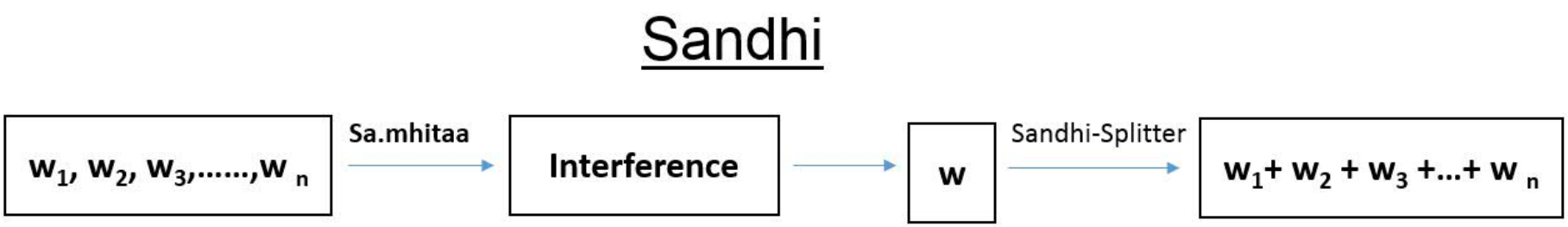
Sanskrit sandhi splitting and merging tools

Shubham Bhardwaj
Neelamadhav Gantayat
(Mini Project)
under the guidance of
Prof. Rahul Garg
Prof. Sumeet Agarwal

February 24, 2016

Outline

Sandhi



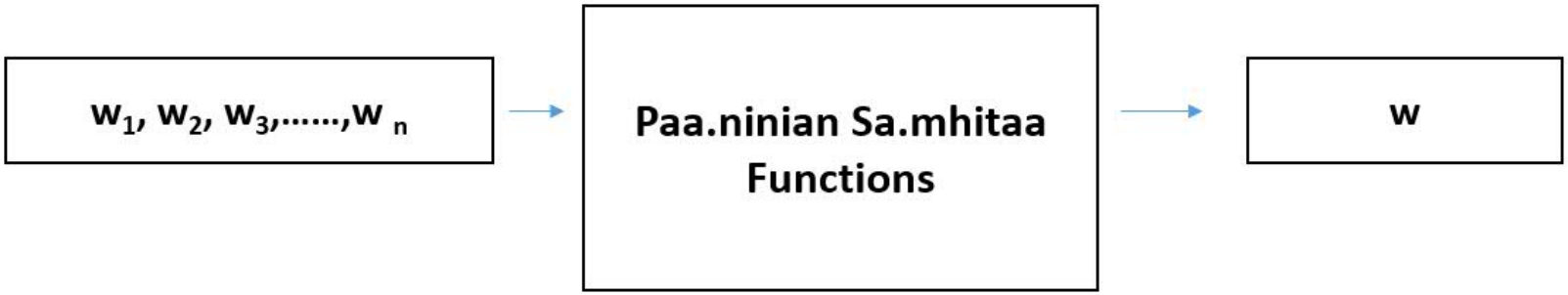
For every i , w_i is a word ⟶ External Sandhi

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.ninian Sandhi Sutras - 271
- Paa.ninian Sutras codified in the form of Sets and Functions

Automatic Evaluation of Sandhi Splitters and Generators

- Sanskrit Sandhi Analyzer and Generator[1]
  (Dr. G.N. Jha, Special Centre for Sanskrit Studies, JNU)
- Sandhi-Splitter and Generator[2]
  (Dr. Amba Kulkarni, Department of Sanskrit Studies, University of Hyderabad)
- The Sanskrit Reader Companion and the Sandhi Engine[3]
  (Dr. Gerard Huet ,Computational Linguistics, INRIA, France)

[1] http://sanskrit.jnu.ac.in/sandhi/viccheda.jsp?itext=
[2] http://52.25.246.194/scl/
[3] http://sanskrit.inria.fr/DICO/reader.fr.html

Sandhi Corpus

- 40 different sandhi-split corpora available on the University of Hyderabad website[4]
- Problems - Insufficient splits, no splits, wrong splits, typos

Text (first 100 cases)	**Errors**
Vinodini	42%
Manjusa	44%
Vyutpattivada	58%
Dutvakyam	8%
Short-stories	9%

- Creating a Srimad Bhagvad Gita Corpus

[4] `http://sanskrit.uohyd.ac.in/Corpus/`

Need for a transliteration tool

- Various data sources of Sanskrit literature were encoded in different scripts.
- Different tools use different transliteration schemes for input and output
- From our survey identified some web tools and some standalone tools - No bulk conversion option
- Need a tool which can convert from one script to another in bulk (Files and folders)

Sanskrit Transliteration

Devanāgarī	IAST	Harvard-Kyoto	ITRANS	Velthuis	SLP1
अ	a	a	a	a	a
आ	ā	A	A/aa	aa	A
इ	i	i	i	i	i
ई	ī	I	I/ii	ii	I
उ	u	u	u	u	u
ऊ	ū	U	U/uu	uu	U
ए	e	e	e	e	e
ऐ	ai	ai	ai	ai	E
ओ	o	o	o	o	o
औ	au	au	au	au	O
ऋ	ṛ	R	RRi/R^i	.r	f
ॠ	ṝ	RR	RRI/R^I	.rr	F
ऌ	ḷ	lR	LLi/L^i	.l	x
ॡ	ḹ	lRR	LLI/L^I	.ll	X
अं	ṃ	M	M/.n/.m	.m	M
अः	ḥ	H	H	.h	H
अँ			.N		~

Figure: Example: One step conversation[5]

- Among these SLP1 script has complete encoding scheme for Devanagari

[5] https://en.wikipedia.org/wiki/Devanagari_transliteration

Sandhi Corpus Format

- Manually curated corpus from University of Hyderabad [6]

Formats

- Single words split into multiple words

चार्थे => च+अर्थे

- Multiple words split into multiple words

सङ्केतो लक्षणा => सङ्केतः+ लक्षणा

- Multiple words span across multiple lines

व्याख्यां विज्ञहितां => व्याख्याम्+ विज्ञहिताम्+
विज्ञहितां कुर्वे => विज्ञहिताम्+ कुर्वे

[6] http://sanskrit.uohyd.ac.in/Corpus/

Processing corpus

- Filtered out single word, multi split instances
- Combined multi lines into single line and converted them to multi word instance.
- One way of ensuring correctness is checking for the split words in the dictionary

Dictionary Creation

Various types of dictionary, example: [7]

- Sanskrit to Sanskrit dictionary
- Sanskrit to English dictionary
- English to Sanskrit dictionary

Dictionary words creation

- Parsed "Sanskrit to Sanskrit" and "Sanskrit to English"
- Created a comprehensive list of words from all the dictionary.

[7] http://www.sanskrit-lexicon.uni-koeln.de/

Filtering Corpus

- Identified all the sandhi splits from the corpus where all the split words are available in the dictionary
- out of 97674 samples we identified 15325 samples where all the splits are present in the dictionary

Evaluation

- Sanskrit Sandhi Analyzer and Generator - 1523/7217 (21.10%)
- Sandhi-Splitter and Generator - 4248/7217 (58.86%)
- The Sanskrit Reader Companion and the Sandhi Engine - Not yet done

Future Work

- Validate dictionary output
- Check for Sandhi splitting and Sandhi Merging tools apart from the above three
- Create corpus using Sandhi Merging tool by merging words or dictionary.
- Manually create Sandhi corpus from "Sri Mad Bhagwat Gita"
- Create our own Merging and splitting tool if required.