International Journal of Allied Research in Engineering and Technology (IJARET)
Volume.14, Number 1; January-2023;
ISSN: 2836-5631| Impact Factor: 6.34
https://zapjournals.com/Journals/index.php/ijaret
Published By: Zendo Academic Publishing

COMPUTATIONAL TOOL FOR SANSKRIT SAMASA ANALYSIS: A STEP TOWARDS PRESERVING AND NOURISHING SANSKRIT TEXTS

[1]Biswanath Mishra

Article Info

Keywords: Sanskrit, Samasa Analysis, Computational Tool, Preservation, Nourishment, POS Analysis, Indian Intellectual Tradition, Automatic Translation

Abstract

Sanskrit language has played a crucial role in communicating the unbroken knowledge tradition of the Indian intellectual tradition from the Vedic times to the present. The language is a treasure trove of information, but it is being threatened by the lack of a systematic approach for preserving and nourishing the texts. The sheer volume and variety of data require a good amount of computerization, and a networked approach across centers is necessary for data exchange in real-time. In this regard, the development of a computational tool for Sanskrit Samasa Analysis is essential. Sanskrit language has been analyzed by Shastric Scholars without the stage called "Part of Speech" (POS). The whole process involves many steps, but it does not stress upon POS analysis. However, the development of a computational tool for Samasa Analysis is important as it helps in automatic translation from Sanskrit to Indian languages, which is highly desirable. The computational tool will be a semi-automatic system and will use a human being's knowledge about the world to take a decision about which analysis is correct. The tool will help scholars in word-split, Markup for Sandhi, and Samasa analysis. This paper presents the methodology for developing a computational tool for Sanskrit Samasa Analysis. The paper highlights the need for the tool and the methodology used for its development. The paper discusses how the tool will use a semi-automatic system towards the end, and not all scholars creating content need to be engaged in this. The paper also discusses how the tool will be online and can be accessed at any time.

INTRODUCTION

Sanskrit language has been the carrier of the Indian intellectual tradition, and its texts are essential for understanding the country's cultural heritage. However, the language is being threatened by the lack of a systematic approach for preserving and nourishing the texts. The nature of data varies from scripts (for texts) to

[1] Computer Teacher, Shree Sadashiva Campus Puri

voice data (stored on tape), and there is a need for a networked approach for data exchange in real-time. The texts are related to shastras, and as such, require great skill/expertise for keying in to avoid errors.

The lack of a systematic approach for preserving and nourishing Sanskrit texts has been a major concern for Sanskrit Scholars. The texts require word-split, markup for Sandhi, and Samasa analysis, which is a time-consuming process. This has led to the development of a computational tool for Sanskrit Samasa Analysis. The computational tool will be a semi-automatic system and will use a human being's knowledge about the world to take a decision about which analysis is correct. The tool will help scholars in preserving and nourishing Sanskrit texts.

Sanskrit language has been analyzed by Shastric Scholars without the stage called "Part of Speech" (POS). The whole process involves many steps, but it does not stress upon POS analysis. The computational tool for Samasa Analysis will help in automatic translation from Sanskrit to Indian languages, which is highly desirable. The tool will be online and can be accessed at any time.

The grammar of the Sanskrit language has a complex verbal system, rich nominal declension, and extensive use of compound nouns. It was studied and codified by Sanskrit grammarians from the later Vedic period (roughly 8th century BCE), culminating in the Pāṇinian grammar of the 6th century BCE. Word is considered as the most basic unit of the linguistic structure. Word is a sequence of characters delimited by space. Word consists of a complex set of more primitive parts. The study of morphology is concerned with the construction of words from more basic meaningful units called morphemes. The process of analyzing the given word to extract the information encoded in the word is called as morphological analysis. Morphological analysis deals with the segregation of word into morphemes. By identifying the morphemes of a given word, the form (syntax) and meaning (semantics) of the word can be understood. Morphological Analyzer (MA) is a tool which analyzes the given word into its root, affixes and feature values of the grammar like number, gender, person etc. If the word is not found in the dictionary, then it is assumed to be a complex form which can be further broken down into its derivational constituents that is known as Samasa. The scope of present research is to develop a ***Samasa Analyzer*** for sanskrit based on Paninian formulations. While some attempts have been made to develop string segmentation engine based on adhoc processing. There is no Samasa nalyzer which comprehensively analyses a Sanskrit Text according to Paninan Approach . The present work and the associated algorithm will be helpful in solving this long over due problem in Sanskrit Natual Language Processing (NLP) . Samasa analyser is a critical module for any natural language system for sanskrit.

It is because of the synthetic nature of Sanskrit in which words can be combined together to form a larger string of words. So, before processing Sanskrit input text and extracting morphological and syntactical information from it, these conjugated words need to be segmented into their constituents. An automated *sandhi* analysis is a pre-requisite for complete analysis of Sanskrit input text as it will simplify the Sanskrit text and this simplified text can be basis for doing Part of Speech (POS) analysis and doing further grammatical analysis of the text. This complete analysis of Sanskrit text can be used in various NLP applications like Sanskrit- Indian Language Machine Translation System (MTS), tagging of large text corpora, spell checker for Sanskrit, building a Sanskrit text search engine etc. This work, besides being an essential component in NL system of Sanskrit, will also be useful for self-reading and understanding of Sanskrit text.

SIGNIFICANT AND OBJECTIVES

All the Sanskrit Universities and Vidypaeetha-s have large quantities of data both in terms of volume and type. The nature of data varies from scripts (for texts) to voice data (stored on tape) etc. These data have lot of volue to entire mankind and need to be preserved and nourished. The very nature of this type of system (viz., data exchange in real-time and very continuous) calls for good amount of computerisation and that too networked across centres. The texts are related to *shastras* and as such, require great skill/expertise for keying in to avoid errors. Word-split, Markup for *sandhi and samasa* and also analysis are being tried with programs based on DESIKA and hence, we may use a semi-automatic system towards the end and not all scholars creating content need be engaged in this.

The present document gives a detailed description of the tags which have been defined for the tagging schemes and elaborates the motivations behind the selection of these tags. The objectives of this study are :

- To build a reverse Samasa rule base and example base of Pāninian *Samasa* rules for identification and analysis of Samasa .
- To adapt Monier Williams Sanskrit Digital Dictionary (MWSDD) of Louis Bontes for analysis purpose .
- To adapt available e-corpora and customize them for *Samasa* analysis purpose .
- To build a servlet based online Java engine which will consult the rule base, example base and the linguistics resources to analyze vowel *Samasa* in a Sanskrit text, and will be used in any other Sanskrit processing application .

METHODOLOGY

It is noteworthy here that Sanskrit language has been analysed by Shastric Scholars with out the stage called "POS". (Part of Speech) . The whole process which involves many steps, does not stress upon POS analysis. Hence there was a feeling among Sanskrit Scholars that no POS tagging is required for Sanskrit Languagae analysis . A human being, based on his/her knowledge about the world, takes a decision about which of these analysis is correct. This decision helps him to know which is the main verb and accordingly based on the Samasa. The process involves

- Identifying the word boundaries
- Knowing the meaning
- Analyse each split word at morphological leave
- Do the samasa(compound) analysis Segmentation (samasapadacchedadh)
- Syntactic knowledge with relation .

The process flow of the system is as follows: input Sanskrit text

↓

viccheda eligibility tests

(pre-processing)

↓ subanta processing

↓

search of *sandhi* marker and *samasa* patterns

(*samasa* rule base)

↓

generate possible solutions

(result generator)

↓

search the dictionary

↓

search the results in the corpora (if not found in the dictionary)

↓

output (segmented text)

Conclusion :-

Sanskrit is the communicator of an unbroken knowledge tradition from the vedic times to the present times. Modern Indian languages can benefit from profound knowledge of the texts of Indian intellectual tradition by being able to access these texts in a cost effective manner. Therefore automatic translation from Sanskrit to Indian languages is highly desirable. And no automatic translation from Sanskrit is possible without building such analysis tools .The development part of the synopsis is partial at this point and is likely to be stable with extensive lexical resources. Subsequent research to make this program more efficient will always be going on. The system will be online and it can be accessed at any time .

Reference

Allen, W. Sidney. 1965, Phonetics in Ancient India , London: Oxford University Press. Allen, W. Sidney. 1972

The Theoretical, Phonetic, and Historical Bases of Word Junction in Sanskrit The Hauge: Mouton & Co. Publishers.

Bakharia, Aneesha. 2001, Java Server Pages New Delhi: Prentice Hall of India Private

Limited.Bharati, Akshar, Rajeev Sangal and Vineet Chaitanya. 2004, Natural language Processing: A Paninian Perspective , New Delhi: Prentice Hall of India Private Limited.

Bhat, D. N. Shankara. 1972 ,Sound Change , Poona : Bhasha Prakashan. Academy of Sanskrit Research, Melkote, http://www.sanskritacademy.org/About.htm

Anglabharti, IIt,Kanpur, http://www.cse.iitk.ac.in/users/langtech Anubharti, IIt,Kanpur http://www.cse.iitk.ac.in/users/langtech Anusaaraka, http://www.iiit.net/ltrc/Anusaaraka/anu_home.html

RCILTS, Utkal University , http://www.ilts -utkal.org

MITCHELL, T. M. 1997. Machine Learning. McGraw-Hill

Och, F. J., Ueffing, N., and Ney, H. (2001). An efficient A* search algorithm for statistical machine translation. In Data-Driven MT Workshop.

T. Dhanabalan, T. V. Geetha, UNL Deconverter for Tamil, International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies, Convergences 2003

Natural Language Processing James Allen, Pages: 23--109 Pearson Educations

Computational Linguistics – An International Handbook on Computer Oriented language research and Applications – D.Gruyter