International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 9 (2015) pp. 24173-24178

http://www.ripublication.com

Part-of Speech Tagger For Sanskrit: A State of Art Survey

Sharadha Adinarayanan
IV B.Tech Computer Science and Engineering
School of Computing, SASTRA University
Thanjavur, India
Sharadha.adinarayanan@gmail.com
N.Sri Ranjaniee
IV B.Tech Computer Science and Engineering
School of Computing, SASTRA University
Thanjavur, India
nsriranjaniee@gmail.com
Naren. J
Assistant Professor – II
Computer Science and Engineering
School of Computing, SASTRA University
Thanjavur, India
naren@cse.sastra.edu

Abstract

POS Tagging is becoming popular now-a-days as it is being applied in many languages of the world. The State of Art Survey on POS Tagging for Sanskrit Language is done with the intention to expose the beauty in the language's Grammar and also to know the way in which Natural Language Processing Techniques could be used in doing the same. The Survey focuses on the Grammar, Models that are being used in general for POS Tagging and its applications to Sanskrit Language as a whole.

Keywords: Sanskrit,Part-of-speech tagging,morphological analysis

Introduction

POS Tagging is the process of attaching the appropriate tags to the words in unprocessed sentences. It is also called as grammatical tagging or word category

disambiguation. The paper revolves about creating a central database for root words and to which many sub tables for each grammatical category can be mapped. A separate table for rules is also created. The manually annotated corpus is searched when the traversal through different table ends. Thus appropriate tags are assigned for the annotated text.

NLP In Sanskrit

Sanskrit a language of classical literature, rich and abundant, oldest language in the world is considered to be the origin of many other languages. Panini, known for Sanskrit grammar formulated the 3,996 rules in Sanskrit Ashtadhyayi, a work which deals with morphology, syntax and semantics in Sanskrit grammar. The language has a rational derivative power. In Sanskrit, Etymology is a method used to find the derivatives. Etymology helps in understanding that every word has a meaning; every single word has a meaning built into the word itself, i.e. meaning of that word is contained in the root hidden in the word. Hence a Sanskrit scholar need not use the dictionary everytime to find the meaning of any word. Panini scientific theory introduces the period of Classical Sanskrit.[6]

A. Lexical Analysis:

Lexical Analysis is a process of breaking a sequence of characters into its smallest meaningful unit. Example: In a sentence, words are the smallest units called tokens, whereas for words, letters are smallest unit. To perform lexical analysis a lexer / tokenizer / scanner is used. In Sanskrit, an addition of prefix or suffix to a word may result in another meaningful word. For Example: adding a prefix ‘ah’ to the word vigna becomes the opposite of the word. There can be infinite word formations possible which are well structured.

B. Syntactical analysis:

It is the process of analyzing words in a string according to the formal grammar rules of that particular language. The language has wide ranges of conjugations, declensions with flexible sentence structure. There are rules in Sanskrit, which can be applied to the words in sentences, so as to recognize each one of them according to their grammar. Also, every word in Sanskrit has a unique root which helps in identifying the part of speech. Further classification is done according to the available rules of Sanskrit grammar.

C. Morphological analysis:

Morphology is identifying and analyzing the structure of a given language’s morphemes and other linguistic units such as affixes, root words, parts of speech, intonations and stresses. Since Sanskrit is a systemized oriental language, the identification of a sentence or the morphological analysis can be done as follows:

identifying<root><derivativesofroot><object><action><prefix/suffix> in the sentence.

D. Phonology:

The writing of Sanskrit letters is directly linked with its sound and pronunciations. Thus it is phonetic and there is a natural, logical and easy-flowing way to proceed with understanding and learning Sanskrit. There are no silent letters in Sanskrit compared to that with English. Each letter produces a vibrant resonance in the sounds well-expressed in Sanskrit speaking.

E. Semantic analysis:

It is the process of correlating syntactic structures from phrase, clause, sentence and paragraphs level to the level of writing their language independent meanings. Sanskrit is a language which is very natural, its structure is systematic and versatile in its usage even while relating its purity and uniformity.

Grammar In Sanskrit

Devanagiri script is mostly used script for Sanskrit. Two major classifications of alphabets in Sanskrit language are: vowels (Swaravarna) and consonants (Vyajnjanavarna). Words are classified as noun, verb, indeclinable noun and verb can also be a part of following categories with respect to their usage: pronoun, adjective, adverb.[7][8]

Noun (shabda)

1. Suvantapada -masculine , feminine, neuter
2. Taddhitapada.

Verb (dhaatu)

1. Tinatapada- Parasmaipadi, Aatmanepadi, Ubhayapadii
2. Krridantapada
3. Nijanta
4. Sannanta
5. Yannta
6. Naamadhaatu

Indeclinable (Avyaya)

1. Avyaya
2. Upasarga
3. Nipaata

Also Noun and Pronoun can be classified according to Case-Ending (Vibhakthi),singular,dual,plural and first person,second person and third person.

Models In POS Tagging

There are two models in POS tagging

- Supervised model
- Unsupervised model

Both these two models can be further classified into Rule-based, Stochastic, and Hybrid models.

In the paper a comparative study between rule-based and Stochastic Approach is done to check the accuracy.

A. Rule-based model:

In Rule-based approach, there are two steps to find the correct tag. In first step, all the possible tags for every word in a given sentence are found. The Second Step focuses on using a set of rules to assign the suitable tag for every word in the given input. Here the accuracy depends up on the way in which the grammatical rules are framed and also the presence of word in the database. Hence for achieving high accuracy, an extensive set of hand written rules to find the exact tag of a word in the sentence is required.[9]

B. Stochastic Model:

The Model uses "Pick the most appropriate tag for this word" formula. Various models used in stochastic approach are HMM, MEMM and CRF Model. Stochastic Model uses frequency and probability to find the correct tag. At first, it the most frequently used tag for a particular word in the given sentence is found. Then the information to assign the correct tag for the word is used. The Probability is computed by the N-Gram Method which clears the ambiguity. An N-Gram Method could be unigram, bigram and trigram method. The disadvantage of the model is, all set of tags for every word in a given sentence which are not applicable to the grammar rules of language but has been chosen for POS tagging is found. For a given sentence, Hidden Markov Model uses the tag sequence which amplifies the formula:

P(word | tag) * P(tag | previous n tags)

C. Hybrid model:

It is the combination of both rule-based and stochastic models. The part-of-speech tagger using hybrid model delivers high accuracy compared to individual rule-based or stochastic models. There are two main tasks in hybrid model, at first, it uses a set of hand written rules provided by rule-based approach, and then the result of it will be

further subjected to probability and frequency analysis of stochastic model to assign the appropriate pos tags to every word in a given sentence.

Literature Survey for Sanskrit

The paper by Oliver Hellwig has information about the development of a stochastic tagger for unpreprocessed Sanskrit text. Markov Model is used for tokenizing the text, while Hidden Markov Model performs POS tagging. Modified form of Viterbi Algorithm is also used. The Tagger is used for digitization of Sanskrit Texts. How the design of tagger is influenced by the linguistics problems present in Sanskrit language is discussed in the article and short passage texts are used to evaluate performance of the tagger. For testing the accuracy and performance of the tagger, five passages namely Lingapurana, Visnusmriti, Mulamadhyamakarika, Gitagovinda, Kamasutra were examined.[1]

The paper by R Muni Prasanthi .et.al have proposed and implemented Tree Tagger for Sanskrit Language. ie, annotation of text with Part-Of-Speech tagging is done using a tool called Tree Tagger. POS Tagging is implemented using the Training and testing phase. Suitable tags are assigned for annotated texts in Sanskrit.[2]

The paper by Namrata Tapaswi .et.al have presented a rule-based POS Tagger for Sanskrit Language. Rules are stored in database. Sanskrit sentences are parsed so as to be assigned appropriate tag to each word using suffix stripping algorithm, wherein the longest suffix is searched from suffix table and tags are assigned. The results are tested for 15 tags and 100 words. 90% accuracy using Rule-based approach for POST for Sanskrit is achieved.[3]

The paper by Amba Kulkarni focuses on the Deterministic Dependency parser for Sanskrit. Depth First Traversal Technique is used wherein morphological analysis of words is represented by the nodes of the graph. Dynamic programming is guaranteed by stacks of intermediate results. Also an interface described for users, which has many parsing techniques, wherein a user, can choose the parse technique for implementation. Canonical form of input is considered, Left-Right Deterministic Parsing with Dynamic Programming is used while Pruning is done using Shallow Parser.[4]

Acknowledgment

I am highly grateful and express my earnest and humble thanks to my project internal guide Mr. J. NAREN, Assistant Professor II, Department of CSE, School of Computing, SASTRA UNIVERSITY for his timely help, valuable suggestion, guidance and moral support throughout the completion of the project.

References

[1] Oliver Hellwig “Sanskrit Tagger,a Stochastic and Lexical POS Tagger for Sanskrit”.

[2] R Muni Prashanthi,M.Sirish Kumar,R.J.Rama Sree,”POS Tagger For Sanskrit”International journal of Engineering Sciences Research,Vol-04,(2013), ISSN:2230-8504; e-ISSN-2230-8512.

[3] Namrata Tapaswi, Suresh Jain “Treebank Based Deep Grammar Acquisition and Part-Of-Speech Tagging for Sanskrit Sentences”.

[4] Amba Kulkarni,”A Determnistic Dependency Parser with Dynamic Programming for Sanskrit”.

[5] http://sanskrit.samskrutam.com/en.grammar-tutorial- chapters.ashx

[6] Rijuka Pathak, Somesh Dewangan,”Natural Language Chhattisgarhi:A Literature Survey” International Journal of Engineering Trends and Technology(IJETT)-Volume 12 Number 2– Jun 2014

[7] http://chandanasamskritam.blogspot.in/

[8] http://sanskrit.jnu.ac.in/post/post.jsp

[9] http://nlp.stanford.edu/software/