Naturlal Language Processing and Sanskrit

Deeptanshu Jha¹,  Dr. Rashmi Jha²,  Varun Varshney³

¹CSE Department, Thapar University, Patiala, Punjab, India  
²Associate Professor (CS), GIBS, GGSIPU, New Delhi, India  
³M.S, University of Texas – Arlington, Texas, USA

Abstract

Sanskrit considered being the mother of all languages possesses a rich grammar which was penned by Panini around 2500 years ago. The dual case is a unique and beautiful feature of Sanskrit which is not present in any other language other than Sanskrit. The importance of dual case is ignored in almost all the languages which lead to confusion while processing dual and plural but this ambiguity is also removed in the highly efficient and organized Sanskrit grammar. This research paper focuses on this Indian treasure and suggests its use for to-day's computer applications. It explores diverse key characteristics of NLP, various problems encountered in NLP are and how Sanskrit successfully overcomes these limitations and fulfills all the requirements of a Natural Language Processor.

Keywords: Artificial Language (AI), Inflection, Natural Language Processor (NLP), Sanskrit, Semantic Nets Etc.

1. Introduction

It is a common and reasonable misconception that natural languages are too ambiguous for transmission of many ideas that the artificial languages can handle and process with great precision and mathematical rigor. However, there is one language which despite being natural has no ambiguity and that is the mother of all Indo-European languages, treasure of India "Sanskrit”. The creator of Sanskrit language was Panini who formulated 3,949 rules. Now even a millennium later, Sanskrit has the strongest and simplest grammar of all the natural languages and surprisingly; the most suited language for Artificial Intelligence and Natural Language Processors. It is really commendable that Panini was able to design a language that can make computers understand the concept of human linguistics without any ambiguity even in this day and age. Certain advantages of
Sanskrit mentioned may find its use in advanced Artificial Intelligence for computers. It is also immensely useful for creating a highly efficient Natural Language Processor.

In this article, we will discuss what NLP is, various problems encountered in NLP are and how Sanskrit Grammar effectively handles these limits and restrictions with precision while accomplishing every requirement of a Natural Language Processor.

2. NATURAL LANGUAGE PROCESSING

2.1 Definition

Natural language processing (NLP) is used for communication between computers and human (natural) languages in the field of artificial intelligence, and linguistics. Being concerned with human-computer interaction, NLP works to enable computers to make sense of human language to make interactions with machinery and electronics as user friendly as possible.

2.2 How NLP Works

A structured approach is necessary to accomplish tasks using computers, with binary code being the basis of all interaction. It is a compulsion for computer languages to be unambiguous. Although, human languages do have a structure (called grammar), a high level of ambiguity remains because words can have different meanings depending on the context.

Let us inspect the given English phrase - "I like apple"
1. Does this phrase refer to the brand apple, or to the fruit?
2. Who is "I" in this context?
and now further deepening the linguistic analysis, consider a sentence like:

"Do you see the man with the glasses?"
1. Does that mean - "you see the man using the glasses"
2. Or does it mean - "you see the man who is holding the glasses"
3. Glass here refers to spectacles or normal glass?

Grammatically both the sentences are correct, which meaning was implied here; depends solely on the present context. Thus NLP has to face a lot of ambiguity during its processing and now we will explore how Sanskrit overcomes all of these hurdles to become the best suited language for NLP.

3. SEMANTIC NETS

Translating a sentence into machine acceptable form is not just a map from lexical item to lexical item and since ambiguity is always present, we need ways to bring out the actual meaning of the sentence. Semantic Nets are used to represent semantic relations between concepts using directed/undirected graphs.

Take for example the sentence - "Varun gave the book to Deeptanshu". This information can be stored as a set of triples:
give, agent, Varun
give, agent, book
give, recipient, Deeptanshu
give, time, past
Here, we can see that central significance is given to the verb and is considered to be the distinguishing characteristic of the sentence.

Now let us explore another example:

Varun, an author living in Mayur Vihar, gives the book to Deep, who is a scriptwriter.

Now, if we read this semantic net, we will get a long and confusing English text - "There is a "Varun" who lives in Mayur Vihar, where Mayur Vihar is a subset of ADDRESS-EVENTS, itself a subset of 'ALL EVENTS', and who is an author which is again a subset of 'OCCUPATION EVENTS'...etc.

The extent to which a semantic net is complex and unmanageable is related to the fact that language is - "natural" and diverges from the precise or "artificial". Further on, we will discuss that one minute difference between Sanskrit language and semantic nets is that Sanskrit grammarians were not aware of the diagrammatic representation and thus they developed all abstract notions in grammatical sentences.

4. SHASTRIC SANSKRIT

Several unique and outstanding features of Sanskrit make it extremely suitable for Artificial intelligence and NLP.

4.1 Word Representation of Properties of Objects/Entities

One of the major differences between Sanskrit and other languages is that in all other languages there is a one-to-one correspondence between the words and the objects represented. In Sanskrit, this one-to-one correspondence exists between the words and the associated properties. For example in English, a tree is called a tree and doesn't reflect its properties. On the other hand, in Sanskrit the word वृक्ष (tree) represents properties of trees as well and not only the tree itself. Similarly other words which can define the properties of a tree can be used to denote a tree.

वृक्ष = something that is cut and felled down
तरु = something that floats
पार्स = something that drinks using its feet
So, all these words can be used to denote a tree because a tree has the above properties. There are a number of advantages of this approach:

1) Today English has approximately 500,000 words, most of which is borrowed from other languages. Before an airplane was invented, there was no word for it in the dictionary and was added only after its invention and thus every new word coined has to be put into the dictionary. In Sanskrit even if a word is coined we don’t need to put it in the dictionary because any person well versed in Sanskrit can use the algorithms of Sanskrit grammar to decode its meaning as the word coined will ultimately represent a property of the object.

2) There are virtually infinite words in Sanskrit, whereas in English there will always be a finite number of words in the dictionary. This is so because, words for an object can be coined based on its properties and using the same properties anyone can understand the meaning of the word.

Here we elucidated those words in Sanskrit that represent properties of the object and not the object itself. What separates Sanskrit from other languages is the enormous ratio of words representing properties to words representing objects. Let this ratio be X;

For most languages: $X < 1$ - OR- $X= (approx) 1$ - OR- $X(slightly) >1$ For Sanskrit: $X 100000$

So even Sanskrit has words that represents objects only but their count is negligible in comparison to the words representing properties.

4.2 VIBHAKTI

In this section, we will see how Sanskrit uses programming concepts such as classes, objects, and pointers to shorten the language.

Let us take this sentence as an example:

\[ \text{पूर्व: परिहर्तः: प्रत्ययः द्विपः: पसूः} \]

It means that - A stupid person must be avoided. He is like a two legged animal in front of the eyes.

We can see that Sanskrit language is very economic in its usage of words as compared to English. As mentioned in 4.1 words in Sanskrit represent properties so the given 5 words also represent some properties

\[ \text{पूर्व: = (the property of being) stupid} \]
\[ \text{परिहर्तः: = (the property that makes one) avoidable (by others)} \]
\[ \text{प्रत्ययः = (the property of being) in front of the eyes} \]
\[ \text{द्विपः = (the property of) having two legs} \]
\[ \text{पसूः = (the property of usually being) tethered} \]

In the spoken language we talk about objects and not properties, so we need to force the words to represent objects and this way of making a word (Which represents property) an object is called vibhakti.

So, पूर्वः represents the property of being stupid, but पूर्वः (which is a vibhakti of the word पूर्वः) represents an object/person who is stupid.
Similarly

पतिरूहथ्नः = an object/person who must be avoided
प्रत्यक्षः = an object/person located in front of the eyes
ढङ्गः = an object/creature having two legs
पत्थरः = an object/creature who is tethered = a beast or cattle

Now we have five words representing five different properties and with the help of vibhaktis, we have converted them to represent five objects having those five properties. Now the golden rule of Sanskrit grammar which states that **words having the same vibhakti represents the same object** i.e. the five different words are like pointers that point to the same object because they all have the same vibhakti.

![Figure 2: Basic Object Pointers for First Vibhaktis](image)

All object pointers point to the same object because they all are first vibhaktis.

**Figure 2: Basic Object Pointers for First Vibhaktis**

Thus we see that a word in Sanskrit is like a class in object oriented language (without methods) and its "vibhaktified" form is like a pointer to an object of that class.

### 4.3 Dual Case

This is a unique feature of Sanskrit language which is not supported by most of the languages creating confusion while processing dual and plural. A comparison study between the given four languages is presented below:

1) French
   - Singular Case: chez le garçon - in the boy
   - Dual Case: entre les garçons - between the boys
   - Plural Case: parmi les garçons - among the boys

2) Spanish
   - Singular Case: en el niño - in the boy
   - Dual Case: entre los chicos - between the boys
   - Plural Case: entre los chicos - among the boys
3) English
Singular Case: in the boy
Dual Case: between the boys
Plural Case: among the boys

4) Sanskrit
Singular Case: baalkey (in the boy)
Dual Case: baalkayo: (between the boys)
Plural Case: baalkeshu (among the boys)

Thus we can see that only in Sanskrit there is a clear difference between dual and plural case and thus we can get an error free NLP, whereas in other languages even the human mind can get confused between dual and plural case.

4.4 Inflection Based Syntax
Another unique feature of Sanskrit is its inflection based syntax which makes the overall meaning of a sentence independent on the position of its constituent words. An inflection of a word is a different form of that word and is used for enhancing the meaning of the original word. When we say that English is a weakly inflected language, we mean that English language seldom uses different forms of words to represent enhanced meanings of that word, but on the other hand it uses totally unrelated new words to represent the enhanced meaning whereas in Sanskrit with the help of inflected word we can easily convey the enhanced meaning without the help of any new unrelated words.

For example consider the sentence: The meat was eaten by a dog.
Which when translated in Sanskrit would be:

नांसं खासिं (meat) इति इति (act of eating) हुक्करेय.

Thus here we see that in English to convey the information that the agent is the dog we had to use a new unrelated word - "by" but in Sanskrit there was no use of any new unrelated word because we used the inflected word इति.

Now here in the sentence the word - "was" performs two functions

1) It conveys that the act of eating has already been performed.
2) By appearing after meat it conveys that act of eating was performed on the meat.

So if we try to jumble up the words in the sentence the whole meaning of the sentence will change or in many combinations there will be no meaning at all. But in Sanskrit these two functions are performed by vibhakti and not word order thus jumbling up of the words have no effect on the overall meaning of the sentence.

1) As नांसं (meat) and खासिं (act of eating) have the same vibhakti, so they apply to the same object and we know for sure that खासिं applies to नांसं and not इति, whatever be the order of the words.
2) To show that the act of eating has already taken place, we use the inflected word खासिं which has the information that the act of eating has taken place engrained in it irrespective of the position of खासिं in the sentence.
4.5 Speech Therapy

The Indian book of music Samaveda is a Sanskrit Script and we can implement Sanskrit in coding, like coding a musical instrument. Just like any musical instrument each and every Sanskrit word has its own pronunciation with a certain Frequency, Stress and Amplitude. The pronunciation and ac-cent of Sanskrit words is based on physics and these words have the power of vibration. These properties can help in creating a highly efficient speech therapy for NLP.

For example in English language two words such as "hole" and "whole" have the same pronunciation. Alternately there are words such as enough which are written as "e n o u g h " but pronounced as "enuff " which is very confusing for the NLP but in Sanskrit there is no such ambiguity because all the words are formed with great care and also because there is no colloquial version of the language. It did not face any accent change, so Sanskrit can be very helpful for machine speech therapy.

5. CONCLUSION

In this paper, we tried to describe the reasons to support Sanskrit as a language for Natural Language Processing as compared to other languages because of it vast and intelligent grammar and various properties such as

1) Words describing the properties rather than objects
2) Special attention to dual case
3) Extensive use of vibhaktis
4) Support for high inflection, and
5) Extremely refined pronunciation of each and every word.

Using Sanskrit preferentially for research and development should be explored, rather than making use of a foreign language adding fame to the Indian heritage.

REFERENCES