Designing a Sanskrit Sandhi Splitter

Shubham Bhardwaj

2012EE10480

Supervisors

Dr. Rahul Garg and Dr. Sumeet Agarwal

Sandhi

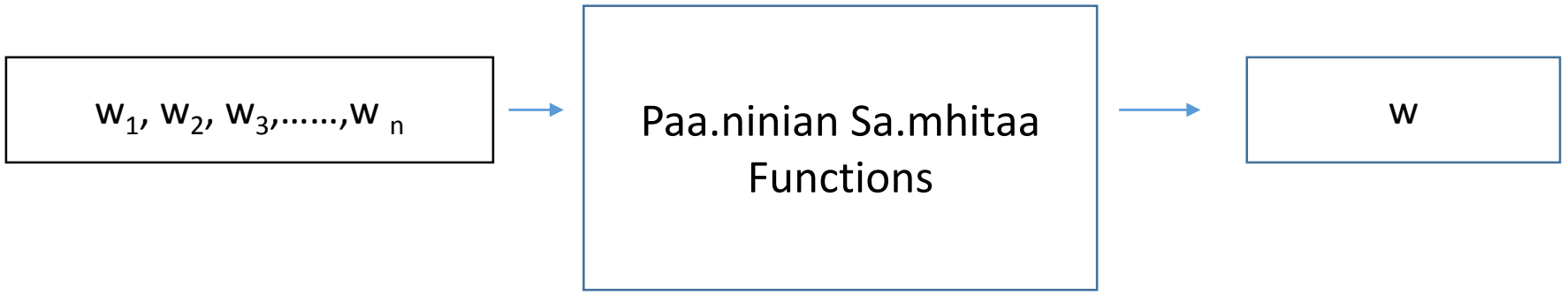
For every i , w_i is a word → External Sandhi

e.g., तस्मै + एतत् → तस्मायेतत्

For some i , w_i is a prefix, verb root or a suffix → Internal Sandhi

e.g. वि + छेद → विच्छेद

What governs this interference?



- A.s.taadhyaayii of Paa.nini
- Sa.mhita governs sutras 73-157 of Chapter 1 of Book 6 and all sutras of Chapter 3 and 4 of Book 8
- Total number of Paa.nian Sandhi Sutras - 271

Existing Sandhi Splitters

- Sanskrit Sandhi Recognizer and Analyzer
  (Dr. G.N. Jha, Special Centre for Sanskrit Studies, JNU)

- Sandhi-Splitter
  (Dr. Amba Kulkarni, Department of Sanskrit Studies, University of Hyderabad)

- The Sanskrit Reader Companion
  (Dr. Gerard Huet ,Computational Linguistics, INRIA, France)

Evaluation of Sandhi Splitters

1. Rule-Based Evaluation

(a) Internal Sandhi Cases

(b) External Sandhi Cases

2. Literature-Based Evaluation

Rule- Based Evaluation

Source of Rules : The A.s.taadhyaayi of Paa.nini Translated into English by Srisa Chandra Basu

Source of Examples :

1. The A.s.taadhyaayi of Paa.nini Translated into English by Srisa Chandra Basu
2. Prau.dh- Rachnaa- Anuvaad Kaumudi by Dr. Kapil Dev Dwivedi
3. Sandhi.h by G. Mahaabaleswar Bhatt

Total Number of Cases :

External Sandhi – 132

Internal Sandhi - 150

Rule-Based Evaluation Results

SANDHI SPLITTER	EXTERNAL SANDHI CASES (132)	INTERNAL SANDHI CASES (150)	OVERALL PERFORMANCE
JNU	21 (15.9 %)	14 (9.3 %)	12.4 %
UoH	48 (36.4 %)	27 (18 %)	26.6 %
INRIA	49 (37.1 %)	6 (4 %)	19.5 %

No. of Cases Not Detected by Any Splitter:

External Sandhi - 62 (46.9 %)

Internal Sandhi - 114 (76 %)

ISSUES

1. Rules not implemented

e.g. कथन्ह्नुते -> कथम् + ह्नुते

2. Strategy of Splitting

दक्षिणायन -> दक्षिण + अयन

उत्तरायण -> उत्तर + अयन

युधिष्ठिरः -> युधि + स्थिरः

3. Limited Corpus

उपाच्छति -> उप + ऋच्छति

Literature-Corpus Based Evaluation

Source of Words : Sandhi Extracted Corpora available at UoH website

Total Number of Cases : 150

SANDHI SPLITTER	CASES DETECTED CORRECTLY	PERFORMANCE
JNU	14	9. 3 %
UoH	96	64 %
INRIA	123	82 %

ISSUE

Compounding creates problems, e.g.

व्याप्यवृत्तितयेदानीमित्यस्यानन्वयाद् → व्याप्यवृत्तितया+इदानीम्+इत्यस्य+अनन्वयाद्

निरवधिरमलप्रीतिरस्माकमास्ताम् → निरवधिरमलप्रीतिः+अस्माकम्+आस्ताम्

Literature Review

1. *Sandhi Splitter and Analyzer for Sanskrit (With Special Reference to aC Sandhi)*, Sachin Kumar , JNU
2. *From Paa.nini Sandhi to Finite State Calculus*, M.D. Hyman, Max Planck Institute for the History of Science, Berlin
3. *Analysis of Sanskrit Text : Parsing and Semantic Relations*, Pawan Goyal and Vipul Arora and Laxmidhar Behera, IIT Kanpur

Approach to Sandhi Splitting

- Brute-Force Approach

1. Scan every letter.
2. Reverse apply each sandhi rule which leads to that letter.
3. Send the splits for validation.

- Inverse Application of 271 Sutras of Paa.nini
- Codification of Paa.nini Sutras (in the form of Sets and Functions)

Paa.nini Sandhi Functions

A general **sandhi function** is given by :

f (Last(w_1) , First (w_2), Left_Context, Right_Context, Overall_Context)

= (Action , Governing_Rule(Last(w_1) , First (w_2)))

where

Last(w_1) = Last letter of w_1

First (w_2) = First letter of w_2

Left-Context – A umbrella category related to information about the structure of w_1 .

This includes Second_Last(w1), whether w1 is a prefix, a word, verb or noun, etc.

Right-Context - A umbrella category related to information about the structure of w_2 .

This includes Second_(w2), whether w2 is a suffix, verb or noun, etc .

Overall-Context – This includes the sense in which a word is used, the other words among which it is used, etc.

Action Functions

Let Last(w_1) and First (w_2) be denoted by a and b respectively.

1. Aagam

   a + b -> acb वि + छेद = विच्छेद

2. Pre-Aagam

   a + b -> ac + b (c = f(a) or f(b)) प्राङ् + शेते = प्राङ्क् + शेते = प्राङ्क् छेते

3. Post- Aagam

   a + b -> a + cb (c = f(a) or f(b)) मधुलिट् + साये = मधुलिट् त्साये

4. Pre-Aadesh

   a + b -> c + b (c = f(a) or f(b)) प्रति + एकः = प्रत्येकः

5. Post-Aadesh

   a + b -> a + c (c = f(a) or f(b)) वाग् + हरिः = वाग्घरिः

Action Functions (Continued)

6. Ekadesh

   a + b -> c (c = f(a) or f(b)) राजा + ऋषिः = राजर्षिः

7. Pre-Elision

   a + b -> b एषः + ददाति = एष ददाति

8. Post- Elision

   a + b -> a कृष्णर् + ध्धिः = कृष्णर्धिः

9. Pre-Reduplication

   a + b -> aa + b प्रत्यङ् + आत्मा = प्रत्यङ्ङात्मा

10. Post-Reduplication

    a + b -> a + bb कर् + तव्यम् = कर्त्तव्यम्

11. Prakritibhaav हरी + एतौ = हरी एतौ

Observations

1. Special letters, strings and symbols

   Letters - ङ , ञ , ं , ँ , ऐ , औ , etc

   Strings - आर् , आव् , आय् ,च्छ , etc

   Special features – **S** (Avagraha), Space

2. Large number of functions dealing with change of n to .n and s to .s

3. Left and Right Context conditions applicable to a large number of functions – Can help reduce the number of splits

Sandhi Splitting Function

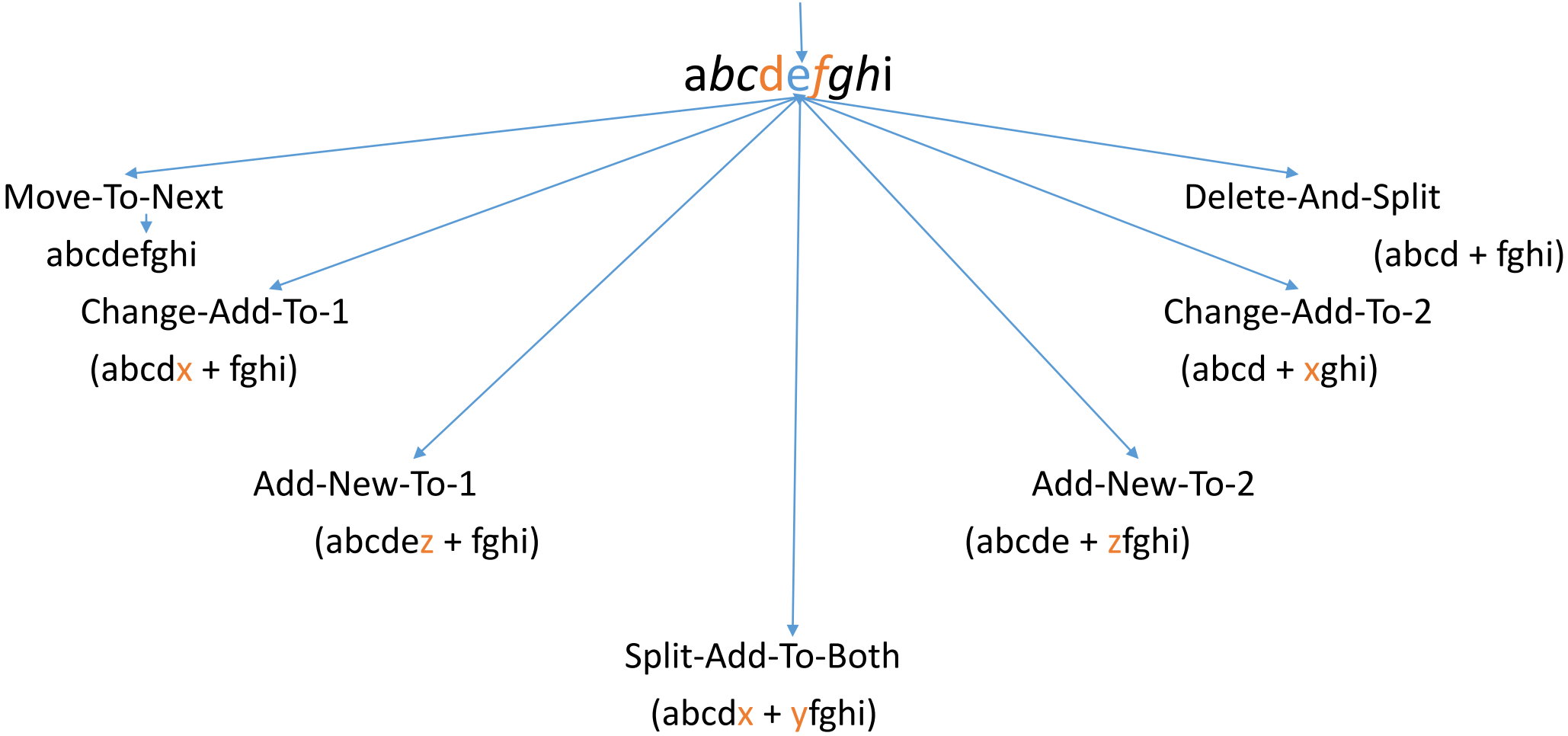
f (Letter, Preceding-Letter, Succeeding-Letter, Preceding-String, Succeeding- String)

= (Action , Inverse-Function)

Actions :

1.	Split-Add-To-Both	राजर्षिः	-> राजा + ऋषिः
2.	Change-Add-To-1	प्रत्येकः	-> प्रति + एकः
3.	Change-Add-To-2	वाग्घरिः	-> वाग् + हरिः
4.	Add-New-To-1	एष ददाति	-> एषः + ददाति
5.	Add-New-To-2	कृष्णर्धिः	-> कृष्णर् + ध्धिः
6.	Delete-And-Split	विच्छेद	-> वि + छेद

Using Sandhi Splitting Function



Proposed Approach for Sandhi Splitting

1. Replacement of .s to s and .n (preceded by r) to n.
2. Checking for the presence of special letters / strings / symbols in the word. If found, call the inverse sandhi function for each of them.
3. Scanning the input for each letter. Call the sandhi splitting function for each letter.
4. Send the splits for validation by the Corpus.

Future Work

1. Sandhi splitting function to be exhaustively laid out for each letter
2. De-compounding tool also to be created and used simultaneously
3. Validation Problem

REFERENCES

1. *The A.s.taadhyaayi of Paa.nini Translated into English* by Srisa Chandra Basu, Indian Press, 1891

2. *Prau.dh- Rachna- Anuvaad Kaumudi* by Dr. Kapil Dev Dwivedi, 2007 Edition , Visvavidyalaaya Prakaashan, Varanasi

3. *Sandhi.h* by G. Mahaabaleswar Bhatt, 2013 Edition , Sanskrit Bharati Prakaashan, Bengaluru

4. *Sandhi Splitter and Analyzer for Sanskrit (With Special Reference to aC Sandhi)*, Sachin Kumar , JNU

5. From Paa.nini Sandhi to Finite State Calculus, M.D. Hyman, Max Planck Institute for the History of Science, Berlin

6. Analysis of Sanskrit Text : Parsing and Semantic Relations, Pawan Goyal and Vipul Arora and Laxmidhar Behera, IIT Kanpur