A Review on Grammar error correction in different domains.

Kiran R. Borade[1] & Prof. N. M. Shahane[2]
[1]M.E. Student, [2]Associate Professor
[1]Department of Computer Engineering,
[1]K. K. Wagh Institute of Engineering Education & Research, Nashik,
Savitribai Phule Pune University, Maharashtra, India.

Received: December 08, 2018 Accepted: January 14, 2019

ABSTRACT: *English as second language (ESL) learners frequently make many grammatical errors while writing. This may be due to lack of understanding correct use of connectives, articles, preposition, spelling mistake, etc. Hence error correction in different domains (article, preposition correction) becomes an important grammar project in English. To correct this, many methods and strategies are proposed which use metaheuristic algorithms and classification algorithm for error identification and correction. In feature extraction many syntactic and semantic features are considered, these features have high influence in Natural language processing (NLP) task. The performance of system mainly depends on feature extraction process, appropriate use of feature selection algorithm and classification algorithm.*

Key Words: *Machine learning, Metaheuristic algorithm, feature extraction, feature selection, connective correction, NLP.*

I. Introduction

Grammar errors are often made in English writing by new English learners. Grammar error can be errors in writing articles, preposition, subordinate connectives and spelling mistakes. Many researches are done in correcting article and preposition errors. Recent research is going on English clause error correction which becomes important project in English grammar. Many features extraction methods are used depending upon the surrounding context of an article, preposition or connective position in sentence.

Grammar of a language is divided into syntax and semantic. Syntax is how words are combined to form a sentence. Retrieving the syntactic information is a primary step in pre-processing English language sentences. The tool which is used for retrieving syntactic structure from a given sentence is called parsing. Syntactic information includes dependency relation, syntactic structure and Part of Speech (POS) tag.

Stanford University proposed a statistical technique for retrieving the syntactical structure of English sentences. Based on this technique a Stanford Parser core NLP toolkit was developed. Stanford Parser is actually used for preprocessing the text in such a format that can be recognized by machine learning algorithm. If coding is done using java languages then Stanford core NLP libraries are used otherwise for python language NLTK libraries are used [9].

Many feature selection techniques are used for selecting the appropriate features for error correction and detection along with the classification algorithm for evaluating the performance of the system [1] [2]. Hence English grammar project in particular domain whether it be articles or connectives uses different strategies to deal with error correction and performance evaluation of system. The performance depends on the feature extraction and selection. Many semantic and syntactic features are considered for feature extraction and many different feature selection algorithms are used for feature selection [2]. They need to be focused because text classification problem consist of many features that are not easy to handle and slows down the speed and requires more memory [3]. The main research going in the field of NLP is speeding up the feature selection task and makes use of less memory by considering appropriate features for text classification.

II. RELATED WORK

There are various strategies and techniques used for error correction under different domains are discussed. Different features are considered to represent the text document as a dataset to be used for machine learning algorithm and feature selection methods for accurate classification that is for correcting the error and for performance evaluation.

A. Preposition and Determiners correction[1][2]

Stanford parser is employed which consists of a competitive phrase structure parser. Preposition selection is prejudiced by parse features which have a strong hold on selection in well formatted text. The

performance of preposition error detection system using parse features is improved, even though errors of learner's text, parse features make small and non significant effect. The input sentence is split into chunks before and after the preposition and both is parsed separately. A preposition model is amplified with tokenization and parse feature. After examining the output of parser shows that parse features can be extracted from ESL data.[1]

To correct errors of preposition and determiners, pipeline of confidence-weighted linear classifiers in system are used. In this system determiner and preposition correction is considered as classification problem. From possible correction based on confusion set confidence weighted linear classifiers are used to predict the correct word from set. Separate classifiers are built for correction of determiner errors, preposition replace errors, and preposition insert and deletion errors [1]. To form an error correction system the classifiers are combined into a pipeline of correction steps. System consists of pipeline of sequential steps where the output of one step serves as the input to the next step. Feature extraction analyzes the syntactic structure of the input sentences part of speech (POS) tagging, chunking, and parsing and relevant instances are identified for correction all noun phrases (NP) for correction of determiner. Determiner error correction is treated as a multiclass classification problem. A classifier is trained to predict the correct article from a set of possible article choices (a, the, an) as per the sentence context.[2]

B. Neural Networks for Correcting misuse of English Article[3]

Words surrounding the context of article are considered in feature extraction. Instead of considering features relying on human skill and prior knowledge in NLP, this approach simply helps in system automation. Depending on this approach both an error annotated corpus and an error non-annotated corpus is trained. It is possible to learn a strong statistical model on sufficient examples of error type. This system focuses on article error correction using Convolution Neural Network (CNN). The preprocessing module extract surrounding context of an article based on its position in sentence. In English, there are rules to use a, an or the considering properties of word immediately after the or a/an. These rules implemented are used to revise the output of CNN module.

C. English article error detection for non-native speakers[4]

A maximum entropy classifier is trained to select among a/an, the, or zero i.e no article for noun phrases (NPs), based on a set of features extracted from local context of each article. The system uses features based on local context in the form of words and POS tags to compute the probability that the NP will have a, an, the, or 0 article. The system's performance is evaluated in 2 ways: On held-out data from the same corpus as the training set, and on essays written for the Test of English as a Foreign Language by native speakers of Japanese and Russian.

D. Reranking technique for Grammatical Error Correction[5]

Grammatical error correction methods that employ statistical machine translation (SMT) have been proposed for dealing with many grammatical errors. An SMT system generates instances with scores for all sentences and selects the sentence with the highest score as the correction result. In SMT system 1-best result is not always the best result. Ranking approach is proposed for grammatical error correction. Also to re-score N-best results of the SMT and reorder the results reranking approach is used. When we use the discriminative reranking with features, both precision and recall increases.

E. Feature Selection Tool using Parallel Genetic Algorithm[6]

Number of discriminate features matter for higher classification performance. The smaller number of features reduces the problem of dimensionality and hence improves the performance. DWFS (Dynamic wrapper feature selection) is a tool that allows selection of features for a variety of problems. DWFS follows the wrapper approach and applies search strategy based on Genetic Algorithms (GA). A parallel GA implementation evaluates simultaneously large number of collections of features. DWFS integrates various filter methods that are applied as a pre-processing step in the feature selection step. According to the application requirements weights and parameters in the fitness function of GA can be adjusted. Experiments done using heterogeneous datasets from biomedical applications demonstrate that DWFS is fast and efficient in reduction of the number of features without sacrificing performance as compared to several existing methods.

F. Hybrid Feature Selection method using instance learning and cooperative subset search [7]

In many domains the problem of selecting most useful features from lots of features in a low sample size data set arises. In such classification tasks feature subset selection is a key problem. It is common to use filter methods. Because of correlation between genes due to which complexity of new feature techniques are in research hence ignorance in the correlations between genes which are prevalent in gene expression data and standard wrapper algorithms cannot be applied. Additionally, existing methods are not especially able

to handle the small sample size data which is one of the main cause of instability of feature selection. In order to deal with these issues, a new hybrid, filter and wrapper, based approach is proposed based on instance learning. A cooperative subset search (CSS), is used with a classifier algorithm to represent an evaluation system of wrappers. Comparison results show that existing approach is better than other methods in terms of accuracy and stability of the subset.

G. Feature Selection and Confidence Tuning for English article error correction [8]

In this paper an approach is proposed to solve problem of English article error correction, which is as important for a large proportion in grammatical errors. Genetic algorithm is employed for feature selection with confidence tuning for error correction. Machine learning based approach is considered which works on error annotated corpus and proposed a strategy to solve correction problem of article. At first, large numbers of related syntactic and semantic features are extracted from the context of connectives. With the help of genetic algorithm, a best feature subset is selected out which reduces feature dimensionality. For each testing instance, according to the predicted scores generated by the classifier, given approach measures the difference between scores in order to enhance the precision to a certain category.

H. Stanford Core NLP Toolkit[9]

The Stanford CoreNLP toolkit is designed, which is an extensible pipeline that provides core natural language analysis. It is used for semantic and syntactic analysis of sentences. It is easy for users to get started with framework, and to keep framework small, so it is easily comprehensible, and can easily be used as a component within the large system that a user is developing. It helps in analyzing the syntax and semantics of text data or sentences. It helps in feature extraction process it helps by letting know the meaning of the sentence according to which we can extract the features.

I. Two Stage Feature Selection Method for Text Categorization[10]

Text classification is to assign the documents to one of the predefined classes according to their contents. Text classification consists of many features which is a pattern recognition problem, where important step is feature selection. Still researches are going in which researchers are proposing new feature selection methods for text classification. Two-stage based feature selection methods are constituted by combining filter based feature selection methods along feature transformation methods and wrapper based feature selection method. The main objective is to do two-stage feature selection methods analysis for text classification considering different views. Filter-based feature selection methods and feature set construction methods are considered in the first stage of two-stage feature selection and in second stage feature transformation methods such as Principle Component Analysis (PCA) and wrapper-based feature selection method such as genetic algorithm (GA) is used. kNN classifier is used for text classification .

J. Correction Model for English Subordinate Clause Connective error [11]

New English learners make many grammatical mistakes in English subordinate clause connectives. It is not possible for checker to check each sentence for detecting and correcting connective errors. Hence for English subordinate clause connective error correction, an automatic error correction model is implemented. This model is nothing but a combination model of GA and kNN. At first, syntactic and semantic features are extracted based on sentence context and then feature selection is done using metaheuristic algorithm GA and learning algorithm kNN. Technologies are combined for reducing automatic feature selection time. This hybrid algorithm is used to increase accuracy and for fast classification. Finally, proposed model performance is compared with other models.

III. CONCLUSION

In this work we have studied various feature extraction and selection techniques in different domains. These techniques have high influence on system performance. English grammar project mainly focuses on correction of articles, prepositions, clause connectives and spelling error correction. Many features are considered as n-gram features, dependency features, POS features, etc. For feature selection many metaheuristic algorithms like genetic algorithm, PSO, firefly algorithm, binary gray wolf optimization algorithm, etc are considered. For classification, algorithms like kNN, Nave Bayes, SVM, etc are used for error correction and performance evaluation. Performance measures such as accuracy, f-measure, precision, etc are considered. The main problems still exist such as optimal error correction, high memory usage and time consumption.

IV. ACKNOWLEDGMENT

I would like to express my gratitude to my guide Prof N. M. Shahane Associate Professor, Computer Engineering, K.K.W.I.E.E.R., Nashik for giving me a moral support, valuable guidance and encouragement in making this literature survey. A special thanks to Prof.Dr. K. N. Nandurkar, Principal and Prof.Dr. S. S. Sane,

Head of Department of K.K.W.I.E.E.R., Nashik for their kind support and suggestions. I would like to extend my sincere thanks to all the faculty members of the department of computer for their help.

References

[1] Tetreault, Joel, Jennifer Foster, and Martin Chodorow. "Using parse features for preposition selection and error detection." In Proceedings of the acl 2010 conference short papers, pp. 353-358. Association for Computational Linguistics, 2010.

[2] Dahlmeier, Daniel, Hwee Tou Ng, and Eric Jun Feng Ng. "NUS at the HOO 2012 Shared Task." In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pp. 216-224. Association for Computational Linguistics, 2012.

[3] Sun, Chengjie, Xiaoqiang Jin, Lei Lin, Yuming Zhao, and Xiaolong Wang. "Convolutional neural networks for correcting English article errors." In Natural Language Processing and Chinese Computing, pp. 102-110. Springer, Cham, 2015.

[4] Han, Na-Rae, Martin Chodorow, and Claudia Leacock. "Detecting errors in English article usage by non-native speakers." Natural Language Engineering 12, no. 2 (2006): 115-129.

[5] Mizumoto, Tomoya, and Yuji Matsumoto. "Discriminative reranking for grammatical error correction with statistical machine translation." In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1133-1138. 2016.

[6] Soufan, Othman, Dimitrios Kleftogiannis, Panos Kalnis, and Vladimir B. Bajic. "DWFS: a wrapper feature selection tool based on a parallel genetic algorithm." PloS one 10, no. 2 (2015): e0117988.

[7] Brahim, Afef Ben, and Mohamed Limam. "A hybrid feature selection method based on instance learning and cooperative subset search." Pattern Recognition Letters 69 (2016): 28-34.

[8] Xiang, Yang, Yaoyun Zhang, Xiaolong Wang, Chongqiang Wei, Wen Zheng, Xiaoqiang Zhou, Yuxiu Hu, and Yang Qin. "Grammatical error correction using feature selection and confidence tuning." In Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 1067-1071. 2013.

[9] Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. "The Stanford CoreNLP natural language processing toolkit." In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55-60. 2014.

[10] Uysal, Alper Kursat. "On Two-Stage Feature Selection Methods for Text Classification." IEEE Access 6 (2018): 43233-43251.

[11] Huang, Guimin, Chuang Wu, Sirui Huang, Hongtao Zhu, Ruyu Mo, and Ya Zhou. "An english subordinate clause connective correction model based on genetic algorithm and k-nearest neighbor algorithm." In Progress in Informatics and Computing (PIC), 2017 International Conference on, pp. 302-306. IEEE, 2017.