Intelligent Sanskrit

Author: Ashish Kumar Sahani

This was first published in June 2013 issue of RENews , a Monthly Magazine of IIT Madras

The Sci-Fi movie Ironman-III was released a few weeks ago. The protagonist of the movie, Tony Stark, is a rich and ingenious industrialist. The visuals of Tony's lab are a treat to the eyes of any engineer. I was particularly fascinated by Tony's intelligent computers which executed direct spoken commands in natural English. Although the story line of the movie is completely fictional, yet it palpably depicts a long standing dream of computer scientists, to enable computers to understand natural languages spoken by humans. Natural language processing (NLP) involves many challenging steps including speech recognition, syntax decoding and understanding the semantic relationships among words in a sentence. NLP is considered as an AI-Hard problem [1] .

There is a widespread belief that natural languages are unsuitable for computer processing. The expression of ideas in natural languages is generally ambiguous and cumbersome which computer languages, like C or Java, can render with great precision and mathematical rigor. Yet, perhaps most of us will not be able to enjoy poetry or drama written in C! Artificial languages are not suitable for carrying subjective ideas while natural languages are not designed for objectivity. But this apparent dichotomy probably is not true for all natural languages. There is at least one language, Sanskrit, that was a living spoken language for at least 2000 years and has a considerable literature of its own.

Even a cursory study of this language reveals the almost mathematical structure of its grammar which allows it to express ideas in most unambiguous and precise manner. Contrarily, it is also the carrier of one of the most terse philosophical texts in the world, Vedas and Upanishads, with highly subjective expressions. This can be attributed to its unique capacity of expressing both the subjective and objective ideas with equal ease. It's words are not merely dictionary based, but are derived from a few root sounds (about 200) which are the most natural sounds associated with the actions that they are intended to express. When these roots are combined with appropriate suffixes and prefixes they can give rise to millions of words. Mastery over a handful of basic sounds and structural combination rules (*sandhi*) enables the speaker to express any idea; if you have something to say, you can always find a word for it. You can even form your own vocabulary and the idea will still be communicated. In this article I will try to elucidate a few apparent reasons that perhaps make Sanskrit the strongest

candidate for NLP.

Generally, the first step in NLP is recognition of the spoken word transmitted via sound. The current state-of-art systems in speech recognition are nowhere near the capabilities depicted in the above mentioned Sci-Fi flick. We have yet to go a long way before reliable voice recognition systems become a part of our life. Most voice recognition research has been focused towards English, which is a non-phonetic language. It is obvious that phonetic languages, having direct correspondence between sounds and symbols, are more suitable for speech recognition. All major Indian languages are phonetic except for Tamil, which is partially phonetic † . While we will probably confuse the computer, when we will say **ch** emistry and **ch** at with **ch**carrying different sounds for both words, but there will be no issues if we say the same thing to a speech recognition algorithm designed for a phonetic language like Hindi, where we will sayकेमिस्ट्री and चाट which have well defined characters for each symbol. On a lighter note, it is not possible to have competitions like spelling-bee in Indian languages. Sanskrit, being strongly phonetic, has an obvious advantage over English as a candidate for NLP.

Apart from being phonetic all its words are actually derived from few root sounds, by adding appropriate suffixes and prefixes. This provides it with a huge vocabulary where new words can be manufactured by the speaker at will. Whenever a listener hears a new word he need not go back to the dictionary to find its meaning; if he knows the roots and the proper decoding techniques he can intelligently guess the meaning of any new word. For example suppose you found a word hitherto unheard of like: उपनीसत् ; we may decode it as follows:

उपनीसत् = उप + नी + सत्

उप = Near

नी = Take

सत् = Truth

So, उपनीसत् **means something which takes near the Truth** . This capability which is full blown in Sanskrit is seen even in other Indian languages to some extent. This malleability to form desired word makes it the ideal vehicle of poetic and philosophic expressions. In NLP this feature can be fruitfully exploited to recognize a vast number of words with a small database of root words, suffixes and prefixes.

The third and probably the most important character of Sanskrit language is the tightly held and well defined semantic relationship between words. To understand this concept we will have to understand a few basics of noun and verb forms in Sanskrit. All forms of any noun are represented as a 7×3 matrix where rows represent the auxiliary action over the noun, known as *Vibhakti* , and columns denoting the number. It is important to note here that Sanskrit has an additional dual case. Let's try to understand the noun forms by taking one simple example of noun बालक (boy) (Table I).

Table I: बालक शब्दः (अकारान्त पुल्लिङ्ग)

Vibhakti	**Singular**	**Dual**	**Plural**	**Hindi auxialiary action**	**English Case name**
1 st	बालक :	बालकौ	बालका :	बालक ने	Agent

2 nd	बालकम्	बालकौ	बालकान्	बालक को	Object
3 rd	बालकेण	बालकाभ्याम्	बालकै :	बालक केद्वारा	Instrumental
4 th	बालकाय	बालकाभ्याम्	बालकेभ्य :	बालक केलिये	Dative
5 th	बलाकात्	बालकाभ्याम्	बालकेभ्य:	बालक से	Point of Departure
6 th	बालकस्य	बालकयो :	बालकाणाम्	बालक का	Possessive
7 th	बालके	बालकयो :	बालकेषु	बालक मे	Locality

At the first glance we can immediately see the perfect structural arrangement of all 21 forms of noun in a matrix arrangement, with position of each element corresponding to a unique auxiliary action over the noun. For demonstrating a complete example I am also including the first two rows of the table for the noun पुस्तक (book) (Table II).

Table II : पुस्तक शब्दः (अकारान्त नपुंसकलिङ्ग)

Vibhakti	**Singular**	**Dual**	**Plural**	**Hindi auxialiary action**	**English Case name**
1 st	पुस्तकम्	पुस्तके	पुस्तकानि	पुस्तक ने	Agent
2 nd	पुस्तकम्	पुस्तके	पुस्तकानि	पुस्तक को	Object

Similarly, each tense of all verb forms is arranged in 3×3 matrices. The three rows represent the person of the action and the three columns the represent number. I elucidate the form of present tense of पठ् (read) root in Table III.

Table III : पठ् धातु (लट् लकार् – Present Tense)

Person	**Singular**	**Dual**	**Plural**
Third	पठति	पठतः	पठन्ति
Second	पठसि	पठथः	पठथ
First	पठामि	पठावः	पठामः

Now, we have just enough resources to understand the basics of semantic network in Sanskrit. Let's take one example:

English Sentence: Boy is reading a book.

Hindi Translation: बालक किताब को पढ़ता है |

It's quite easy to convert the sentence to Sanskrit by looking at Table I, II and III. Boy is 3rd person Singular number so we choose matrix element (1, 1) i.e. पठति for expressing the act of reading, being

performed by a third person. There are two nouns, boy and book. Boy is the agent of the action and is singular number so we choose (1, 1) element from the matrix shown in Table I i.e. बालक :. Book is the object and thus second row and first column has to be selected from the auxiliary matrix ofपुस्तक noun which is पुस्तकम् (from Table II).

Thus the sentence in Sanskrit is: बालक : पुस्तकम् पठति |

Let's now do a bit of stress testing to see how much damage each language can withstand when we try to modify the semantic structure by simple permutation of words:

English Sentense: Boy is reading a book.

Modification 1: Boy is a book reading. (Incorrect)

Modification 2: Is boy a book reading. (Incorrect)

Modification 3: Book is reading a boy. (Incorrect)

Modification 4: Reading is a book boy. (Incorrect)

In the above modifications we see that even with mild modifications in sequence of words, semantic structure of English breaks down very easily. Now let's consider the Hindi sentence:

Hindi Translation: बालक किताब को पढ़ता है |

Modification 1: किताब को बालक पढ़ता है | (Correct)

Modification 2: पढ़ता है किताब को बालक | (Correct)

Modification 3: पढ़ता किताब को बालक है | (Correct)

Modification 4: बालक को किताब पढ़ता है | (Incorrect)

Here we can see that Hindi is much less fragile and can take quite a bit of torture on the syntax. Only place it breaks down is when we move the term " को " which comes after the object (किता ब) and place it after the Agent (बालक). In Sanskrit this "को " is built into the object term itself (Second row of noun forms) and there is no breakdown when we shuffle the words in a sentence.

Sanskrit Translation: बालक : पुस्तकम् पठति |

Modification 1: पुस्तकम् बालक : पठति | (Correct)

Modification 2: पठति पुस्तकम् बालक : | (Correct)

Modification 3: बालक : पठति पुस्तकम् | (Correct)

Modification 4: पुस्तकम् बालक : पठति | (Correct)

Thus the position of a word in a sentence does not matter in Sanskrit. This is because each noun form carries the information of auxiliary action along with it; पुस्तकम् doesn't mean only book but पुस्तक को (Book as the object). Similarly, each verb form carries the person information along with it. This tightly held semantic structure makes Sanskrit probably the best language for NLP. Additionally, in Sanskrit most words carry gender and number information as well.

As in every other natural language, Sanskrit also has many exceptional cases, but here *astadhyayi* of Panani [2] comes to our rescue. It is a phenomenal work of Sanskrit grammar with 3959*sutras* and lists all rules and exceptions in Sanskrit in highly compact and quite terse form. The *sutras* are totally non-redundant and succeeding principles are many times derived from the preceding once like mathematical equations. Panini's rules can be encoded into the NLP system for Sanskrit to take care of all exceptions.

If we try to visualize a NLP system designed in Sanskrit then it will have the following major steps in its execution:

1. Detection of separate words from the recorded or live speech signals.
2. Words are converted to a phonetic script.
3. Words are broken down into prefix, suffixes, nouns and verbs.
4. Root words are found for each word. For example if detected verb is पठासि then root is found as पठ्.
5. Generation of table of the verb or noun forms of root words.
6. Finding the complete semantic relationship and meanings by scanning the position of the words in the tables.
7. Checking for grammatical or syntactical exceptions using Panini's rules.

The above steps are only speculative and determining the exact challenges will require more rigorous analysis. But, it can be seen that modern field of AI shares a mystical relationship with this ancient language. Prof. Rick Briggs of NASA wrote in one of his remarkable paper in this field [3] that, "(Sanskrit is the only language) where we find that a natural language can serve as an artificial language also, and that much work in AI has been reinventing a wheel millennia old." Unfortunately, like most of our other prized inheritances, we have completely neglectedSanskritic studies in India. Most institutes of Sanskrit studies are in pathetic condition with scholars being paid paltry amounts as stipends [4, 5] . In our own institute, IIT Madras, we have courses being offered on French, German and even Chinese, but Sanskrit never features in our course lists. It is unfortunate and perhaps even dangerous to neglect this language that is so important for the cultural unity of this country.

† Tamil written in *Grantha* script is fully phonetic, with one to one correspondence between sounds and symbols.

References:

1. http://en.wikipedia.org/wiki/AI-complete
2. http://en.wikipedia.org/wiki/P%C4%81%E1%B9%87ini
3. Knowledge Representation in Sanskrit and Artificial Intelligence, Rick Briggs, AI Magazine, Vol. 6, No. 1, (1985)
4. http://www.sanskrit.nic.in/scholarship12-13.jpg
5. http://www.sanskrit.nic.in/Notification%202012-13.pdf

Acknowledgements: The author owes all his knowledge of Sanskrit to Sri Nikesh Rajagopalan , faculty of Samskrita Bharatiat IIT Madras. The author also thanks Vijay K . Gurugubelli , Ph.D. scholar at Dept. of EE, IIT Madras, for proof reading the article and giving valuable suggestion based on his long association with Sanskrit language.