EED310 Mini-Project

on

Automated Evaluation of Sandhi Splitters

by

Shubham Bhardwaj

(2012EE10480)

Assisted by : Neelamadhav Gantayat

under the guidance of

Dr. Rahul Garg and Dr. Sumeet Agarwal

BTP Part 1 Recap

- Manual evaluation of Sandhi Splitters (Rule-based and Literature-based)
- Codification of Paa.ninian Sa.mhitaa Rules as Sets and Functions
- Suggesting an algorithm of sandhi splitting more efficient than the brute force approach

Sandhi

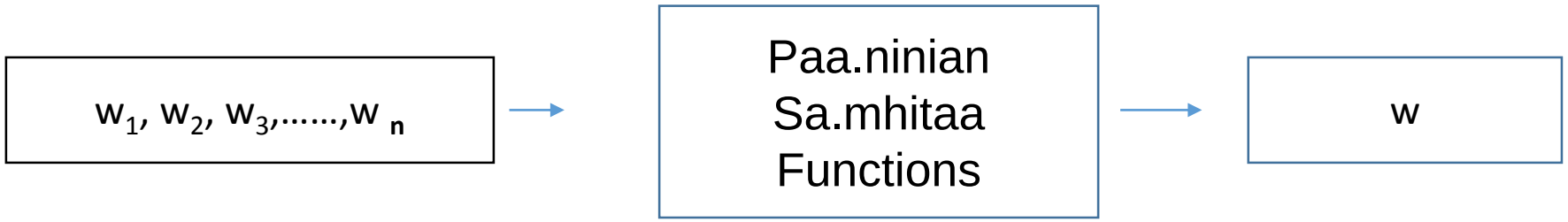
For every i , w_i is a word → External Sandhi

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)

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 %)

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 %

Mini P's Objective

Evaluation of Sandhi Splitters on Large Corpora

- Automation Required

--Developing a tool for all the three splitters

--Getting Reliable Corpora

Developing Automated Evaluator

- Automation -

--Sending requests and getting output (done last semester)

--Extracting results (done only for JNU)

--Evaluation / Matching against the right splits

- All three steps completed this semester for all the three splitters by Neelamadhav Gantayat
- Automatic Evaluation Tool Ready

Getting Reliable Corpora

- 40 Sandhi Split Corpora Available on UoH website
- Lots of Issues with Each Corpus

- No Split

  श्रीमद्भगवद्गीतासु -> श्रीमद्भगवद्गीतासु (श्रीमत् + भगवत् + गीतासु)

- Insufficient Splits

  श्रोत्रादीनीन्द्रियाण्यन्ये -> श्रोत्रादीनि +इन्द्रियाणि +अन्ये

- Wrong Splits वृकोदरः -> वृकः + उदरः (वृक + उदरः)

- Typos श्रेयोनुपश्यामि (श्रेयोऽनुपश्यामि) -> श्रेयः + अनुपश्यामि

Problem of Insufficient Splits: Just an Illustration

Text (first 100 cases)	Insufficient Splits
Vinodini	42 %
Manjusa	44 %
Vyutpattivada	58 %
Dutvakyam	8 %
Short-Stories	9 %

Corpora for Automatic Evaluation

- Rule-based Internal and External Sandhi Corpus (Fool-Proof)
- Creating a Srimad Bhagvad Gita Corpus (Fool-Proof)
- Corpus of A.s.taadhyaayii Sutras (Word-Length Filtering)
- All 40 Corpora available at the UoH website (Dictionary Filtering)

Sutra-based External and Internal Sandhi Corpus

- Created last semester
- Taking at least one example for each of 271 Paa.nini Sa.mhitaa Sutras
- Total number of cases : 282

  External Sandhi -132

  Internal Sandhi - 150

Results of Evaluation of Rule-Based Corpus

Total Number of Cases: 282

	Automated Evaluation	Manual Evaluation
JNU	32 (11.4 %)	35 (12.4 %)
UoH	51 (18.1 %)	75 (26.6 %)
INRIA	41 (14.5 %)	55 (19.5 %)

Lenient Evaluation and Intermediate Results

नयनम् ->	ने + अनम्	But	ने + अनम ?
प्रौढः ->	प्र + ऊढः	But	प्र + ऊढ
वृक्षश्शेते ->	वृक्षस्+शेते	But	वृक्षः + शेते

Srimad Bhagvad Gita Corpus

- Manually being created
- Why not use the Existing Corpus?
- Cases up to Ch. 2 - 431

Type of Mistake	No.
Typos	41
Insufficient Splits	92
Wrong Split	10

Results of Automatic Evaluation of Bhagvad Gita Corpus

	A_1 [157]	A_2 [270]	A_3[168]	A_4 [164]	A_5 [113]	Total [872]
JNU	2 (1.3)	10 (3.7)	11(6.5)	4(2.4)	9 (7.9)	36 (4.1)
UoH	61 (38.8)	115(42.6)	74(44)	78(47.6)	48(42.5)	376(43.1)
INRIA	95 (60.5)	160 (59.3)	93(55.4)	88(53.7)	60(53.1)	496 (56.9)

Filtering for Creating Highly Reliable Corpora

1. **UoH Corpus**

Restricting to those cases where the RHS occurs in the dictionary

तुमुलो व्यनुनादयन् -> तुमुलः+वि+ अनुनादयन्

(Cannot be located in Dictionary, so Rejected)

सर्वान्बन्धूनवस्थितान् -> सर्वान् + बन्धून् + अवस्थितान्

(Cannot be located in Dictionary, so Rejected)

शब्द इव -> शब्दः+ इव	(Included)
नार्हति -> न + अर्हति	(Included)
अस्तमितो भगवान् -> अस्तम् + इतः + भगवान्	(Included)

Results of Automatic Evaluation of All Available Corpora

Total Number of Cases :18326

Splitter	No. of Cases Correctly Identified
JNU	3215 (17.5 %)
UoH	11405 (62.2%)
INRIA	13416 (73.2 %)

Filtering for Creating Highly Reliable Corpora

2. A.s.taadhyaayii Corpus

प्रथमचरमतयाल्पार्धकतिपयनेमाश्च -> प्रथमचरमतयाल्पार्धकतिपयनेमाः+च

(Insufficient Split)

उदुपधाद्भावादिकर्मणोरन्यतरस्याम् -> उदुपधात्+भावादिकर्मणोः+अन्यतरस्याम्

(Insufficient Split)

Selecting words with fewer letters to decrease the probability of insufficient splits

तद्धितश्चासर्वविभक्तिः -> तद्धितः+च+च+असर्वविभक्तिः (Full Split)

वृद्धिरादैच् -> वृद्धिः+आदैच् (Full Split)

विज इट् -> विजः+इट् (Full Split)

Results on A.s.taadhyaayii

Total Number of Sutras – 3,959

Sutras where Sandhi Split Applicable – 2,700

No. of letters (<=)	**10**	**20**	**30**	**40**	**50**	**All**
	(93)	(571)	(1512)	(2045)	(2302)	(2700)
JNU	4 (4.3)	10 (1.75)	17 (1.12)	18 (0.88)	18 (0.78)	18 (0.66)
UoH	21 (22.6)	100 (17.5)	226 (14.9)	263 (12.86)	263 (11.42)	263 (9.74)
INRIA	29 (31.2)	195 (34.15)	378 (25)	444 (21.7)	460 (19.9)	507 (18.7)

Analysis of Errors (For Srimad Bhagvad Gita Case)

1. Rules not Implemented

स कौन्तेयः -> सः + कौन्तेयः

2. Optional Results Not Included

वर्तन्त इति -> वर्तन्ते + इति (Main Result : वर्तन्तय् इति -> वर्तन्तयिति)

3. Strategy of Splitting Not Correct

परायणाः -> पर + अयनाः

प्रतिष्ठित -> प्रति + स्थित

युधिष्ठिरः -> युधि स्थिरः

4. Corpus Not Sufficient

अच्छेद्यः -> अ + छेद्यः

उपार्च्छति -> उप + ऋच्छति

Solutions and Future Work

- Implementing all rules, including optional cases
- All the 271 Paa.nini Sutras codified last semester – **Sandhi Tool Implementation**
- Revising the strategy of Splitting
- Using the Algorithm for Sandhi Splitting designed last semester – **Sandhi Splitter Implementation**
- Prefix and Suffix Analysers - **Implemention of Other Parts of A.s.taadhyaayii**
- **Publishing the results**

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