International Journal of Engineering Research and Development
e-ISSN: 2278-067X, **p**-ISSN: 2278-800X, www.ijerd.com
Volume 6, Issue 7 (April 2013), PP.88-91

Natural Language Processing Tasks for Marathi Language

Pratiksha Gawade, Deepika Madhavi, Jayshree Gaikwad, Sharvari Jadhav
I. T. Department, Padmabhushan Vasantdada Patil Pratishthan's College Of Engineering.
Sion (East), Mumbai-400 022

Abstract:- Natural language processing is process to handle the linguistic languages to understand the language in to more detailed format. It will providing us will the complete grammatical structure of the word of linguistic language into English. In the present system such as dictionary we only have the word with its meaning but it does provide us grammar. In our proposed system we are presenting the parsing tree which shows the complete grammatical structure .It will show all the tags chunks etc which help to understand the language in more detailed way.

Keywords:- NLP, Morphological analyser, inflection rules, parse tree.

I. INTRODUCTION

1.1Major tasks in NLP

The following is a list of some of the most commonly researched tasks in NLP. Note that some of these tasks have direct real-world applications, while others more commonly serve as subtasks that are used to aid in solving larger tasks. What distinguishes these tasks from other potential and actual NLP tasks is not only the volume of research devoted to them but the fact that for each one there is typically a well-defined problem setting, a standard metric for evaluating the task, standard corpora on which the task can be evaluated, and competitions devoted to the specific task.

1.1.1:Morphological segmentation:

Separate words into individual morphemes and identify the class of the morphemes. The difficulty of this task depends greatly on the complexity of the morphology (i.e. the structure of words) of the language being considered. English has fairly simple morphology, especially inflectional morphology, and thus it is often possible to ignore this task entirely and simply model all possible forms of a word (e.g. "open, opens, opened, opening") as separate words. In languages such as Turkish, however, such an approach is not possible, as each dictionary entry has thousands of possible word forms.

1.1.2Natural language generation:
Convert information from computer databases into readable human language.

1.1.3Part-of-speech tagging:

Given a sentence, determine the part of speech for each word. Many words, especially common ones, can serve as multiple parts of speech. For example, "book" can be a noun ("the book on the table") or verb ("to book a flight"); "set" can be a noun, verb or adjective; and "out" can be any of at least five different parts of speech. Note that some languages have more such ambiguity than others. Languages with little inflectional morphology, such as English are particularly prone to such ambiguity. Chinese is prone to such ambiguity because it is a tonal language during verbalization. Such inflection is not readily conveyed via the entities employed within the orthography to convey intended meaning

.

Parsing: Determine the parse tree (grammatical analysis) of a given sentence. The grammar for natural languages is ambiguous and typical sentences have multiple possible analyses. In fact, perhaps surprisingly, for a typical sentence there may be thousands of potential parses (most of which will seem completely nonsensical to a human).

1.2Inflection rules:

An inflection[7] expresses one or more grammatical categories with an explicitly stated prefix, suffix, or infix, or another internal modification such as a vowel change. For example, the Latin ducam, meaning "I will lead", includes an explicit suffix, -am, expressing person (first), number (singular), and tense (future). The use of this

suffix is an inflection. In contrast, in the English clause "I will lead", the word "lead" is not inflected for any of person, number, or tense; it is simply the bare form of a verb .The inflected form of a word often contains both a free morpheme (a unit of meaning which can stand by itself as a word), and a bound morpheme (a unit of meaning which cannot stand alone as a word). For example, the English word "boys" is a noun that is inflected for number, specifically to express the plural; the content morpheme "boy" is unbound because it could stand alone as a word, while the suffix "s" is bound because it cannot stand alone as a word. These two morphemes together form the inflected word "boy".

Words that are never subjected to inflection are said to be invariant; for example, "should" is an invariant item: it never takes a suffix or changes form to signify a different grammatical category. Its category can only be determined by its context.

Requiring the inflections of more than one word in a sentence to be compatible according to the rules of the language is known as concord or agreement. For example, in "the boy dance", "boy" is a singular noun, so "dance" is constrained in the present tense to use the third person singular suffix "s".

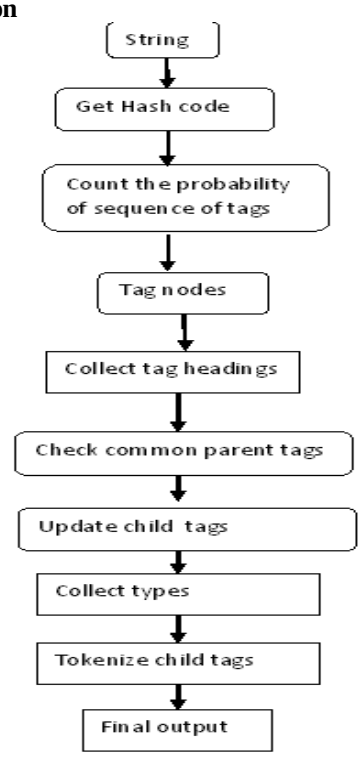
II. RELATED WORK

Morphological analyzer:

Morphological analyzer[1] for Marathi is the program design for developing the morphemes of the given word. The system will describe the grammatical structure of the given translated word. A morphological analyser forms the foundation for many NLP applications of Indian Languages. In this paper, we propose and evaluate the morphological analyser for Marathi, an inflectional language. The morphological analyser exploits the efficiency and flexibility offered by finite state machines in modelling the morph tactics while using the well devised system of paradigms to handle the stem alternations intelligently by exploiting the regularity in inflectional forms. We plug the morphological analyser with statistical pos tagger and chunker to see its impact on their performance so as to confirm its usability as a foundation for NLP applications.

III. ALGORITHM

A:Algorithm for parse tree generation

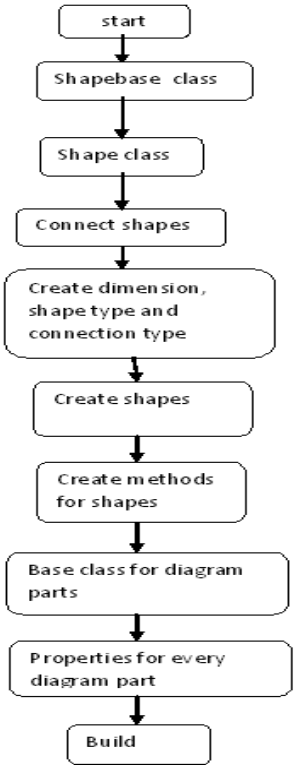


Steps

1) Get the string.
2) Get the Hash code of the string.
3) Count the probability of sequence of tags.
4) Get the Tag nodes.
5) Gets the array of Tag Headings.
6) Get the common parent tags of child tags.
7) Update the child tags of the parents.

8) Get the type of headings.
9) Get the token values from child tags.
10) Show the final output.

B:Algorithm for tree control



IV. EXPERIMENTAL RESULT

In this paper, we present the morphological analyser for Marathi which is official language of the state of Maharashtra (India). Marathi is the language spoken by the native people of Maharashtra. Marathi belongs to the group of Indo-Aryan languages which are a part of the larger of group of Indo-European languages, all of which can be traced back to a common root. Among the Indo-Aryan languages, Marathi is the southern-most language. All of the Indo-Aryan languages originated from Sanskrit.
The morphological analyser takes the Marathi input then converts it into the English language. We gets its type, lexicon, morphemes.

Example 1:
Consider an example for analyser
Lapavun thevane = Lapavun+thev+ne(thevane) ---→ to put in secret place
Tree is as follows

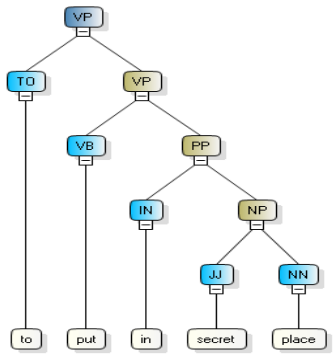


Fig1: Example1

VB : Verb, base form
VP: Verb phrase
PP: Post position
JJ: Adjective
IN: Preposition/subordinate
NP: Noun phrase
NN :Noun, singular or mass

Example 2:
samanjas asne = saman+jas(samanjas)+asne ---> having good sense
Parse tree

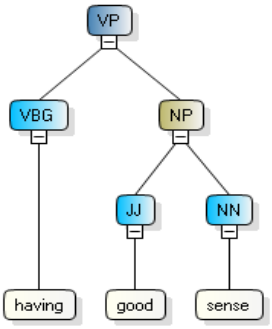


Fig2: Example 2

VP: Verb phrase
VBG: Verb, gerund/present participle
JJ: Adjective
NN: Noun singular
NP: Noun phrase

V. CONCLUSION

We presented a high accuracy morphological analyzer for Marathi that exploits the regularity in the inflectional paradigms while employing he Finite State Systems for modelling the language in an elegant way. We gave detailed description of the morphological phenomena present in Marathi. The classification of postpositions and the development of morph tactic FSA is one of the important contributions since Marathi has complex morph tactics. As a next step the morphological analyzer can be further extended to handle the derivation morphology and compound words. We are presenting the etymological structure of the word which defines the origin of that word. We plan to develop a hybrid system using methods to handle unknown words and to improve the overall accuracy of the system. In the meantime, more analysis will be added to the system to cover aspects which might have eluded us so far. Even consideration of sentence also taken into account to develop the further models using this kind of analysis.

REFERENCES

[1]. "An improvised Morphological Analyzer for Tamil:A case of implementing the open source platformApertium". Parameswari K. Unpublished M.Phil. Thesis. Hyderabad:University of Hyderabad. 2009.

[2]. " Design and Implementation of a Morphology-based Spellchecker forMarathi, an Indian Language". Dixit, Veena, Satish Dethe, and Rushikesh K. Joshi. 2006.
[3]. "James Allen. Natural Language Understanding". Pearson Education, Singapur, second edition, 2004.
[4]. "Natural Language Processing: A Paninian Perspective". Bharati, Akshar, Vineet Chaitanya, and Rajeev Sanghal 1995.
[5]. " Natural Language Processing – A Paninian Perspective". R. S. Akshar Bharati and V. Chaitanya,1995.
[6]. "Two-level Morphology: a general computational model for word-form recognition and production". Koskenniemi, Kimmo 1983.
[7]. "A Paradigm-Based Finite State Morphological Analyzer for Marathi", Pushpak Bhattacharyya.
[8]. "Natural Language Processing". Utpal Sharma. Department of Computer Science and Information
[9]. Technology, Tezpur University, Tezpur-784028, Assam ,India.