करोटी	कर+उटी
कपालमाळा	कपाल+उमाळा

13th Rule: A“अ” With a “ऋ”:

देवर्षी	देव+ऋषी
अग्रणी	अग्र+ऋणी
अंजिणी	अंज+ऋणी

14th Rule: A“आ” With a “ऋ”:

परमैश्वर्य	परम+ऐश्वर्य
देवैश्वर्य	देव+ऐश्वर्य

15th Rule: A“अ” With a “ओ”:

जलौघ	जल+ओघ
कलमकारी	कलम+ओकारी
करारनामा	करार+ओनामा
कडोसरी	कड+ओसरी

16th Rule: A“अ” With a “औ”:

वृक्षौदार्य	वृक्ष+औदार्य

17th Rule: A“ि” With a “आ”:

अतिक्रमण	अति+आक्रमण
इत्यादी	इति+आदी

18th Rule: A“ी” With a “अ”:

प्रीत्यर्थ	प्रीती+अर्थ

Step 5: Display the result with Sandhi name.

Our module has been developed in **python3.3** environment The web interface also has been given using Python. We are using Marathi dictionary for word checking.

IV. EXPERIMENTS AND RESULTS

We have tested our Tool on more than 150 words. Using the Rule based algorithm we have reported an accuracy of 70-80% depending upon the number of rules that has been implemented.

V. CONCLUSION

In this paper, we have presented the rule-based technique for sandhi-vichcheda (Ach sandhi) of Marathi compound words. Using the rule based algorithm, we have reported an accuraly of 70-80% based on the rules that have been considered. We can increse the accuracy by incorporating more words in the database used.

VI. REFERENCES

[1] Ganpatrao Navalkar,2001. 'The Student's Marathi Grammar', AES Publication.

[2] Gerard Huet. 'Towards Computational Processing of Sanskrit', http://pauillac.inria.fr/~huet/PUBLIC/icon.pdf

[3] TDIL, MIT, GOI website, http://tdil.mit.gov.in/nlptools/ach-nlptools.htm

[4] Amba Kulkarni . Anusaaraka, http://www.iiit.net/research/ltrc/Anusaaraka/anu_home.html.

[5] Kumar, Sachin, Girish Nath Jha. 2005, “A Paninian Sandhi Analyzer for Sanskrit” In the Souvenir Abstracts of Platinum Jubilee International Conference of the Linguistic Society of India, University of Hyderabad, Hyderabad, pp. 36-37.

[6] Uma Maheshwar Rao, Amba P. Kulkarni, Christopher Mala, Parameshwari K. 'Telugu Sell-Checker', http://sanskrit.uohyd.ernet.in/faculty/amba/PUBLICATIONS/ITIC-ss.pdf.

[7] Priyanka Gupta,Vishal Goyal. Implementation of Rule Based Algorithm for Sandhi-Vicheda Of Compound Hindi Words, International Journal of Computer Science Issues, Vol. 3, 2009, pp. 45-49.

◈◈◈