Sanskrit Verb Argument Valence: A Computational Analysis

Subhash Chandra

subhash.jnu@gmail.com
Special Centre for Sanskrit Studies, Jawaharlal Nehru University, New Delhi

Dr. Girish Nath Jha

girishj@mail.jnu.ac.in
Special Centre for Sanskrit Studies, Jawaharlal Nehru University, New Delhi

Abstract

In this paper authors present a methodology to develop a "Sanskrit Verb Argument Analysis System (SVAAS)" for ascertaining verb's arguments and their semantic compatibility with the verb. The work involves two major goals - to develop a knowledge-base for verb-expectancy and to map semantic compatibility (or logic of it in the real world) of the arguments with the verbs. The object language taken for this work is Sanskrit. The data entry is being done with the help of an interface consisting of Sanskrit ontological tree. After the completion of this work, the system will be able to distinguish between different types of entities like human, animal, animate and inanimate etc.

Keywords:

Sanskrit, argument, compatibility, ontological tree, expectancy, valency, verb categorization, valency analyzer, Panini

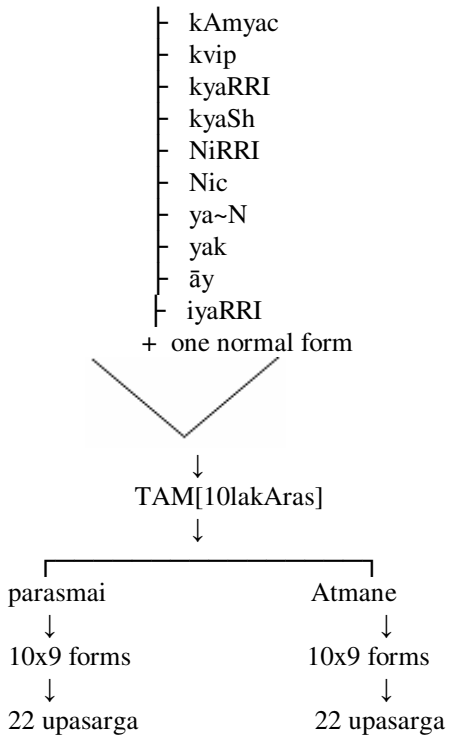
1. Introduction

The relationship between verbs and their arguments is a widely debated topic in linguistics. This comprehensive research and development provides an overview of this important research area, exploring theories of how verb semantics can determine the morpho-syntactic valence of its arguments. Assuming a close connection between verb arguments and syntactic structure, the analysis constructs a bridge between first argument (subject) and second or other argument (object) for semantic and syntactic research. To understand verb argument valence by machine, we need to represent real world knowledge in the machine. Verb is a very important part in a sentence. All other parts of speech are related with the verb with some predefined association. Typically the first, the second and other arguments of the verb uses noun or pronouns only. To identify the valid subject and object in sentence for other processing by machine is very challenging. Human brain knows a verb has certain expectancy for its arguments whereas noun has a mutual compatibility for associating with a specific verb but without representing real world knowledge machines cannot identify the valid subject and object. This paper presents a method to represent knowledge in machine for identification of valid subject and object in a Sanskrit sentence.

2 Sanskrit Verb

Sanskrit has about 2000 verb roots classified in 10 morpho-semantic classes called *gaNas*, and can also be further sub-classified as regular (un-derived) forms and the derived forms. Further, these can have two major semantic classes called *Atmane* and *parasmai* based on who the beneficiary of the action is and also whether something is a universal fact. Further there are conjugated forms in 10 *lakAras* (or TAM - Tense / Aspect / Mood) and 3 x 3 PN (person and number combinations, and can also be potentially prefixed with 22 prefixes. Finally there could be in-numerable *nAmadhAtus* (nominalized verbs). The following tree reproduced from Mishra & Jha (2004) illustrates the distribution of Sanskrit verbs [1]

VR [2000]
- san
- kyac



There is almost no knowledge database related to this area in Sanskrit. Notable exceptions can be VerbNet [2, 3, 4, 5], FrameNet [6, 7, 8] and WordNet [9]. The Sanskrit WordNet [10, 11] is being developed at IIT Bombay. Bandyopadhyay (1998) describes a system on the Bangla verb types, valency and problems in compounding. The purpose of the paper is to describe the ontology of each Bangla Verbs (BV) type by drawing its factual existence in the world of Bangla discourse [12]. Institute of Linguistics, Romanian Academy is also engaged in similar work [13]. Horak Et.al [14] have described valency representation for deep linguistic processing. Atsushi et al (2003) [15] have discussed an empirical method to detect incorrect verb valency occurring in paraphrasing Japanese sentences. Their error detection model involves an ensemble of two error detection models that are separately trained on a large collection of unlabeled positive examples and a small collection of labeled negative examples. Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lucica, Zagreb, Croatia [16] University of Gent, Blandijnberg [17] and Institute of Formal and Applied Linguistics, Charles University in Prague [18] are also involved in the field of verb argument valency.

3. Previous Works

After lot of searching we have not get any knowledge database related to this area excluding VerbNet [3, 4, 5, 6], FrameNet [7, 8, 9] and WordNet [10]. The Sanskrit WordNet [11, 12] is also being developed by Malhar Kulkarni at IIT Bombay. Special Centre for Sanskrit Studies, JNU is also doing related works and published a paper [13].

We got lot of paper that they are claiming of development. We got the following papers over the Internet:

Bandyopadhyay (1998) describes a system in the article on the Bangla verb type and especially on compound verbs, the rule of selectional restrictions of which is discussed here. The purpose of the paper is to 'describe' to understand the ontology of each Bangla Verbs (BV) type by drawing its factual existence in the world of Bangla discourse [14]. Institute of Linguistics, Romanian Academy, Bucharest, Romania is also engaged in these types works [15], Ales Horak Et.al [16]. Advanced Inst. Sci. and Technology discussed regarding a paper they propose an empirical method to detect incorrect verb valences occurring in paraphrasing Japanese sentences. Their error detection model involves ensembling of two error detection models that are separately trained on a large collection of unlabeled positive examples and a small collection of labeled negative examples [17], the Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lucica, Zagreb, Croatia [18] University of Gent, Blandijnberg [19] and Institute of Formal and Applied Linguistics, Charles University in Prague [20] are involved in the field of verb argument valence.

4. Verb Arguments Valence

The verb arguments valence depends on the following principles-

4.1 Expectation of Verb

Intransitive verbs expect one noun as an argument (subject), and transitive verbs expect more than one nouns as arguments with valid compatibility between subject and object/s –

Expectation of Verb			
ITRANS	rAmaH	utpIThikAm	khAdati
English	Ram	Table	eats

Table-1 Expectancy of Verb

The above sentence is wrong because 'Ram' cannot eat 'table'. The verb 'eat' expects an animate as subject and an edible thing (for humans in this case) as an object. Here 'Ram' is a member of the animate class, so this rule is correct for this sentence, but 'table' is not edible for humans. As a result, the compatibility match fails. In other words, if subject occurs as animate and object as an edible object in a sentence in the context of eating then the sentence is acceptable otherwise sentence is not acceptable.

4.2 Compatibility of arguments

The compatibility of verb arguments is another parameter. Verb expects compatible subject/object/s for action to be complete. For example -

Compatibility of noun			
ITRANS	KagaH	AkAshe	uDDayate
English	The bird	in the sky	is flying
Compatibility of noun			
ITRANS	naraH	AkAshe	uDDayate
English	Man	in the sky	is flying

Table-2 Compatibility of Noun

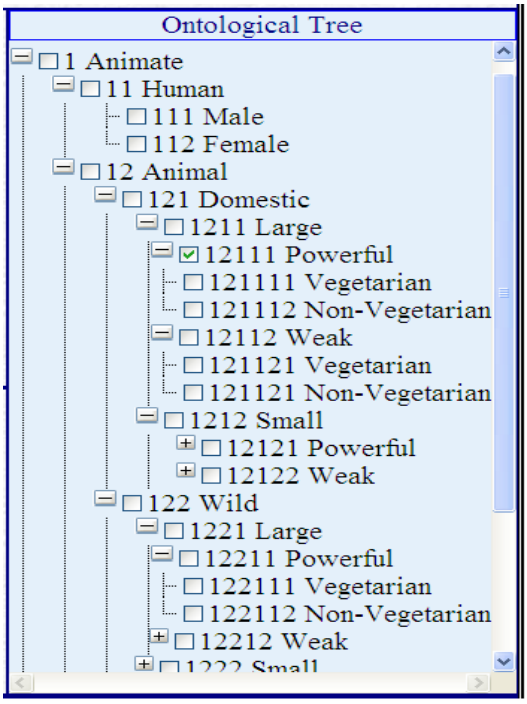
In the examples above, while the first example is acceptable, the second is not. The main difference between the sentences is compatibility of noun. 'Bird' and 'man' are both animate nouns, but there are differences between two nouns in term of whether they can fly or not. This research is trying to solve these type problems.

5. Methodology

We are representing the knowledge in machine using the ontological method [19]. So we have developed an ontological tree with the help of available previous tree structure, corpora of Sanskrit and inherent knowledge based on Sanskrit verb expectancy (subject and object structure). First, we studied corpora and verb requirements for the development of ontological tree. Next, we built an ontological tree to describe these entities and relations. We then assessed the ontological tree for data entry of knowledge-base for Sanskrit noun and pronouns with respect to selected verbs. After the completion of knowledgebase, we developed a rule base called argument valence mapper database. For knowledge-base we require following data-

- an ontological tree (developed by us taking the help from other available ontological trees, Sanskrit corpora analysis and human intelligence)
- Monier-Williams Dictionary (MWD) for Sanskrit
- Sanskrit Verb list
- Sanskrit processing tools developed at JNU
- Sanskrit named entity list collected from the internet and telephone directory
- Developing a data entry user interface for data entry

We are using ontological method for developing knowledge database. The ontological class has been completed through manual effort. A sample of ontological class:



"Fig 1: Ontological Tree"

This is not the final class hierarchy but only a sample for developing ontology. It is customizable through program as per requirements.

6. Database and Rule files required for System

After defining ontological classes we have developed following databases.

6.1 Knowledge Database of Sanskrit Noun and Pronouns

Knowledge Database of Sanskrit Noun and Pronouns (KDSNP) is knowledge database with property information of each base word (right now nouns and pronouns only) collected from corpora and MWD for Sanskrit. The database has id, word, ontological information, numerical value for the ontological information and updated date etc. in the following format:

Word			Ontology
ID	ITRANS	English	
1	gau	Cow	121111
2	aham	I	11
3	tvam	You	11
4	si.nha	Lion	122112

Table-3 Verb Argument Valence Database (VAVD)

Here the number sequence encodes the property set of the word. For example, if the numerical value of any word is 121111 then, it will be *Animate →Animal →Domestic →Large →Powerful →Vegetarian.* If the value is 11 then *Animate → Human*.

6.2 Rule Database for Argument Valence Mapping

Argument Valence Mapper Database (RDAVM) is subject and object mapper database. The database is in following format:

Sr.	Verb Id	Subject	Object
1	VB1	11	31
2	VB1	12	31/122111
3	VB1	122112	11/1211/1212

Table-4 Argument Valence Mapper Database (AVMD)

All acceptable combinations of argument for the specific verb are given in the above Table-8. Suppose VB1 is representing the Sanskrit verb *khAd*, then the subject line represents all possible subjects for *khAd* such as 11 for human, 12 for animal, 122112 for *Animate →Animal →Domestic →Large →Powerful →non-Vegetarian* etc. Object line of table-8 also represents the compatible objects for above mentioned subject such as if the subject occurs as 11 then object must be 31, 122111, 1211, 1212 etc. (follow the Ontological Tree in Fig.1) in the specific sentence according to the verb *khAd*.

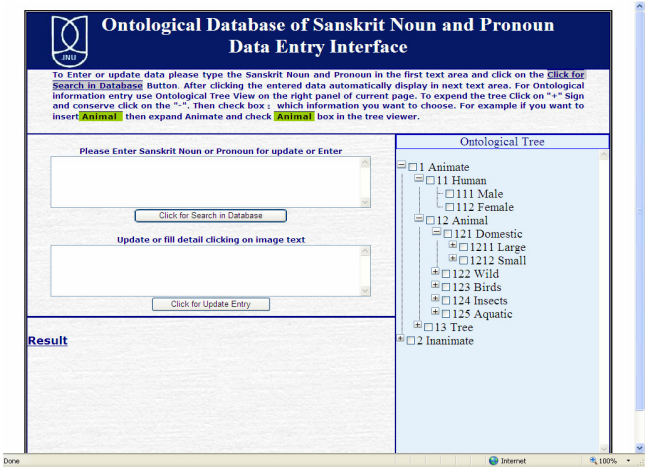
7. User Interface and Technical Details

7.1 User Interface

We have developed web based interfaces for data entry, analysis of verb expectation etc. The details are as follows.

7.1.1 Data Entry Interface

The data entry user interface is web based. The interface allows users to modify existing entry and enters new words with ontological information in database with the help of ontological tree viewer which is displayed on the right panel of the interface. In the interface, there are two text areas and one tree viewer. The first text area is for searching whether the entry exists or not, the second is for updating or inserting in the database and the tree viewer is for ontological information entry. Every node of the tree has been displayed with the check box (□) with the ontological value. After checking in the box of nodes, system automatically returns the ontological value in the second text area. Let us see a sample of ontological class on the data entry interface:



"Fig. 2 : Data entry Interface"

7.1.2 Analysis of Verb Arguments Valence System Interface

This interface accepts Sanskrit sentence, text or corpora in utf-8. User can also upload a Unicode text file through file browse. After clicking on the button, the system returns analyzed data with acceptable subject and object tags.

7.2 Technical Detail

Web based data entry interface has been developed in Python. We are using the MySQL database and text files for the backend and Python Server Pages (PSP) for the front-end and Python for the programs. All the

entries would be done through this interface created by us. There are two options for input – either type the data and search in database. If entry found then user can update the information if entry not found, then then the user can enter word with details. The second option is that the user can upload file with all information and the system will automatically update the database. For the Sanskrit Verb Argument Analysis System (SVAAS) has been developed in python, front end in PSP and the backend is in MySQL.

8. Analysis of Verb Arguments Valence

The SVAAS is under development at this point. The analysis is done with the help of KDSNP and RDAVM. In the RDAVM, we have defined valid argument-1 and argument-2 combination with respect to specific verbs. First, the system gives the ontological information of each input word with the help of RDAVM then checks all ontological information in the RDAVM. If any combination of arguments is found in the RDAVM then the system returns it as valid combination of argument, otherwise it labeled invalid.

Conclusion

The accuracy of SVAAS depends on the size of above mention databases and corpora. This is an effort to providing knowledge to the machine for verb argument valency. The current work is in developmental stage at this point, so it does not cover all Sanskrit verbs. Though, this approach does not promise a complete solution, yet it may be a very effective model for language processing in general. It will be a model for Indian languages for computational processing and can be very useful system for knowledge representation in machines which has been very challenging till date.

References:

[1] Mishra Sudhir K, Jha Girish Nath, 2004, Identifying verb inflections in Sanskrit morphology, in proc of SIMPLE 04, IIT Kharagpur. pp. 79-81.

[2] Karin Kipper-Schuler. (2005). VerbNet: A Broad Coverage, Comprehensive Verb Lexicon. Ph.D. thesis, University of Pennsylvania.

[3] Palmer, Martha. (2009). Semlink: Linking PropBank, VerbNet and FrameNet. Proceedings of the Generative Lexicon Conference (GenLex-09). Pisa, Italy.

[4] Loper, Edward, Szu-ting Yi and Martha Palmer. (2007). Combining Lexical Resources: Mapping Between PropBank and VerbNet. Proceedings of the 7th International Workshop on Computational Semantics. Tilburg, Netherlands.

[5] Kipper, Karin, Anna Korhonen, Neville Ryant and Martha Palmer. (2006). Extending VerbNet with Novel Verb Classes. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy.

[6] Fillmore, Charles J. (1982). Frame semantics. In Linguistics in the Morning Calm, Seoul, Hanshin Publishing Co., 111-137.

[7] Fillmore, Charles J. (1976). Frame semantics and the nature of language. In Annals of the New York Academy of Sciences, Conference on the Origin and Development of Language and Speech, Volume 280, 20-32.

[8] Ruppenhofer, Josef, Collin F. Baker and Charles J. Fillmore. (2002). The FrameNet Database and Software Tools. In Braasch, Anna and Claus Povlsen (eds.), Proceedings of the Tenth Euralex International Congress. Copenhagen, Denmark. Vol. I: 371-375

[9] Christiane Fellbaum. (1998). WordNet: An Electronic Lexical Database. Language, Speech and Communication, The MIT Press, ISBN-10:0-262-06197-X, ISBN-13:978-0-262-06197-1.

[10] Malhar Kulkarni, et al. (2010). Introducing Sanskrit Wordnet. The 5th International Conference of the Global WordNet Association (GWC-2010), 31st Jan - 4th Feb, 2010,

[11] Malhar Kulkarni and Pushpak Bhattacharyya. (2009). Verbal roots in the Sanskrit WordNet. Sanskrit Computational Linguistics, Lecture Notes in Computer Science:328--- 338, Berlin/Heidelberg. Springer-Verlag.

[12] Debaprasad Bandyopadhyay. (1998). Valency of Bangla Verb and the Problem of Compounding. Vahuvacana, Journal of the Linguistic Research Center, Jadavpur, Kolkata. 46-64.

[13] Ana-Maria Barbu. (2008). First Steps in Building a Verb Valency Lexicon for Romanian. Text Speech and Dialogue,

Lecture Notes in Computer Science by Springer Berlin Heidelberg with ISBN: 978-3-642-00154-3, Volume 5246/2008, Page: 29-36

[14] Alesh Horak, Karel Pala, Marie Duzi and Pavel Materna. (2007). Verb valency semantic representation for deep linguistic processing. Proceedings of the Workshop on Deep Linguistic Processing, Association for Computational Linguistics Morristown, NJ, USA, 97-104.

[15] Fujita Atsushi, Inui Kentaro and Matsumoto Yuji. (2003). Automatic Detection of Verb Valency Errors in Paraphrasing. Joho Shori Gakkai Kenkyu Hokoku, Japan, ISSN:0919-6072, VOL.2003;NO.76(NL-156); 53-60.]

[16] Kresimir Sojat, Kristina Vuckovic, Marko Tadic. (2009). Extracting Verb Valency Frames with NooJ. Proceedings of the 2009 NooJ conference and workshop, Tozeur, Tunis

[17] Devos, Filip and et al. (1996). Contrastive Verb Valency and Conceptual Structures in the Verbal Lexicon.. Contrastive Semantics and Pragmatics, Volume 1: Meanings and Representations, Language Sciences, Canada, v18 n1-2, ISSN-0388-000, 319-38.

[18] Viktor Bielicky and Otakar Smrz. (2008). Building the Valency Lexicon of Arabic Verbs. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08, Marrakech, Morocco, European Language Resources Association (ELRA), isbn = 2-9517408-4-0

[19] John F. Sowa, *Knowledge Representation: Logical, Philosophical and Computational Foundations*, Brooks Cole Publishing Co. Pacific Grove, CA, 1999.