International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 11 | Nov 2018 www.irjet.net p-ISSN: 2395-0072

Survey on Grammar Checking and Correction using Deep Learning for Indian Languages

Neethu S Kumar[1], Supriya L P[2]

[1]MTech, Dept. of Computer Science & Engineering, Sree Buddha College of Engineering, Pathanamthitta, Kerala
[2]Assistant Professor, Dept. of computer science & Engineering, Sree Buddha College of Engineering, Pathanamthitta, Kerala

---

Abstract - A grammar checker is one of the basic Natural Language Processing tools for any language. The grammar checker is widely used for detecting and correcting the sentence during a writing process. There are different kinds of grammar checkers. This paper describes a survey on grammar checker using deep learning for Indian languages. Grammar checking is a fundamental task for the writing process. The grammar consists of many rules including past, present, and future. There are different grammar checker for different languages which aims to improve the accuracy for minimum error. This survey concludes with different features of existing grammar checking.

Key Words: Natural Language Processing, Grammar, Grammar checker, Rule-based, Statistical, Hybrid

1. INTRODUCTION

Language is a communication between human beings. Human natural language can be defined as an interchangeability process between human beings. Grammar is elements in language and it contains sets of rules. Words are the basic grammatical units and these grammatical units combine together to form sentences. These sentences are formed by using some grammar rules. Grammar is a set of rules and these rules are used to form sentences. There are many grammatical errors occurring during the writing process.

One of the main objectives of communication is to share information. This information can be defined in written-form or vocal-form. The most important in information content form is the validity of sentences in the language. Morphemes, phonemes, words, phrases, clauses, sentences, vocabulary and grammar are the blocks of language. All valid sentences of a language must follow the rules of that language. A Sentence is the combination of different words. Sentences with various types of errors are written by language learners of different backgrounds. Sentences can be classified into mainly three. First, simple sentences, which is a collection of one or more arguments. This sentence contains clause and mostly verb root and does not contain question words and negation. Second, complex sentences, which contain two clauses, having interdependence between main and dependent or subordinate clause. Third, compound sentences, which contain multiple clauses.

Natural Language Processing is the one the subfield of artificial intelligence, which is the interaction between the computer and human languages. Most of the natural language processing based on handwritten rules. Grammar checking is one of the most common technology of natural language processing. There are many grammar checkers are used for different languages. The Grammar checker is a program which is used to check whether the sentence is grammatically correct or not. Many different types of grammar checker based on different approaches. They are Rule-based checking, statistics-based checking and hybrid checking. Most of the existing grammar checking are style checking, checking uncommon words and complicated sentence structure.

1.1 Statistical Grammar Checker

In statistical grammar checker, which use an annotated corpus. The annotated corpus is maintained from different journals, magazines or documents. It ensures that the correctness of sentences by checking the input sentences with corpus. Here, there are mainly two ways to check the input sentence. First input text is directly checked with corpus and it check whether the sentence is matched with input text and it is tagged as grammatically errors otherwise checked the sentence is correct or incorrect. The second way is, the maintained corpus are generating some rules and the input sentence is checked by using these rules. When the corpus is maintained or add new data there is no update for the rules. This approaches has some disadvantage is that it is difficult to find the error in sentence and recognize the error in the system.

1.2 Rule Based Grammar Checking

Most commonly used approaches is rule-based grammar checking. In rule-based grammar checking, the input sentence is checked by rules formed from the corpus. But in statistical approach, rules are manually generated. In the rule-based approach, the rules are easy to configure and also to modify these rules. One of the significant advantages of this approach is to handle the rules by one who does not have programming language and it also provides a detailed error message. The main characteristics of this approach are to handle all features of language and sentences also need to be completed and also it can easily handle the input sentence.

1.1 Hybrid Grammar Checking

The hybrid approach combines both rule-based and statistical grammar checking. It is more robust and also achieves higher efficiency.

2. LITERATURE SURVEY

This section will provide the detailed study of different existing grammar checkers.

2.1 ENGLISH GRAMMAR CHECKER

Many grammar checker are developed including the English language in Microsoft word processor. There is a number of English grammar checker researches. One of these grammar checker research is done by Daniel Naber [1]. The aim of an English grammar checker is to design a grammar and style checker. Both the grammar checker and style checker can be used as standalone and word processor systems. The grammar checker and style checker takes a text as input text and returns a list of possible errors. To detecting and correcting the errors, each word of the sentence is assigned to its part of speech tag (noun, verb, adjective, adverb, and determiner). Many of them have a different part of speech tags depending on their context. If the number of part of speech tag is increased, it is more difficult to find the right tag for a given occurrence of a word. After part of speech tagging, each input sentence is divided into chunks (noun phrases).

The English grammar checker has 54 pre-defined grammar rules, which are a sequence of tokens to be matched. After chunking, the text is matched against all these pre-defined error rules. English grammar checking with dependency parsing is a possible extension for grammar and style checker. The English grammar checker detects 42 errors. But the grammar and style checker identifies different errors in the input sentence when the errors are independent of one another. This type of errors cannot be detected in MS word and the MS word cannot specify more than one error for one sentence.

2.2 AFAN OROMO GRAMMAR CHECKER

Afan Oromo grammar checker is one of the most research is done for language in Ethiopia [2]. This grammar checker is based on a rule-based approach. There mainly 123 different rules are used in this grammar checker to identify the errors occurring in the given sentence. There are mainly five components of Afan Oromo grammar checker. First one is tokenizer, which the input sentence is divided into words. The second one is, part of speech, it assigns each word into a part of speech tag. The third one is stemmer, which accepts the tagged words and provides root and affixes for the tagged word. These three steps used to remove certain types of roots and affixes. The fourth one is grammatical relation finder, which assigns grammatical relations between each word. These words include subject and verb, subject and adjective, main verb and subordinate verb in terms of tense. Based on the rules there re agreement between words. The rules take roots and affixes that are produced by a stammer, which is to check the agreement between words. This grammatical information is presented by the roots and affixes rules. The fifth and final one is suggestion creation, which suggests the corresponding sentence when the error occurs. Afan Oromo grammar checker based on rule-based approach, which tested based on the number of errors are detected and the number of errors correcting by the system.

2.3 PORTUGUESE GRAMMAR CHECKER

One of the developed grammar checkers is for Portuguese languages is COGrOO [3]. It is a Portuguese grammar checker which is based on CETENFOLHA a Brazilian Portuguese morph syntactic annotated corpus. The researcher designed to solve some problem including nominal and verbal agreement, nominal and verbal government and misuse of adverb and adjectives. The Portuguese grammar checker contains some rules set in the system. They are local rules and structural rules. These two error checker check for the local and structural errors. Local rules consist of a sequence of word rules and the structural rules consist of complex rules including nominal and verb agreement, nominal and verbal government, misuse of adverb and adjectives. Words in the sentence are tagged by using the part of speech. And then chuck the tagged words to noun and verbal phrases using the chunker. Finally, the grammatical relation finder identifies the relation between the noun and verbal phrases and their grammatical roles including subject, object, and verb.

2.4PUNJABI GRAMMAR CHECKER

Punjabi grammar checker is the type of grammar checker. Indian and Pakistan is most commonly using the Punjabi language. It is a part of Indo- Aryan family of language. Punjabi grammar checker is discussed in [4]. Punjabi grammar checker is based on a rule-based approach. Most commonly occurring a grammatical error in the Punjabi language is modifier and noun agreement, subject and verb agreement, noun and adjective, an order of modifier of the noun in a noun phrase, the order of verb in a verb phrase. There are some steps to detect the errors. First is tokenization, it is done the pre-processing task on the input sentence and it will perform the morphological analysis in this phase. The part of speech tagger is engaged to disambiguate in the tags using a rule-based approach and the sentence is grouped into phrases based on the phrase chucking rule. And last the grammatical errors to the phrases and the text will be identified and correction suggested by using the grammatical error checking rule.

2.5 RULE-BASED AMHARIC GRAMMAR CHECKER

Amharic grammar checker is a rule-based approach is used, which is used to check whether the given sentence is grammatically correct or not [5]. This rule-based grammar checker contains some defined grammar rules. This checker is developed by using manually constructed rules. In rule-

based Amharic grammar checker accepts the input text and this text contains a subject, object and verb order as an input. The input sentences split the sentences to words. These Amharic subjects do not have a subject marker. The rule-based Amharic grammar checker contains some series of steps to checking the grammar. First identifies the noun as the subject and then noun as an object. And every word in sentence analyzed using the morphological analyzer. There is some grammatical relation inbuilt based on the analysis. These relations will be matched against the rules. The rule-based Amharic grammar checker accepts the sentence as an input and then identifies grammatically correct or not and it shows corresponding suggestion for output.

2.6 SWEDISH GRAMMATIFIX GRAMMAR CHECKER

Grammatifix is one of the projects, which is done by Swedish speakers. It is developed as a Swedish grammar checker [6]. The study of grammatifix starts with different existing grammar checker of different languages. The Swedish grammar checker collecting for error types and discovery of new types. Error type is based on the classification of defined error types. Grammatifix checks noun phrase internal agreement, verb chain consistency. In Swedish grammar checker, to detecting the error by using different error technologies. To the detection of syntactic errors, using a constraint grammar formalism. To detect punctuation and number formatting convention violations using the regular expressing based technique. Grammatifix is commonly referred to as an error treatment method and it does not detect compound words mistakenly.

2.7 NEPALI GRAMMAR CHECKER

The Nepali grammar checker is one of the grammar checker developed to the Nepali language. The detailed explanation of Nepali grammar checker in [7]. This grammar checker is based on a rule-based approach. In Nepali grammar checker, the input sentence tokenized into words and these words is for part of speech tagged through the morphological module. The next step is the part of speech tagger tags untagged tokenized words. Then the chunker and parser identify the phrases from the part of spoken words. And the chunker and parser require production rules and part of speech tags of the input text, it will return chunks and phrases. These phrases are assigned to grammatical roles like a subject, object, and verb. Finally, the syntax is checked and the Nepali grammar checker deals with simple sentences.

2.8 ICELANDIC GRAMMAR CHECKER

The Icelandic grammar checker is based rule-based approach [8]. In Icelandic grammar checker, first input sentence passes through parsing for part of speech tagging and syntactic analysis. After part of speech tagging, the system is finding the process to each rule. If the system finds some ways, the input text is not in accordance with relevant rules, an error is generated. The system only detects only errors and correction suggestions. But it does not detect stylistic errors.

2.9 HINDI GRAMMAR CHECKER

Hindi grammar checker is one the grammar checker in Indian languages. L. Bopche, G. Dhopavkar, and M. Kshirsagar explain the methods for Hindi grammar checker [9]. This grammar checker consists of the lexicon for morphological analysis and rule-based system. First, the input sentence passes through all processes such as tokenization, analysis, and part of speech tagging. The part of speech tags contains a set of rules. This grammar checker checks the sentence and detects the errors. It does not provide any suggestion for the errors.

3. CONCLUSION

In this literature survey, different approaches, methodologies are reviewed with different concepts. The main aim of this literature survey is to study the features and methods used for different existing grammar checkers. There are many different types of grammar checker for different languages. Each system has different methods and different approaches. This survey concludes with the study of the features of the grammar checker. And also study the most commonly occurring grammatical errors. There are many grammar checker existing in foreign languages but in Indian languages a limited number of the grammar checker. Hence our future research work aims to develop grammar checker for Indian languages.

REFERENCES

[1] Daniel Naber. “A Rule-Based Style And Grammar Checker”. Diplomarbeit. Technische Fakultät Bielefeld, 2003.

[2] Tesfaye, Debela. “A Rule-Based Afan Oromo Grammar Checker”. Jimma Institute of Technology. Ethiopia: Vol. 2, No. 8, 2011.

[3] Kinoshita, Jorge; Nascimento, Laнs do; Dantas ,Carlos Eduardo. ”CoGrOO: a BrazilianPortuguese Grammar Checker based on the CETENFOLHA Corpus”. Universidade da Sro Paulo (USP), Escola Politйcnica. 2003.

[4] Singh, Mandeep; Singh, Gurpreet; Sharma, Shiv. “A Punjabi Grammar Checker”. Punjabi University. 2nd international conference of computational linguistics: Demonstration paper. 2008. pp. 149 – 132.

[5] Aynadis Temesgen Gebru, ‘Design and development of Amharic Grammar Checker’, 2013.

[6] Domeij, Rickard; Knutsson, Ola; Carlberger, Johan; Kann, Viggo. “Granska: An efficient hybrid system for Swedish grammar checking”. Proceedings of the 12th Nordic conference in computational linguistic, Nodalida- 99. 2000.

[7] “A prototype of a grammar checker for Icelandic”, available at www.ru.is/~hrafn/students/BScThesis_Prototype_Icelandic_GrammarChecker.pdf .

[8] Bal Krishna Bal, Prajol Shrestha, "Architectural and System Design of the Nepali Grammar Checker", www.panl10n.net/english/.../Nepal/Microsoft%20Word%20-%208_OK_N_400.pdf

[9] LataBopche, GauriDhopavkar, and ManaliKshirsagar, "Grammar Checking System Using Rule Based Morphological Process for an Indian Language", Global Trends in Information Systems and Software Applications, 4th International Conference, ObCom 2011 Vellore, TN, India, December 9-11, 2011

BIOGRAPHIES

Neethu S Kumar received the Bachelor's Degree in Computer Science and Engineering from Sree Buddha college of Engineering, Kerala, India in 2017. She is currently pursing Master's Degree in Computer Science and Engineering in Sree Buddha College of Engineering, Kerala, India.

Prof. Supriya L P. has more than 12 years of experience in teaching, Research and industry. She completed her post-graduation in Computer Science from Madras University in 2003. She received her M.Phil. From the department of computer Science in 2007, Annamalai University specialized in image processing .She received her Master of Engineering (M.E) degree from School of Computing, Sathyabama University, Computer Science and Engineering in 2009. At present she is pursuing her PhD. She started her career as a faculty of Computer Science in 2004 at Chennai. She has got a number of publications in conferences and Journals national/international.