Syntactic Analyzer using Morphological Process for a Given Text in Natural Language for Sense Disambiguation

Gauri Dhopavkar
Research Scholar, CSE dept., GHRCE
Faculty, CT dept., YCCE
Nagpur, INDIA.
gaurid.manoj@gmail.com

Manali Kshirsagar
VP(Academy), ADCC Infocad Pvt.Ltd.
Professor, CT dept., YCCE
Nagpur, INDIA
manali_kshirsagar@yahoo.com

Abstract— **In this paper, we present the work related to syntactic annotation of Marathi text using Ruled–based approach which is very essential in Sense Disambiguation of a natural language text. We have implemented a system for generating and applying natural language patterns to overcome the sense ambiguity problem. We manipulate the grammatical structure of sentence to give the correct output for Marathi Language. Some patterns describe the main constituents in the sentence and some, the local context of the each syntactic function. We present the results of our work and discuss possible refinements of the method from a linguistic point of view. This paper also discusses the morphological analysis method used for Marathi Language. Morphological Analyzer is designed to find a root word of a given word and can be used in Gender Recognition as well during the syntactic analysis for a given sentence.**

Keywords—Morphological Analysis; Syntactic Analysis; Named Entity Recognition; Machine Translation

I. INTRODUCTION

In a Natural Language, Syntactic analysis must exploit the results of morphological analysis to build a structural description of the sentence. The goal of this process, called parsing, is to convert the list of words that forms the sentence into a structure that defines the units that are represented by that list. The important thing here is that a sentence is converted into a hierarchical structure and that the structure corresponds to meaning units when semantic analysis is performed. The syntactic analysis of natural language is performed in many levels of natural language understanding task. Therefore, syntactic analysis is very important. [1]

In Natural language tasks like Machine Translation, the incorrectness in Syntax may lead to severe ambiguity resulting into inaccurate translations. As efficient parsers are not available for Marathi language, it is difficult to detect and correct the nuisance in the syntax of the sentence. Hence it is necessary to use some technique with the help of which we can manipulate the grammatical structure of sentence to give the correct output. Hence we are use the "Rule Based Approach". With the help of this method; we are able to detect the morphological features of the sentence along with the tense categories. It becomes easy to shuffle the structure of the sentence to give the correct structural format if the inputted sentence is syntactically incorrect. We present our work focusing on Syntactic Analysis. The paper is divided into sections. Section I is Introduction which provides introduction about our work and its need. Section II focuses on Literature Survey (review). Section III explains the method used in our work to deal with sense ambiguity. Section IV is Conclusion which provides conclusion of the work presented in the paper and Future scope.

II. LITERATURE SURVEY

The literature review of syntactic analysis shows that development of morphological analysis and generation as well as natural language parsing work has been successfully done for languages like English, Chinese, Arabic and European languages using various approaches for last few years. Literature shows that there are a very few number of attempts for Indian languages and still many are an ongoing processes.

Comparing with foreign languages, the work carried out for Indian languages is less but it is significant. This section provides survey of various developments contributed towards natural language parsing in Indian languages.

Authors of [1] have developed a morphological analyzer system for Punjabi using Lexicon based morphological analyzer.

In this paper authors have identified drawbacks of other methods and also focussed on issues related with corpus in Punjabi language.

Natural Language constructs for Venpa class was implemented by Bala Sundara Raman L, Ishwar S, and Sanjeeth Kumar Ravindranath in 2003 [5]. This was designed for Tamil Poetry using Context Free Grammar. They used Push-Down Automata parser to parse the CFG(context free grammar) in the proposed system. Authors claimed it to be very efficient system.

978-1-4799-4236-7/14/$31.00 ©2014 IEEE

Authors Antony P J, Nandini J. Warrier and Dr Soman K P have developed statistical syntactic parsers for two South Dravidian languages. They developed these parsers for Kannada and Malayalam languages in 2010 [3][4]. Authors used Penn Treebank structure which is the well known grammar formalism to create the corpus for statistical syntactic parsers. The parsing system was trained using corpus (Treebank based) which consisted of around 1,000 sentences. These sentences were carefully constructed. Annotation of Corpus was done earlier. The developers used their own POS tagger to assign suitable tags to each and every word in the training and test sentences. As per authors' claim, the proposed syntactic parser was implemented using supervised machine learning and probabilistic context free grammars (PCFG) approaches. Training, testing and evaluation were done by support vector method (SVM) algorithms. Experiment shows that the performance of the proposed system is significantly good and has very competitive accuracy. [3]

Akshar Bharati and Rajeev Sangal described a grammar formalism called the 'Paninian Grammar Framework'. The authors claim that this formalism can be successfully applied to all free word Indian languages. They have described a constraint based parser for the framework. Paninian framework uses the notion of karaka relations between verbs and nouns in a sentence. [6]

[8] Authors K. Saravanan, Rajani Parthasarathi, T. V. Geetha detail their work regarding the Tamil parser in their paper. The parser identifies syntactic constituents of a sentence and represents it using a parse tree. The ambiguity present at the morphological and syntactic level make Parsing complex.

In Paper [9], authors describe how they have used hybrid approach to solve the task of dependency parsing. In this work, authors have used grammar driven approach along with controlled statistical strategy for achieving high performance and robustness. Modularity is taken into consideration for dealing with dependency parsing. The task of dependency parsing is achieved using modularity, where specific tasks are broken down into smaller linguistic sub-tasks.

[10] Authors report their contribution in designing a Lexicon parser for performing syntactic and semantics analysis of Devanagari script sentence. Hindi wordnet is also used in this work. Authors follow rule based approach in designing the parser. The parser generates a parse tree which is obtained by using semantic relationship. The parser identifies grammatical meaning of the words in the given Hindi sentence. Authors have used new 150 sentences for testing purpose and accuracy is calculated at different levels.

III. METHODOLOGY IMPLEMENTED

As in Marathi language(Language commonly spoken by people living in State of Maharashtra, a state in country India) the syntax of the sentence plays a very vital role, incorrect syntax can change even the meaning of sentence leading to ambiguity. Hence, we have designed the syntactic analyzer for increasing the efficiency of working with Marathi language tasks. The syntactic analyzer designed uses the "Rule based approach". If the inputted sentence is grammatically correct as per Marathi grammar, then it will show the output as correct sentence along with its tense and morphological features. Otherwise, it will show the output as incorrect sentence and then manipulate the result using rules to correct it according to the language rules of tenses.

The methodology of the system which have implemented is shown in the block diagram (flowchart) below:

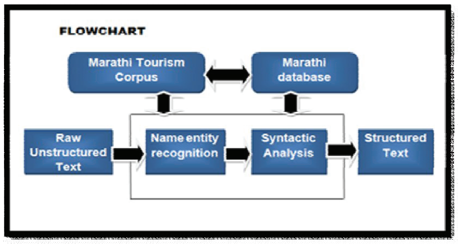


Fig. 1. Block Diagram of the system[7]

First we give the input as raw unstructured text which might contain ambiguities. After this, Named entity recognition is performed. In Named entity recognition the words present in sentence are identified for whether they are Pronouns, objects, thus separating verbs or helping verbs, then using syntactic analyzer the unstructured text gets converted in the structured format (called as structured text).

As we are implementing the Syntactic analyzer for Marathi so we cannot use English parsers. The parser can be used for English but not for Indian languages. In this project we are using the ruled-based approach to design the parser. Here, we check the syntactic structure of a sentence which the user is has input. We first to categorize the sentences in some specific format like the Subject, Object and Verb. The syntactic structure of a sentence is checked and then it is determined whether the inputted sentence is syntactically correct or not. If the sentence is syntactically incorrect then using syntactic analyzer the incorrect sentence will be converted into a correct sentence. As we use the morphological process we are showing the morphological analysis also. Here we perform the root word and gender analysis.

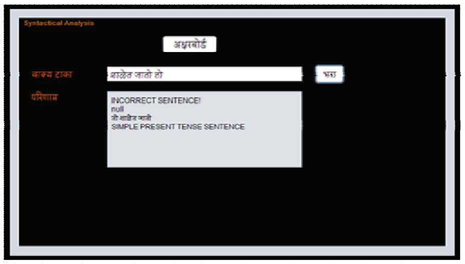
In root word and gender analysis, the analyzer determines the root word, means the word from which a given word is derived. In gender analysis we are determining the gender of the sentence which the user is going to give as the input.

Here implementation of all the tenses is done. Depending on its format, if the sentence is in (SOV) format then output will be displayed as correct sentence plus the tense of the sentence. But if it is not in (SOV) format the output will be displayed as incorrect sentence plus using syntactic analyzer incorrect sentence gets converted into the correct sentence. Here are examples of all the Tenses:

A. Simple Present Tense(For Correct Sentence)

English follows Subject-Verb-Object (SVO) structure while Marathi follows Subject-Object-Verb (SOV) structure [2]. Along with Marathi other Indo-Aryan language like Hindi, also follow the SOV structure. Here in this example, first the format of (SOV) is checked as the given input sentence is in Subject-Object-Verb (SOV) format output gets displayed as correct sentence along with its tense. In addition to this whether the sentence is Singular/Plural and first person, second person or third person also gets displayed. In this example as तो is Singular and it is also third person, so the output is displayed as Singular sentence and Third person (3^{rd} person).

B. Simple Present Tense (For Incorrect Sentence)



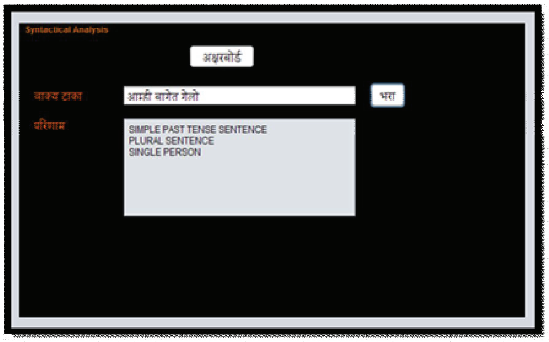
शाळेत जातो तो

INCORRECT SENTENCE!
null
तो शाळेत जातो
SIMPLE PRESENT TENSE SENTENCE

Fig. 2. Result of incorrect sentence (Simple present tense)

First the format of (SOV) is checked. As the given input sentence is not in Subject-Object-Verb (SOV) format output gets displayed as incorrect sentence along with its tense. In addition to this whether the sentence is Singular/Plural is displayed. As the provided input is an incorrect sentence using syntactic analyzer the incorrect sentence gets converted into the correct sentence. Hence correct sentence also gets displayed.

C. Simple Past Tense (For Correct sentence)



आम्ही बागेत गेलो

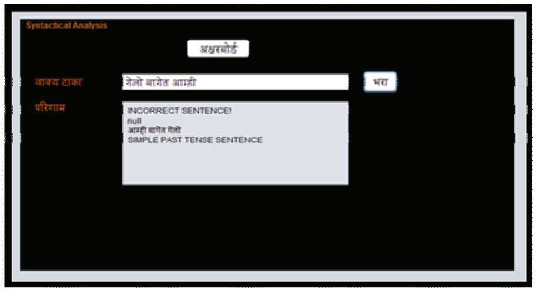
SIMPLE PAST TENSE SENTENCE
PLURAL SENTENCE
SINGLE PERSON

Fig. 3. Result of correct sentence (Simple past tense) [7]

In addition to this whether the sentence is Singular/Plural and first person, second person or third person also gets displayed.

In this example आम्ही is Plural and it is also first person. So the output is displayed as Plural sentence and Single person (1^{st} person).

D. Simple Past Tense(For Incorrect sentences)



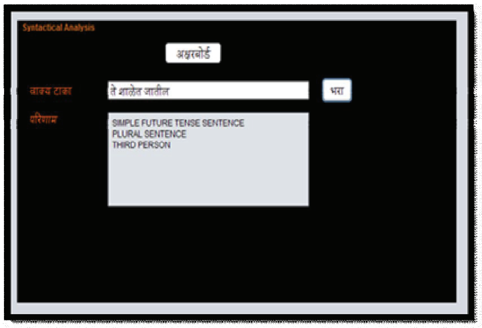
गेलो बागेत आम्ही

INCORRECT SENTENCE!
null
आम्ही बागेत गेलो
SIMPLE PAST TENSE SENTENCE

Fig.4 Snapshot for Simple Past Tense

In the example shown in the figure, (*went in the garden we*) as the given input sentence is not in Subject-Object-Verb (SOV) format output gets displayed as incorrect sentence along with its tense. As the provided input is a incorrect sentence using syntactic analyzer the incorrect sentence gets converted into the correct sentence. Hence correct sentence also gets displayed.

E. Simple Future Tense(For Correct sentence)



ते शाळेत जातील

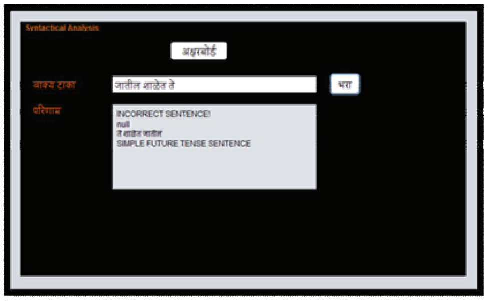
SIMPLE FUTURE TENSE SENTENCE
PLURAL SENTENCE
THIRD PERSON

Fig. 5. Result of correct sentence (Simple future tense)

Here as in this example also, first the format of (SOV) is checked as the given input sentence is in Subject-Object-Verb (SOV) format output gets displayed as correct sentence along with its tense. In addition to this whether the sentence is Singular/Plural and first person, second person or third person also gets displayed. In this example as ते (*They*)is Plural and it is also third person so the output is displayed as Plural sentence and third person (3^{rd} person).

F. Simple Future Tense

Here in this example, first the format of (SOV) is checked. As the given input sentence is not in Subject-Object-Verb (SOV) format output gets displayed as incorrect sentence along with its tense. As the provided input is a incorrect sentence using syntactic analyzer the incorrect sentence gets converted into the correct sentence. Hence correct sentence also gets displayed.



जातील शाळेत ते

INCORRECT SENTENCE!
null
ते शाळेत जातील
SIMPLE FUTURE TENSE SENTENCE

Fig. 6. Example of Incorrect sentence for Simple Future Tense

IV. CONCLUSION AND FUTURE SCOPE

As detailed in the previous sections, we have implemented the syntactic analyzer for Marathi language, with the help of "Rule Based Approach". Some users of Marathi may not be aware of the correct syntactic structure of the sentence as they are new in Marathi language usage. Understanding of Marathi syntactic structure is important for better solution of any NLP application like word sense disambiguation. The analysis process starts with word level analysis like, morphological analysis. Morphological analysis includes root word analysis and gender analysis, then using this we have implemented syntactic analysis which includes detection of tenses and its correction if needed(in cases of incorrectness) along with providing information about whether the sentence is singular or plural.

Our system is capable of identifying incorrectness in a sentence related to Tenses, SOV order, morphology. For all the incorrect samples the machine provides correct output which is compared with Manual output. The system output exactly matches with human output prediction. The system is checked for sample corpus of around Five Hundred sentences.

However this system allows input of single and simple sentence only. In future it may be updated for accepting complex sentences also which have more probability of ambiguity.

More syntactic patterns may be added in future to the database of this system to get the accurate results of sense disambiguation.

REFERENCES

[1] Gagan Bansal, Satinder Pal Ahuja, Sanjeev Kumar Sharma, Improving Existing Punjabi Morphological Analyzer, Research Cell: An International Journal of Engineering Sciences ISSN: 2229-6913 Issue Dec. 2011, Vol. 5.

[2] Mugdha Bapat,Harshada Gune, Pushpak Bhattacharyya, A Paradigm-Based Finite State Morphological Analyzer for Marathi, Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), pages 26–34,the 23rd International Conference on Computational Linguistics (COLING), Beijing, August 2010.

[3] Antony P J, Nandini. J. Warrier, Dr. Soman K P, Penn Treebank-Based Syntactic Parsers for South Dravidian Languages using a Machine Learning Approach, International Journal of Computer Applications (0975–8887) Volume 7 No.8 , October 2010.

[4] Antony P J and Soman K P, Computational Morphology and Natural Language Parsing for Indian languages: A Literature Survey, International Journal of Computer Science & Engineering Technology (IJCSET), ISSN: 2229-3345, Vol. 3 No. 4, April 2012.

[5] Bala Sundara Raman L, Ishwar S, Sanjeeth Kumar Ravindranath, Context free grammar for Natural Language constructs-An implementation for a Vempa class of Tamil poetry, Tamil Internet 2003, Chennai, Tamilnadu, India

[6] Akshar Bharati and Rajeev Sangal, Parsing free word order languages in the Paninian framework,www.ldc.upenn.edu/acl/P/P93/P93-1015.pdf

[7] Aditi Muley, Manaswi Pajai, Priyanka Manwar, Sonal Pohankar, Gauri Dhopavkar, Syntactic Analyser using Morphological process for a given text in Marathi language, International Journal of Scientific Research in Computer Science Applications and Management Studies, Volume 3, Issue 2 (March 2014).

[8] K. Saravanan, Rajani Parthasarathi, T. V. Geetha, Syntactic Parser for TamilTamil Internet 2003, Chennai, Tamilnadu, India.

[9] Akshar Bharati, Samar Husain, Meher Vijay, Kalyan Deepak, Dipti Misra Sharma and Rajeev Sangal Constraint Based Hybrid Approach to Parsing Indian Languages, 23rd Pacific Asia Conference on Language, Information and Computation, pages 614–621.

[10] Swati Ramteke, Komal Ramteke, Rajesh Dongare, Lexicon Parser for syntactic and semantic analysis of Devnagari sentence using Hindi wordnet, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 4, April 2014.