Samāsa-Kartā:
An Online Tool for Producing Compound Words using IndoWordNet

Hanumant Redkar, Nilesh Joshi,
Sandhya Singh, Irawati Kulkarni,
Malhar Kulkarni and Pushpak Bhattacharyya

Center for Indian Language Technology,
Indian Institute of Technology Bombay, India.

8th International Global WordNet Conference (GWC 2016)
Bucharest, 27-30 January 2016

Outline

- What is *Samāsa*?
- Types of *Samāsa* in Sanskrit
- IndoWordNet as a Resource
- *Samāsa-Kartā* – The *Samāsa* Producer
- Components of *Samāsa-Kartā*
- Salient Features of *Samāsa-Kartā*
- Future Enhancements to *Samāsa-Kartā*
- Summary

What is *Samāsa?*

- *Samāsa* or a compound word is constructed from two or more words to form a single word.

- The meaning of this word is derived from each of the individual words of the compound.

Samāsa in Sanskrit

- A *samāsa,* (समास) is defined as पृथगर्थानामेकार्थीभावः समासः
  (pṛthagarthā nāmekārthībhāvaḥ samāsaḥ)
  i.e., placing together two or more words so as to express a composite sense, which is a compound composition.

- Example,

  शिवपत्नी *(śivapatnī)* = शिव *(śiva)* + पत्नी *(patnī)*

  ➔ (wife of *śiva* and a benevolent aspect of *devī*)
  =
  (a major divinity in the later Hindu pantheon)
  +
  (a married woman)

Types of *Samāsa* in Sanskrit

- **अव्ययीभाव** (***Avyayībhāva***)
  - पूर्वपदार्थप्रधान, *pūrva-padārtha-pradhāna.*
  - Here, first member has primacy.
  - E.g., यथाशक्ति (*yathāśakti*, in accordance with one's strength).

- **तत्पुरुष** (***Tatpuruṣa***)
  - उत्तरपदार्थप्रधान, *uttara-padārtha-pradhāna.*
  - Here, second member has primacy and the first component is in a case relationship with another.
  - E.g. सन्ध्याकालः (*sandhyākālaḥ*, evening time).

Types of *Samāsa* in Sanskrit

- **द्वन्द्व** (***Dvandva***)
  - उभयपदार्थप्रधान, *ubhaya-padārtha-pradhāna.*
  - Here, both members have primacy .
  - E.g., रामलक्ष्मणभरतशत्रुघ्नाः (*rāmalakṣmaṇabharata śatrughnāḥ*, Ram and Laxman and Bharat and Shatrughn).

- **बहुव्रीहि** (***Bahuvrīhi***)
  - अन्यपदार्थप्रधान, *anya-padārtha-pradhāna*
  - Here, both members refers to a thing which in itself is not part of the compound.
  - E.g., गजाननः (*gajānanaḥ*, one whose face is that of an elephant).

IndoWordNet as a Resource

- IndoWordNet is a linked structure of 18 Integrated WordNets of major Indian languages.
- In this paper, we have taken **Sanskrit WordNet** as a resource.
- Sanskrit is an Indo-Aryan language and is one of the ancient languages.
- The roots of most of the languages in the Indo European family in India can be traced to Sanskrit.

Samāsa-Kartā

- The *Samāsa-Kartā*, also known as Compound Word Producer is an online tool developed to produce compound words.

- The produced words are formed using a semi-automatic rule based system which takes two words from IndoWordNet database as an input.

- The new word which is produced, is another word, which falls under any of the four types of *samāsas* mentioned earlier.

Samāsa-Kartā

- *Samāsa-Kartā* produces compounds between -
  - Noun-Noun (NN-NN)
  - Noun-Adjective (NN-JJ)
  - Adjective-Adjective (JJ-JJ)
  - Adjective-Noun (JJ-NN)

- *Samāsa-Kartā* does not deal with the following combinations -
  - Noun-Adverb (NN-RB)
  - Adverb-Noun (RB-NN)
  - Verb-Verb (VM-VM)
  - Verb-Noun (VM-NN)
  - Noun-Verb (NN-VM)

Interface of *Samāsa-Kartā*

Samāsa-Kartā

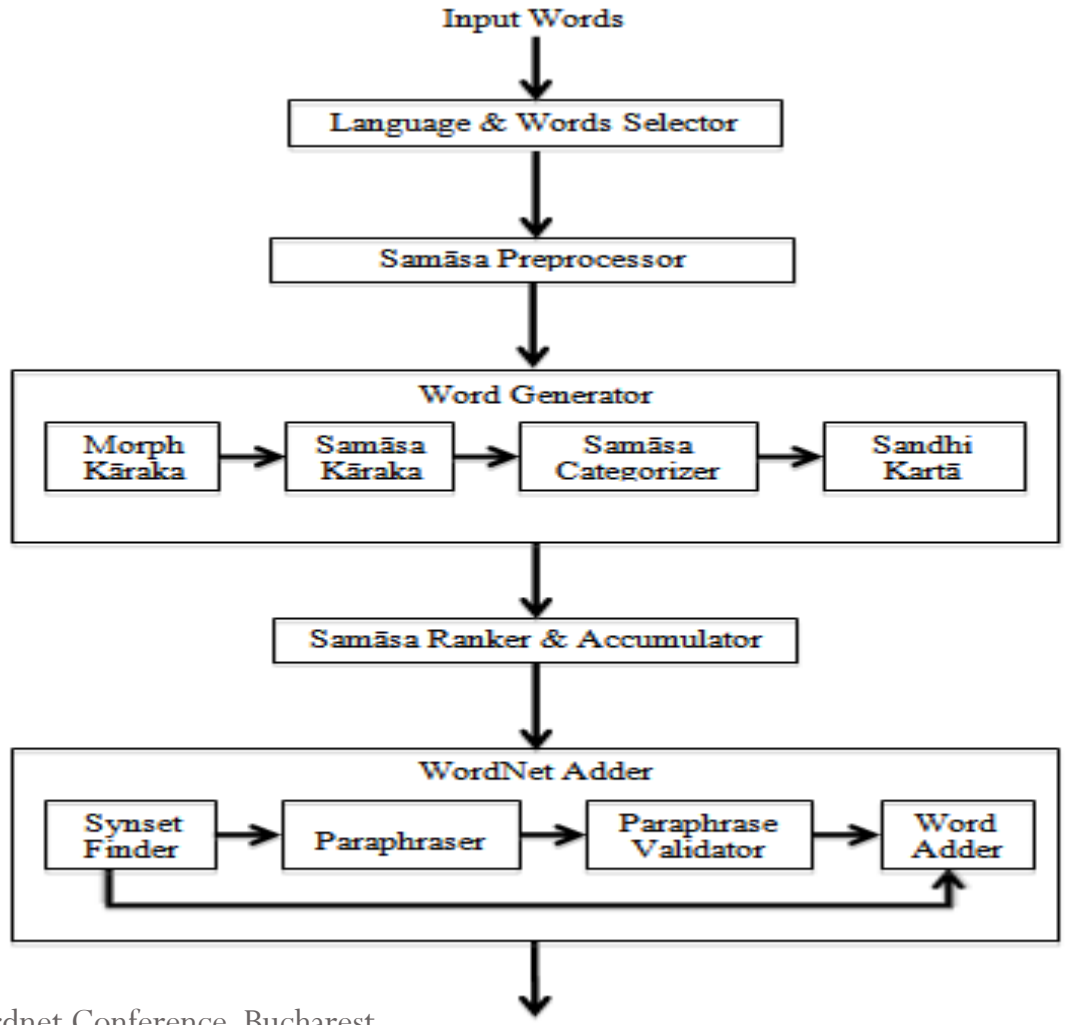
The Compound Word Producer

Language: Sanskrit

Word1: मन्दः Gender: Male	Word2: मतिः Gender: Female
Sense 1 ☑ Synonyms: **मन्दः**, तुन्दपरिमृजः, आलस्यः, शीतकः, अनुष्णः, शीतलः, कुण्ठः, अनाशुः Gloss: अवश्यकर्तव्येषु अप्रवृत्तिशीलः। Example(s): "मन्दः किमपि न प्राप्नोति।"	Sense 1 ☐ Synonyms: मतम्, दृष्टिः, **मतिः**, धीः Gloss: किमपि वस्तु कमपि विषयं वा अधिकृत्य कृतं चिन्तनम्। Example(s): "अस्माकं मतेन भवताम् इदं कार्यं न समीचिनम्।"
	Sense 2 ☑ Synonyms: **मतिः**, बुद्धिः, धी, प्राज्ञता Gloss: निश्चयात्मिकान्तःकरणवृत्तिः यस्याः बलेन चिन्तयितुं शक्यते। Example(s): "धनलाभार्थे अन्यस्य मत्या जीवनाद् भिक्षाटनं वरम्।"
	Sense 3 ☐ Synonyms: मतम्, अभिप्रायः, सम्मतिः, दृष्टिः, बुद्धिः, पक्षः, भावः, मनः, धी, **मतिः**, आकुतम्, आशयः, छन्दः Gloss: केषुचित् विषयादिषु प्रकटीकृतः स्वविचारः। Example(s): "सर्वेषां मतेन इदं कार्यं सम्यक् प्रचलति।"

Generate Words

Components of *Samāsa-Kartā*



Components of *Samāsa-Kartā*

- Language and Words Selector Module
  - The user selects input language and words taken from IndoWordNet database.
  - The system displays their corresponding synset.
  - The user chooses words to be compounded from both the selected synsets.
  - The user clicks on **Generate Words** *button*, which initiates the *Samāsa* Preprocessor.

Components of *Samāsa-Kartā*

- *Samāsa* Preprocessor Module
  - Performs a check whether the input words are valid to form a *samāsa* or not.
  - It checks for part-of-speech of each input word and validates if the combinations like NN-NN, NN-JJ, JJ-NN, JJ-JJ. can be formed as compound.

Components of *Samāsa-Kartā*

- Word Generator Module

  - has four sub modules
    - *Morph-Kāraka*
    - *Samāsa-Kāraka*
    - *Samāsa* Categorizer
    - *Sandhi-Kartā*

Components of *Samāsa-Kartā*

- Morph-*Kāraka* Sub Module
  - Converts each input word in its root form using standard morphological rules.

Components of *Samāsa-Kartā*

- *Morph-Kāraka* conversion example

स्वरान्त-शब्दाः (*svarānta-śabdāḥ*) (vowel-ending words)		**व्यञ्जनान्त-शब्दाः** (*vyañjanānta-śabdāḥ*) (consonant-ending words)	
अकारान्त (*akārānta*)	:दन्म→ दन्म (*mandaḥ → manda*)	**चकारान्त** (*cakārānta*)	वाक् → वाच् (*vāk → vāc*)
आकारान्त (*ākārānta*)	विद्या → विद्या (*vidyā → vidyā*)	**जकारान्त** (*jakārānta*)	भिषक् → भिषज् (*bhiṣak → bhiṣaj*)
इकारान्त (*ikārānta*)	मतिः → मति (*matiḥ → mati*)	**तकारान्त** (*takārānta*)	भगवान् → भगवत् (*bhagavān → bhagavat*)
ईकारान्त (*īkārānta*)	नदी → नदी (*nadī → nadī*)	**दकारान्त** (*dakārānta*)	शरद् → शरद् (*śarad → śarad*)
उकारान्त (*ukārānta*)	भानुः → भानु (*bhānuḥ → bhānu*)	**नकारान्त** (*nakārānta*)	आत्मा → आत्मन् (*ātmā → ātman*)
ऋकारान्त (*ṛkārānta*)	माता → मातृ (*mātā → mātṛ*)	**सकारान्त** (*sakārānta*)	तेजः → तेजस् (*tejaḥ → tejas*)

Components of *Samāsa-Kartā*

- *Samāsa-Kāraka*
  - Here, the standard *samāsa* rules are applied on root words received from *Morph-Kāraka.*
  - This is done to check if words are eligible to form a *samāsa* or not.

आत्मन् (ātman) + शक्ति (śakti)	षष्ठी (ṣaṣṭhī) 2.2.8. shows that these words are eligible to form Samāsa of the type ṣaṣṭhī *tatpuruṣa.*
आत्म (ātma) + शक्ति (śakti)	नलोपः प्रातिपदिकान्तस्य (*nalopaḥ prātipadikān tasya*) 8.2.7. shows that न् (*n*) should be removed from the word आत्मन् (*ātman*)

  - So, words आत्म (*ātma*) and शक्ति (*śakti*) is sent to *Samāsa* Categorizer for further processing.

Components of *Samāsa-Kartā*

- *Samāsa* Categorizer

  - Identifies category and sub category of a *samāsa* like *Avyayībhāva*, *Tatpuruṣa*, *Dvandva* & *Bahuvrīhi* as per the *samāsa* rules.

  - Generates paraphrased information using gloss of input words which will be used in the WordNet Adder for paraphrasing of compound words.

Components of *Samāsa-Kartā*

- *Sandhi-Kartā*
  - It joins two words together which are passed through *Samāsa* Categorizer following the sandhi rules of the language into consideration.
  - A list of all possible combinations of the selected synset words are produced.

Components of *Samāsa-Kartā*

- *Sandhi-Kartā* conversion example

स्वरान्त-शब्दाः (*svarānta-śabdāḥ*) (vowel-ending words)	
अकारान्त (*akārānta*) (words ending with a)	देव + ईश → देवेश (*deva + īśa → deveśa*)
आकारान्त (*ākārānta*) (words ending with ā)	विद्या + आलय → विद्यालय (*vidyā + ālaya → vidyālaya*)
इकारान्त (*ikārānta*) (words ending with i)	प्रति + उत्तर → प्रत्युत्तर (*prati + uttara → pratyuttara*)
ईकारान्त (*īkārānta*) (words ending with ī)	नदी + ईश → नदीश (*nadī + īśa → nadīśa*)
उकारान्त (*ukārānta*) (words ending with u)	भानु + उदय → भानूदय (*bhānu + udaya → bhānūdaya*)
ऋकारान्त (*ṛkārānta*) (words ending with ṛ)	मातृ + ऋण → मातॄण (*mātṛ + ṛṇa → mātṝṇa*)

व्यञ्जनान्त-शब्दाः (*vyañjanānta-śabdāḥ*) (consonant-ending words)	
तकारान्त (*takārānta*) (words ending with ta)	भगवत् + गीता → भगवद्गीता (*bhagavat + gītā → bhagavadgītā*)
दकारान्त (*dakārānta*) (words ending with da)	शरद् + हविष् → शरद्धविष् (*śarad + haviṣ → śaraddhaviṣ*)
नकारान्त (*nakārānta*) (words ending with na)	आत्मन् + शक्ति → आत्मशक्ति (*ātman + śakti → ātmaśakti*)
सकारान्त (*sakārānta*) (words ending with sa)	मनस् + रथ → मनोरथ (*manas + ratha → manoratha*)

Components of *Samāsa-Kartā*

- *Samāsa* Ranker and Accumulator
  - Words from Word Generator module are ranked and accumulated together as per the most frequent usage of words in the original WordNet synsets.
  - *Samāsa* Ranker Algorithm is used to rank the accumulated *samāsas.*

Components of *Samāsa-Kartā*

- WordNet Adder Module
  - Synset Finder
  - Paraphraser
  - Paraphrase Validator
  - Word Adder

Components of *Samāsa-Kartā*

- Synset Finder
  - The lexicographer checks if the intended synset already exists in the WordNet.
  - If yes, then the words are directly appended to the intended synset's vocabulary.
  - If not, then it passes through the ***Paraphraser to create gloss of the compound word*** which will help in creating new synset**.**

Components of *Samāsa-Kartā*

- Paraphraser
  - Automatically generates gloss of the intended synset on the basis of input words.
  - This gloss or a concept definition of a synset is given to Paraphrase Validator for further processing.

Components of *Samāsa-Kartā*

- Paraphrase Validator
  - The lexicographer checks if the paraphrased gloss is properly generated.
  - If not, it is created / edited manually by using the three principles of synset creation, *viz.,* principle of minimality, coverage and replaceability.

Components of *Samāsa-Kartā*

- Word Adder
  - The lexicographer fills-in other synset information like examples, gender, *etc.* and adds to the WordNet database
  - The resultant *Samāsas* will either be the member of an existing synset or it can be a new synset altogether.

Samāsa-Kartā Execution

- Modules like *Samāsa* Preprocessor, Word Generator and Synset Ranker & Accumulator are executed automatically.

- Modules like Word and Language Selector, Synset Finder, Paraphrase validator and Word Adder are handled by users.

Samāsa-Kartā Users

- Lexicographer
  - The main task of lexicographer is to enter words, select synsets, generate compound words and add to a temporary database.

- Validator
  - The main task of validator is to validate the compound words, check if their paraphrase is properly produced, validate it, and update the resultant synset to the IndoWordNet database.

Salient Features of *Samāsa-Kartā*

- *Samāsa* or compounds are created on the flow.
- *Samāsa* in WordNet helps in identifying meaning or concept of a compound occurring in the literature.
- *Samāsa-Kartā* helps in enriching the standard of the language and to simplify the case-ending words in language under consideration.
- It assists in developing vocabulary, which in turn, helps in improving the word count in a language.
- It helps in automatic generation of paraphrases.
- It helps in compound type identification.
- The compound words produced can be helpful to understand the multi-words.

Limitation of *Samāsa-Kartā*

- Used only for words in WordNet.
- Possibility of over generation of compounds.
- In Sanskrit, verbs are in its root form; hence word pairs such as VM-VM and RB-VM are not implemented.
- The word combination NN-RB is not possible as adverbs cannot come as a second word in the compound.

Conclusion

- *Samāsa* is a significant part of most of the languages which is used to express meaning using less number of words.

- The tool *Samāsa-Kartā*, discussed in this paper, is an attempt to improve upon the richness and coverage of a language using a semi-automated approach.

- *Samāsa-Kartā* uses rule based system to form the compounds by passing through various rules of grammar at each sub module and create new compound words along with its paraphrase which can be added to the WordNet.

Future Scope and Enhancements

- The tool can be extended to **other Indian languages** belonging to Indo-Aryan, Dravidian and Sino-Tibetan families *viz*., Hindi, Marathi, Gujarati, Bengali, Konkani, Kannada, *etc*.

- It can also be extended to **other non-Indian languages** like English, German, Italian, *etc.*

- It can be extended to non-WordNet words. This will be useful in the light of development of improving the vocabulary of the language, thus enhancing the richness of the language.

- Some of the major modules of this tool such as *Morph Kāraka*, *Sandhi Kartā*, *Samāsa* Categorizer, Paraphraser, etc. can be made available independently.

References

- Malhar Kulkarni, Irawati Kulkarni, Chaitali Dangarikar and Pushpak Bhattacharyya. 2010(b). *Gloss in Sanskrit Wordnet*. In Proceedings of Sanskrit Computational Linguistics. Jha. G. Berlin: Springer-Verlag / Heidelberg. pp 190-197.

- Neha R. Prabhugaonkar, Apurva S. Nagvenkar, and Ramdas N. Karmali. 2012. *IndoWordNet Application Programming Interfaces.* In 24th International Conference on Computational Linguistics (COLING 2012), p. 237.

- Pavankumar Satuluri, Amba Kulkarni, *Generation of Sanskrit Compounds*, Proceedings of ICON, 2013.

- Priyanka Gupta, Vishal Goyal. 2009. *Implementation of Rule Based Algorithm for Sandhi-Vicheda of Compound Hindi Words*. JCSI International Journal of Computer Science Issues, Vol. 3, 2009.

- Pushpak Bhattacharyya. 2010. *IndoWordNet*. In the Proceedings of Lexical Resources Engineering Conference (LREC), Malta.

- Ramashankar Mishra. 2010. *अष्टाध्यायीसूत्रपाठः*. Motilal Banarasidas publishers pvt. ltd, New Delhi (ISBN 978-81-208-2748-6).

- Venkatesh Prabhu, Shilpa Desai, Hanumant Redkar, Neha Prabhugaonkar, Apurva Nagvenkar, Ramdas, Karmali. 2012. *An Efficient Database Design for IndoWordNet Development Using Hybrid Approach*. COLING 2012, Mumbai, India. p 229.

References

- Amba Kulkarni, Soma Paul, Malhar Kulkarni, Anil Kumar , Nitesh Surtani : Semantic Processing of Compounds in Indian Languages, Proceedings of COLING 2012, Mumbai, December 2012.

- Anil Kumar, Vipul Mittal and Amba Kulkarni: *Sanskrit Compound Processor*, Sanskrit Computational Linguistics - 4th International Symposium, New Delhi, India, 2010.

- George Miller, R., Fellbaum, C., Gross, D., Miller, K. J. 1990. *Introduction to wordnet: An on-line lexical database*. International journal of lexicography, OUP. (pp. 3.4: 235-244).

- Girish Nath Jha, Muktanand Agrawal, Subash, Sudhir K. Mishra, Diwakar Mani, Diwakar Mishra, Manji Bhadra, Surjit K. Singh, *Inflectional Morphology Analyzer for Sanskrit*, Sanskrit computational linguistics. Springer Berlin Heidelberg, 2009.

- Hanumant Redkar, Jai Paranjape, Nilesh Joshi, Irawati Kulkarni, Malhar Kulkarni, and Pushpak Bhattacharyya. 2014. *Introduction to Synskarta: An Online Interface for Synset Creation with Special Reference to Sanskrit*. ICON 2014, Goa, India.

- Malhar Kulkarni, Chaitali Dangarikar, Irawati Kulkarni, Abhishek Nanda and Pushpak Bhattacharyya. 2010(a). *Introducing Sanskrit Wordnet*. In Principles, Construction and Application of Multilingual WordNets, Proceedings of the 5th GWC, edited by Pushpak Bhattacharyya, Christiane Fellbaum and Piek Vossen, Narosa Publishing House, New Delhi, 2010, pp 257 – 294.

Thank You !!!