A Survey on Decoding Sanskrit Language into English: Using Morphological Analysis

Tushar Somwanshi , Amirsohel Shaikh, Prasad Shete, Shekhar Shelar
Under the guidance of Mrs.Abha Pathak
Department of Computer Engineering
DYPIET, Pimpri.

Abstract: The simplest model of a computing device is a finite automation. It has a central processor of finite capacity and it follows the idea of state. It can also be given a formal mathematical definition. Finite automata are used for pattern matching technique in various text editors, for compiler lexical analysis. In this paper, using deterministic Finite automata(DFA) for morphological analysis , we are presenting our work to build dependency parser for Sanskrit language. Another useful thought is the thought of nondeterministic automaton. We can prove that deterministic finite automata, DFA, looks for the same class of languages as NDFA, i.e. they are equivalent formalisms. It is also possible to prove that for given a language L there exists a unequal (up to isomorphism) minimum finite state automaton that accepts it, i.e. an automaton with a minimum number of states. The automata in the examples are deterministic, that is, once their state and input are given, their growth is unequally determined. There are two widespread annotations schemes for parsed structures viz. the constituency structure and the dependency structure. The constituency trees mark the relations because of positions and the dependency relations mark the semantic dependencies.

Keywords: Discourse analysis, Sanskrit, rule based, semantic mapper, Mapping rules.

INTRODUCTION:
Sanskrit has a wealthy tradition of linguistic analysis with strict discussions and arguments on various aspects of language analysis ranging from phonetics grammar logic ritual exegesis and literary theory which is not only useful for analyzing Sanskrit language but it also has much to offer computational linguistics in these areas. The series of conventions in Sanskrit Computational Linguistics the consortium project done by the Technology Development for Indian Languages (TDIL) and the exploration of every individual scholars and the collaborations among them resulted into a) development of several tools ranging from to discourse annotators, lexical resources ranging from annotated corpora. Parsing is the linguistic input; that is, the benefit of grammatical rules and other knowledge sources to analyze the functions of words in the input sentence. Over past 50 years, getting thorough and unambiguous parse of natural languages has been a subject of wide interest in the intelligence. Instead of giving substantial amount of information manually, Machine Learning algorithms are used in possible NLP task.The most important elements in this are state machines, formal rule systems, logic, probability theory and other machine learning tools. These models lend themselves to a low number of algorithms from well-known computational paradigms. One of the most valuable of these are state space search algorithms, and dynamic programming algorithms. The need for unambiguous representation has led to a great effort in stochastic parsing. Sanskrit, also known as Indian Networking language, has a vast collection of literature in nearly all branches of knowledge like astronomy, mathematics, logic, philosophy, medicine, technology, dramatics, literature, poetics. It was the midway of communications for all serious discourses and communications till recent times.

The main reasons behind the difficulty in accessing Sanskrit texts are:

• Sanskrit is influenced by the oral tradition, and therefore the Sanskrit texts are continuous
Strings of characters without any punctuation marks or word or sentence limit. The
Characters at the circumstance of limit undergo euphonic changes making it is hard to 'guess' the limitations.
• Sanskrit is very wealthy in morphology and is inflectional. This also makes it difficult to remember various inflections of a word, which is different from the last character of the word and its gender.
• Despite of substantial vocabulary in today's Indian languages is from Sanskrit, there have been some cases of meaning shifts, meaning enlargement and meaning compression , making it difficult for an Indian to understand the Sanskrit texts faithfully, if he knows the original meaning of the words.

RELATED WORK:

- **Changing Phrase Structures to Dependency Structures in Sanskrit**

- **Author Name- Pawan Goyal Department of CSE Indian Institute of Technology Kharagpur, India – 721302**

  From this paper we Refer-
  - This work focused mainly on conversion from citizenry to dependency structure.
  - The sentences in our dataset are chosen from, which is an authentic book for higher learning of Sanskrit, covering a huge range of grammatical constructions.
  - The tool was tried on a database of 232 sentences and the initial results were reassuring. Explicitly, most of the cases of error were linguistic problems and required further discussion. The phrase labels indicating the case labels are an important extension of the constituency trees to accommodate morphologically rich languages.

- **Designing a Constraint Based Parser for Sanskrit**

  Author Name-
  Amba Kulkarni, Sheetal Pokar, and Devanand Shukl Department of Sanskrit Studies, University of Hyderabad, Hyderabad
  From this paper we Refer-
  - Verbal understanding of any articulation requires the understanding of how words in that utterance are related to each other.
  - Such knowledge is generally available in the form of understanding of grammatical relations. Descriptive grammars describe how a language codes these relations.
  - Hence the knowledge of what information various grammatical relations convey is usable from the generation point of view and but it is not in the case of analysis point of view.

ISSN: 2278 – 1323

- In order to develop a parser based on any grammar one should know absolutely the semantic content of the grammatical relations expressed in a language string, the indication for extracting these relations and whether these relations are articulated explicitly or implicitly.

- **Discourse Analysis of Sanskrit texts**

Author Name-
Amba Kulkarni and Monali Das Department of Sanskrit Studies, University of Hyderabad apksh@uohyd.ernet.in, monactc.85@gmail.com
From this paper we Refer-

- The last decade has seen rigorous hustle in the field of Sanskrit computational linguistics belonging to word level and sentence level analysis.
- In this paper we point out the requirement of special focus on Sanskrit at discourse level owing to specific trends in Sanskrit language in the production of its literature grazing over two millennia.

4. Sanskrit as a Programming Language and Natural Language Processing

Author Name-
Shashank Saxena and Raghav Agrawal C.S C.S, IIET IIET.
From this paper we Refer-

- The significant aspect of our way is that we are not trying to get the full semantics immediately, rather it is derived in stages depending on when it is most appropriate to do so.
- The results we have got are quite encouraging and we hope to consider any Sanskrit text unambiguously.
- We have showed the parsing of a Sanskrit language Corpus employing techniques accomplished and advanced in our previous section.
- Our analysis of the Sanskrit sentences in the form of morphological study and relation study is based on sentences as shown in the paragraphs in previous section.

- **UNL Placed Bangla Natural Text Conversion Predicate Preserving Parser Approach**

Author Name-
Md. Nawab Yousuf Ali, Shamim Ripon and Shaikh Muhammad Allayear
Department of Computer Science and Engineering, East West University Dhaka, Bangladesh

From this paper we Refer-

- This paper presented a unique technique named Predicate Preserving Parsing to convert a natural language text, we considered Bangla, into UNL expressions.
- The UNL expressions conserve the semantic architecture of the natural language texts and can be transformed into any other language using language specific analysis and generation rules, and dictionary entries.

- **ARCHITECTURE**

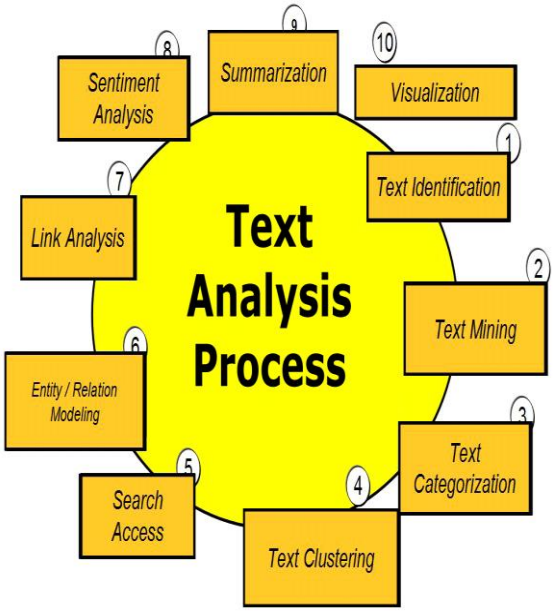

Fig No 1 Text Analysis Processes

Explanation-

Semantic analysis is in the title, and this publication targets marketers, not linguists. You might also have observed that I work for a company that versed in machine learning technology and that there's some computer-y sounding headings a little farther down.

Machine-driven semantic analysis has various real world applications. It helps:

- extract compatible and useful information from huge unstructured data
- find an answer to a question without asking a human
- discover the actual meaning of conversational speech in online posts
- uncover specific meaning of words used in foreign languages disordered with our own

PROPOSED SYSTEM MECHANISM

First , the Sanskrit sentence is taken as input in Devanagari format and changed into ISCII format. Each word is then analyzed using the DFA Tree that is restored by the above block. Following along any path from this DFA tree returns us the root word of the word that we would like to analyze, considering its all features. While evaluating the Sanskrit words in the sentence, we have considered these steps for computation:

1. First, a left-right parsing to distinguish the words in the sentence is accomplished.
2. Second, each word is analyzed against the Sanskrit rules base represented by the DFA trees.

In the primary order given: Each word is examined with the avavya database, next in pronoun, then verb and at last in the noun tree. We did this because, lead ordering is primarily due to the fact pronouns are limited in number compared to the verbs, and verbs are limited compared with large number of nouns that exist in Sanskrit.

ISSN: 2278 – 1323

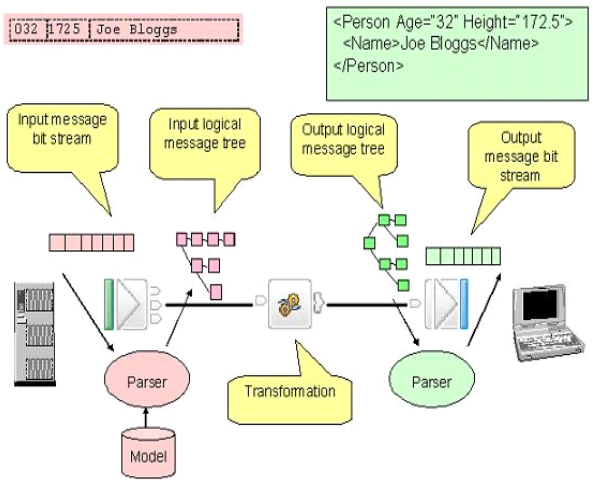

Fig No 2 Parser Flow

TEXT ANALYSIS TECHNIQUES

Computers are very speedy and powerful machines; however, they process texts written by humans in an entirely without thinking, treating them merely as sequences of meaningless symbols. The main objective of language analysis is to obtain an appropriate delegation of text structure and thus make it possible to progress texts based on their content. This is necessary in variety of applications, such as spell- and grammar-checkers, intelligent search engines, text compendia, or dialogue systems. Natural language text can be determined on variety of levels, depending on the actual application setting. Concerned to automatic processing of language data, the analysis level can be distinguished as follows:

- **Morphological Analysis**

Morphological analysis gives a basic comprehension into natural language by studying how to separate and obtain grammatical forms of words emerging through inflection (ie. declension and conjugation). This involves considering a set of tags explaining grammatical categories of the word , most notably, its base form (lemma) and pattern. Automatic analysis of word is done in free text can be used for instance in grammar checker development, and can support corpus tagging, or semi-automatic dictionary assembling. The NLP laboratory has produced a general morphological analyzer for Czech, **HYPERLINK "http://nlp.fi.muni.cz/projekty/ajka/"ajka**, which considers vocabulary of over 6 million word forms. It has further served as a base for a identical analyzer for Slovak, the **fispell** grammar-checker, the **czaccent** converter of ascii text to text with diacritics, and an associated interface for the IM Jabber protocol.

- **Syntactic Analysis**

The goal of syntactic analysis is to conclude whether the given text string is a sentence in the required (natural) language. If it is in given language then, the result of the analysis includes an explanation of the syntactic structure of the sentences. Such formalizations are designed for making computers "recognize" relationships between words (and indirectly between complementing people, things, and actions). Syntactic analysis can be utilized for various purpose like, developing a punctuation resolver, dialogue systems with a natural language interface, or as a basic block in a machine translation system. Czech language is demonstrating rich articulation and free word order and hence requires more grammar rules than other languages. Appropriately, it is one of the languages that are very hard to determine. The NLP laboratory is building the **synt** HYPERLINK "http://nlp.fi.muni.cz/projekty/wwwsynt/" syntactic analyzer. Considering the tests performed on large corpora, the performance of **synt** reaches the recall of 92 % and precision of 84 %. For educational uses, we have a simpler version of **syntactic analyzer HYPERLINK "http://nlp.fi.muni.cz/projekty/zuzana/"**. This analyzer is capable of visualizing several types of derivation trees.

Conclusion

There are three parts in our parser. First part takes care of the morphology. For every word in the sentence taken as input, a dictionary or a lexicon is referred, and associated grammatical information is retrieved. Morphological analyzer is judge considering its speed. We have made a linguistic generalization and deteriorations are given the form of DFA, which will increase the speed of parser. Second part of the parser deals with making Local Word Groups.

References

- Pawan Goyal “Converting Phrase Structures to Dependency Structures in Sanskrit” http://creativecommons.org/licenses/by/4.0/ 2012
- Amba Kulkarni “Designing a Constraint Based Parser for Sanskrit” Indian Linguistic Studies, Festschrift in Honor of George Cardona, Ed. Deshpande, Hook, Motilal Banarasidass, Delhi, 2002.
- Amba Kulkarni “Discourse Analysis of Sanskrit texts” Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA), pages 1–16, COLING 2012, Mumbai, December 2012.
- Shashank Saxena” Sanskrit as a Programming Language and Natural Language Processing” Global Journal of Management and Business Studies. ISSN 2248-9878 Volume 3, Number 10 (2013), pp. 1135-1142
- Md. Nawab Yousuf Ali “UNL Based Bangla Natural Text Conversion Predicate Preserving Parser Approach”Department of Computer Science and Engineering, East West University Dhaka, Bangladesh.
- Rule Based Machine Translation from English to Malayam Rajan, R; Sivan, R; Ravindran, R; Sornan, K.P; Advances in Computing, Control & Telecommunication. Technologies, 2009. ACT’09,International Conference on Digital Object Identifier.
- B. Collins, “Example-Based Machine Translation: An Adaptation- Guided Retrieval Approach”,PhD thesis,Trinity College ,Dublin,1998.
- Cicekli and H. A. Guvenir, “Learning Translation Rules From A Bilingual Corpus”,NeMLaP-2: Proceedings of the Second International Conference on New Methods in Language Processing,Ankara, Turkey.
- L. Cranias, H. Popageorgiou and S. Piperidis, “A Matching Technique in Example-Based Machine Translation “, in Coling. pp. 100-104,1994.
- Gruber T.R, “A Translation Approach to Portable Ontology Specification ,”Knowledge Acquisition, 5(2): 199-220,1993.
- Joakim Nivre and Mario Scholz, “Deterministic Dependency Parsing of English Text,School of Mathematics and System Engineering, Vaxjo University,Sweden,2003.
- Michel Galley and Christopher D. Manning , “Accurate Non-Hierarchical Phrase Based Translation”, Computer Science Department, Stanford University, Stanford

ISSN: 2278 – 1323