SEMANTIC ANALYZER FOR MARATHI TEXT

Pallavi Bagul[1], Prachi Mahajan[2], Archana Mishra[3], Medinee Kulkarni[4], Gauri Dhopavkar[5]

[1]Computer Technology, YCCE Nagpur- 441110, Maharashtra, India
[2]Computer Technology, YCCE Nagpur- 441110, Maharashtra, India
[3]Computer Technology, YCCE Nagpur- 441110, Maharashtra, India
[4]Computer Technology, YCCE Nagpur- 441110, Maharashtra, India
[5]Computer Technology, YCCE Nagpur- 441110, Maharashtra, India

Abstract

This paper represents a Semantic Analyzer for checking the semantic correctness of the given input text. We describe our system as the one which analyzes the text by comparing it with the meaning of the words given in the WordNet. The Semantic Analyzer thus developed not only detects and displays semantic errors in the text but it also corrects them.

Keywords: *Part of Speech (POS) Tagger, Morphological Analyzer, Syntactic Analyzer, Semantic Analyzer, Natural Language (NL)*

---***---

1. INTRODUCTION

Natural Language (NL) is a very important and essential tool to represent information. Computer is not able to understand NL. Now when we say understanding is a mental process, it means how human beings recognize objects (mental and physical) and links between them. Since computer does not have a human mentality, so it cannot understand by definition.

NL processing (NLP) is a general problem and to be more specific we can separate it by categories according to increasing level or complexity of such processing:

- Morphology and morphological processing
- Syntax and syntactical processing
- Semantics and semantic processing

Morphology is a sub-discipline of linguistics that studies word structure. During morphological processing we are basically considering words in a text separately and trying to identify morphological classes in which these words belong to. One of the widespread tasks here is lemmatizing or stemming which is used in many web search engines. In this case all morphological variations of a given word (known as word-forms) are collapsed to one lemma or stem [5].

Syntax as part of grammar is a description of how words are grouped and connected to each other in a sentence. Syntax usually entails the transformation of a linear sequence of tokens into a hierarchical syntax tree. A token is akin to an individual word or punctuation mark in a natural language. Main problems on this level are: part of speech tagging (POS tagging), chunking or detecting syntactic categories (verb, noun phrases) and sentence assembling (constructing syntax tree) [8].

Semantics and its understanding as a study of meaning covers most complex tasks like: finding synonyms, word sense disambiguation, constructing question-answering systems, translating from one NL to another, populating base of knowledge. Basically one needs to complete morphological and syntactical analysis before trying to solve any semantic problem. Formalization of NL leads us to solutions of all these problems [3].

2. APPROACHES

Broadly speaking there are two basic ways in which Semantic Analysis can be carried. They are classified as:

2.1 Supervised Semantic Analysis

The supervised models require a pre-annotated corpus which is used for training the application so as to learn information about the various words, their tags, frequencies, rule sets etc. The performance of the model generally increases when we increase the size of the corpus. Such annotated resources are scarce and expensive to create, motivating the need for unsupervised or semi-supervised techniques.

2.2 Unsupervised Semantic Analysis:

Unsupervised approaches do not depend on pre-annotated corpus instead it relies on distributional similarity of contexts to decide on semantic relatedness of terms, but this information may be sparse and not reliable always [13].

However, unsupervised methods have their own challenges they are not always able to discover semantic equivalences of lexical entries or logical forms or, on the contrary, cluster semantically different or even opposite expressions.

2.3 Semi- Supervised Semantic Analysis:

In this case, groups of unannotated texts with overlapping and non-contradictory semantics provide a valuable source of information. This form of weak supervision helps to discover implicit clustering of lexical entries and predicates, which presents a challenge for purely unsupervised techniques [12].

3. LITERATURE SURVEY

Considerable amount of work has already been done in the field of Semantic analysis for English text. Different approaches along with modifications have been tried and implemented. However, if we look at the same scenario for South-Asian languages such as Marathi and Hindi, we find out that not much work has been done. The main reason for this is the unavailability of a considerable amount of annotated corpora of sound quality, and very high level of ambiguity in those languages. In the following sections, we describe some POS tagging models, Morphological analysis model, syntactic analysis models and semantic analysis model that have been implemented for English and other languages along with their performances.

In the year 2011, Akira Shimazu, Syozo Naito, and Hirosato Nomura of Musashino Electrical Communication Laboratory, Japan developed a Japanese Language Semantic Analyzer that was based on an extended case frame model. The case frame model consists of a relatively large collection of case relations, modalities and conjunctive relations. The analyzer uses a frame type knowledge base for analyzing. It also utilizes plausibility scores for dealing with ambiguities and local scene frames for the prediction of omitted case elements [3].

Ivan Titov and Mikhail Kozhevnikov developed a Bootstrapping Semantic Analyzers from Non-Contradictory Texts in the year 2010. They urged that groups of unannotated texts with overlapping and non-contradictory semantics provide a valuable source of information. This form of weak supervision helps to discover implicit clustering of lexical entries and predicates, which presents a challenge for purely unsupervised techniques [14]. They considered the generative semantics-text correspondence model and demonstrate that exploiting the noncontradiction relation between texts leads to substantial improvements over natural baselines on a problem of analyzing human-written weather forecasts.

A Semantic Analyzer for aiding emotion recognition in Chinese Language is developed by Jiajun Yan, David B. Bracewell, Fuji Ren and Shingo Kuroiwa in the year 2006. The analyzer developed, uses a decision tree to assign semantic dependency relations between headwords and modifiers. It is able to achieve an accuracy of 83.5%. The semantic information is combined with rules for Chinese verbs containing emotion to describe the emotion of the people in the sentence. The rules give information on how to assign emotion to agents, receivers, etc. depending on the verb in the sentence [4].

Semantic analysis is also used for retrieving data from text and applying that to ER modeling. Semantic analysis involves a process whereby meaning representations are created and assigned to linguistic inputs. There are some semantic roles defined which are helpful in interpreting possible elements of the ER model. These semantic roles may also indicate the types of entities from the given text. After recognizing the meaning the particular roles may be defined to identify the elements of ER modeling [9].

Semantic analysis of text and speech by Anssi Klapuri in 2007 states various approaches on semantic analysis: i) statistical approach ii) information retrieval iii) domain knowledge driven analysis [10].

4. METHODOLOGY

4.1 Block Diagram

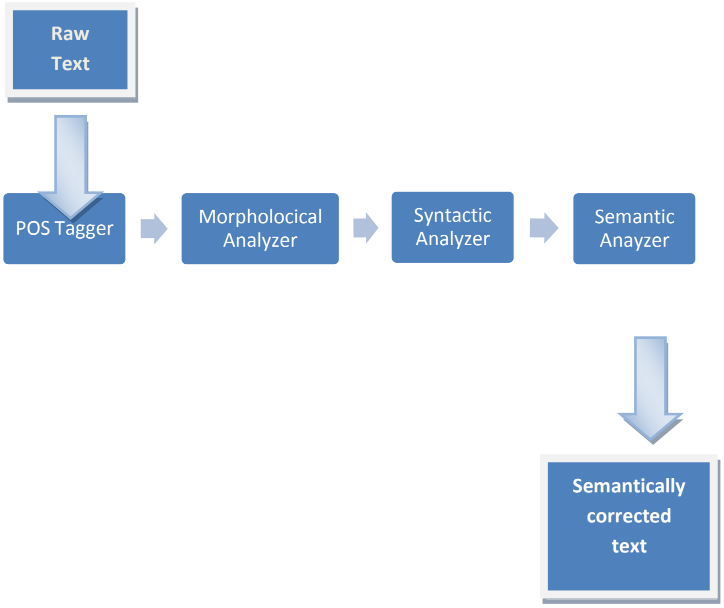

Fig. 1 Block Diagram

4.2 POS Tagger

The input string (Raw Text) is tokenized and a word net is used for detecting the part of speech of each token in the sentence. The wordnet [15] stores the part of speech of each word using 4 bits where 1000 signifies noun, 0100 signifies Adjective, 0010 signifies Adverb and 0001 signifies a verb. We also have combinations like 1100 which means that the designated word can be used both as a noun and as an Adjective, we also have is 0110 which means that the designated word can be used both as an Adjective and as an Adverb.

This ambiguity is resolved using Marathi grammar rules. If we have 0110 ambiguity then if the next token is a noun or an adjective then the ambiguous word becomes an adjective. If the next token is a verb then the ambiguous word becomes an adverb. If we have 1100 ambiguity then if the next token is a noun then the ambiguous word becomes an adjective otherwise it becomes an adverb.

4.3 Morphological Analyzer

The POS Tagger's output is then passed to the Morphological Analyzer which detects the root-word along with the gender and the tense. The first step in this module is to find the root word of each token in the given sentence this is done with the help of the word net.

Using the WordNet the gender of each token in the sentence is detected. The word net is trained using a tourism related corpus. The subject, object and verb in the input sentence is detected. For better accuracy the gender of the sentence is detected with respect to the gender of the subject.

The tense of the given input sentence is detected using the Marathi grammar rules.

4.4 Syntactic Analyzer

Morphological Analyzer's output is used by the Syntactic Analyzer to detect whether the output is syntactically correct or not. It is detected whether the subject, object and verb are placed at proper positions in the sentence. The sentence structure should be in accordance to the Marathi grammar rules where subject comes in first position followed by the object and the verb. If it is not in proper order we change the original order.

Finally we check whether the tense assigned to the sentence is in accordance with the sentence structure if not then the sentence structure is modified such that it retains its meaning and has the correct tense associated with it.

4.5 Semantic Analyzer

The output of the Syntactic Analyzer is used by the Semantic Analyzer to check the semantic correctness of the input text and also correct it if it is found to be incorrect. All the tokens in the input sentence must semantically support each other. This means that if two tokens in the sentence are mismatching then it is not semantically correct and needs to be corrected.

If the tokens in the sentence are antinomies of each other and they are being used in the same context then they need to be corrected as the sentence is not semantically correct. For this a list of all the synonyms and antonyms has to be maintained using the tourism corpus. Finally, the semantically corrected text is presented as output to the user.

5. RESULTS

The output of Semantic Analyzer cannot be obtained before completing POS Tagging, Morphological Analysis and Syntactic Analysis.

5.1 POS Tagger

Input sentence:

लाल फुल निळा आहे .

By this phase, each token i.e. word is assigned its suitable part of speech:

लाल-**Adjective** फुल-**Noun** निळा-**Adjective** आहे-**Verb**

5.2 Morphological Analyzer

In this phase, root word for each token is derived and also gender of the whole sentence is found out.

लाल- लाल, फुल- फुल, निळा- निळा, आहे-असणे.
लाल फुल निळा आहे **.- Neuter sentence**

5.3 Syntactic Analyzer

In this phase, the sentence is checked if it is syntactically correct or not. If not, it is corrected.

लाल फुल निळा आहे **.- Syntactically incorrect sentence**

Corrected sentence: लाल फुल निळे आहे

5.4 Semantic Analyzer

In this phase, the sentence is checked if it is semantically correct or not. If not, it is corrected.

लाल फुल निळे आहे**.- Semantically incorrect sentence**

Corrected sentence: लाल फुल निळे नाही आहे

6. CONCLUSIONS

This paper presents a Semantic Analyzer for Marathi Language which undergoes the phases: POS tagger, Morphological Analyzer, Syntactic Analyzer and Semantic Analyzer respectively. POS tagger used in this work follows rule based approach. This work allows single sentence to undergo semantic analysis.

REFERENCES

[1] Adwait Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," In Proceedings of the Conference on Empirical Methods in Natural Language Processing, University of Pennsylvania, 1996, pp. 133--142.

[2] M. Deroualt and B. Merialdo, "Natural Language Modeling For Phoneme-To-Text Transposition", IEEE transactions on Pattern Analysis and Machine Intelligence, 1986.

[3] Akira Shimazu, Syozo Naito, and Hirosato Nomura, "Japanese Language Semantic Analyzer based on an Extended Case Frame Model", 3-9-11.

[4] Jiajun Yan, David B. Bracewell, Fuji Ren and Shingo Kuroiwa,"A Semantic Analyzer for aiding emotion recognition in Chinese", 2006.

[5] Koskenniemi .K, 'Two –Level Morphology: A general Computational," Model for Word Recognition and Production, University of Helsinki, Helsinki, 1983.

[6] Vishal Goyal, Gurpreet Singh Lehal, "Hindi Morphological Analyzer and Generator," IEEE Computer Society Press, pp. 1156–1159, 2008.

[7] Niraj Aswani, Robert Gaizauskas, "Developing Morphological Analysers for South Asian Languages: Experimenting with the Hindi and Gujarati Languages," In Proceedings of LREC, 2010.

[8] J. P. Thorne, P. Bratley and H. Dewar, "The Syntactic Analysis of English by Machine," International Conference on NLP, 1996.

[9] N.Omar, P.Hanna, P.Mc Kevitt, "Semantic Analysis in the Automation of ER Modelling through Natural Language Processing," Computing & Informatics, 2006. ICOCI '06. International Conference, June 2006.

[10] Anssi Klapuri, "Semantic analysis of text and speech," SGN-9206 Signal processing graduate seminar II, Fall , 2007.

[11] Debbarma, K. Patra, B. G. Debbarma, S. Kumari and Purkayastha, "Morphological analysis of Kokborok for universal networking language dictionary," 1st International Conference on Recent Advances in Information Technology (RAIT), pages 474-477, IEEE., 2012

[12] Sugatu Basu, Arindam Banjeree, and Raymond Mooney, "Active semi-supervision for pairwise constrained clustering," In Proc. of the SIAM International Conference on Data Mining (SDM), pages 333–344, 2004.

[13] Hoifung Poon and Pedro Domingos, "Unsupervised semantic parsing," In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-09), 2009.

[14] Ivan Titov Mikhail Kozhevnikov, "Bootstrapping Semantic Analyzers from Non-Contradictory Texts," Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 958–967, 11-16 July 2010.

[15] http://www.cfilt.iitb.ac.in/wordnet/webmwn/