Proceedings of the 2nd International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7048/35/20232156

The Investigation of Machine Learning in Grammar Correction

Zhiqi Dai[1,a,*]

[1]*English Literature, Fudan University, Shanghai, China*
a. 22300120130@m.fudan.edu.cn
**corresponding author*

Abstract: As English continues to be the dominant language acquired by the majority of the global population, the demand for learning it is on the rise. Within English education, grammar plays a vital role and can now benefit from machine learning-based correctional applications. This paper explores various methods employed in prior research for grammar correction. Notably, the feature extraction method proves to be effective in capturing essential information from the text, resulting in more precise revisions of grammatical errors. The feedback filtering module will select valid improvement advice from users to develop more efficient application. Recurrent Neural Network (RNN) which is a widely-used model can also be adopted to grammar correction due to its memorizing ability. In previous studies, these methods are tested to see their validity in English grammar correction. Results of feedback filtering module show that it can sort out users' advice into "useful" and "useless" so that the modification of the application can be more accurate. In another experiment, the F-0.5 score of RNN is measured with several other models and RNN has apparent advantage over the majority in grammar error detection and correction. Admittedly, however, these methods still have space for further enhancement to provide high precision in correction. Means to eliminate possible errors and inaccuracy are urged to be found, but probably the only way out is the innumerable data fed to computers. This paper offers a comprehensive view of current study progress in the field and encourage new evolution.

Keywords: Grammar Correction, Machine Learning, Artificial Intelligence

1. Introduction

The importance of English education can no longer be ignored nowadays with the growing trend of pursuing further study abroad and the interest of people travelling around the world. While there are limited resorts for people to have access to English education, particularly for those who have already completed their formal education. When the expensive tuition fees of English educational institutions are out of reach, their only recourse is self-study. Fortunately, since the 21st century, online learning is becoming a trend. Education resources are everywhere online: numerous applications and websites targeted at subjects ranging from science, liberal arts to even music all welcome learners to give a try. Technology skills such as Automatic Speech Recognition, automated content generation and embedded experiments are all combined with educational skills to improve learning efficiency and provide more opportunities for learners.

DOI: 10.54254/2753-7048/35/20232156

Language educational technology also plays a nonnegligible role in it because the vocabulary memorizing applications can be the most widely used English learning 'assistant' in China. However, the significance of grammar correction for English learners can never be ignored since message can be easily misunderstood with grammatical problems. English grammar error correction has been a rising topic in Natural Language Processing (NLP) and many researches have been done towards the topic. It can be roughly divided into spoken grammar and written grammar. There are distinct study results in terms of methods and models in those two distinct facets. And this paper will mainly study previous researches in the written English grammar error correction field and point out the overall trend to provide a direction for the research gaps to be filled in the future studies.

By using NLP, computer programming technology can more efficiently detect and correct grammar errors in texts written by nonnative English learners. Regarding to the methods applied in former studies, rule-based method has been used mostly in early grammar correction tools [1] and two other methods more currently used are statistic-based method and depth-based method. However, the effectiveness of either approach relies on the availability of accurate data collected through a reliable system. Traditional English grammar error correction system is based on machine learning and data mining [1]. Chen [2] gave a new attempt to use deep learning technology in order to solve the existing problems of the traditional system. The classification model is the choice of traditional English grammar error correction system [1] and in latest studies of deep models [3], the Transformer model performs better in detection and correction than most grammar models.

By analyzing the previous studies on machine learning applied to English grammar correction, this paper generalizes valid methods, effective models and useful technologies in the latest research paper and synthesizes different ideas of various outstanding researchers to reach a conclusion of what progress the studies in English grammar error correction system have received. The specific framework of this paper is listed as follows: Section 2 introduces the methods; Section 3 discusses the results of the previous studies and the applications; Section 4 is the conclusion.

2. Method

Numerous methods had been adopted into the correction of English grammar application and methods are evolving to boost efficiency and accuracy in the correcting procedure. In this section, methods used by previous studies will be introduced to provide a more comprehensive understanding of how machine learning can be used in English grammar correction.

2.1. Feature Extraction Method

The feature extraction method operates as its name suggests, extracting significant features from raw data for the purposes of analysis, modeling, or prediction [4, 5]. It serves as an indispensable step in machine learning and data mining since the quality and quantity can all directly affect the performance and accuracy of the model. It aims to turn raw data into more representative and interpretable feature to describe and distinguish data. A feature extraction model normally contains four steps: data pretreatment, feature selection, feature extraction and feature dimension reduction. And there are three main domains in which feature extraction method is widely used, including classified variable, pictorial information and text information.

The basic process of feature extraction method in English grammar correction is illustrated in Figure 1. Four methods of feature extraction may be applied: generate n-grams (sequences of adjacent words) to capture local context; identify patterns of part-of-speech tags that commonly occur in incorrect grammar; design rule-based features that capture specific grammar rules or errors and extract features related to specific words or phrases that are indicative of grammar errors.

DOI: 10.54254/2753-7048/35/20232156

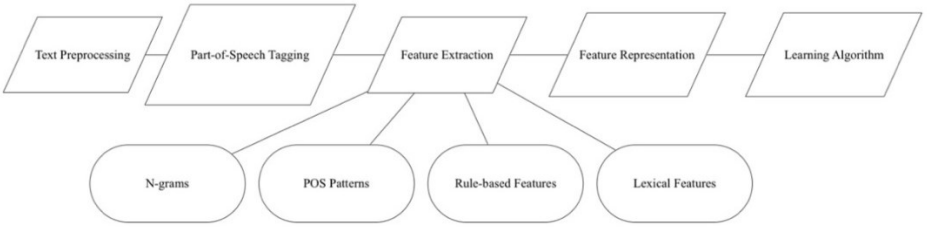

Figure 1: The basic process of feature extraction method in English grammar correction (Photo/Picture credit: Original).

2.2. Feedback Filtering Module—A Deep Learning-based Method

The feedback filtering module lacks widespread recognition in the deep learning field. However, it offers the potential to filter valuable user-written feedback for designers to consider as suggestions and responses regarding their experience with the application. It has larger utility in the field of User Experience, but also can show its power in refining of this English grammar correction application. By processing and analyzing the feedbacks on the correction received from some professionals in English or English native speakers, whose advice can be most valuable in English grammar field, this application can be adjusted and refined to suit the users' need and provide as high accuracy grammar as possible for learners.

The feedback filtering process is shown in Figure 2 [2]. The application will correct the sentence first based on existing rules and send back to the users for feedback. A system where users can provide feedback on the corrections made to their text should also be established. Through feedback filtering process, if the feedback is legal for computer to understand and process, it will be collected into a corpus. Sentences in the corpus will be corrected a second time based on their feedback as a second version. The original corrected sentence will be compared to the second version to assess whether the sentence's quality has improved. If it has, the recommendations will be considered for future corrections.

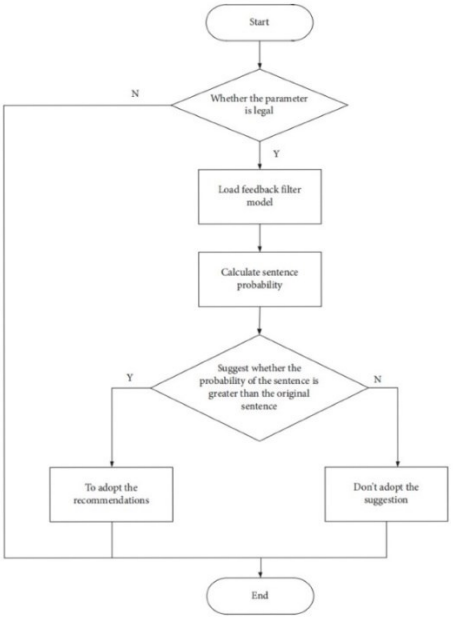

Figure 2: Feedback filtering process (Photo/Picture credit: Original).

DOI: 10.54254/2753-7048/35/20232156

2.3. Recurrent Neural Network (RNN)

RNN is a neural network structure featured as its recurrency in memorizing information inputted in the past. Traditional neural network can only processes information in one-way from input to output, while RNN has the record of the past data and thus develops the ability to propagate in both direction and deal with sequential data such as sentences as further elaborated next. A typical RNN model involves three layers: an input layer, a hidden layer and an output layer. The structure of RNN is illustrated in Figure 3. The cyclic refreshing arrows marked between hidden layer are the key to accomplish its "memory".

Due to the specific competence of RNN, the application in English grammar correction can facilitate in capturing contextual dependencies. The workflow of RNN in English grammar correction is illustrated in Figure 4. With the dataset of paragraph containing both correct and incorrect grammar, next step is encoding. It will translate English into language which RNN can understand and process. Model can thus be generated based on former process after which inferences will be made to highlight the grammar mistakes conducted in the paragraph and correct them. Though RNN is relatively effective in English grammar correction, post-processing is still necessary after all the process done by machine learning in order to adjust the function and measure its accuracy.

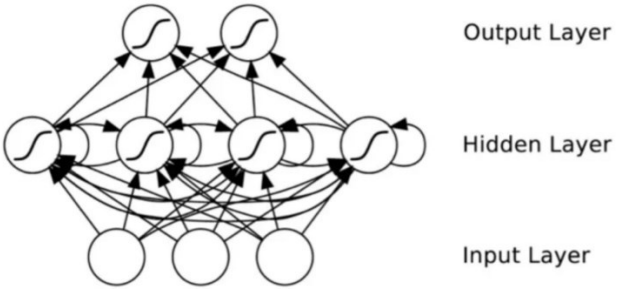

Figure 3: The structure of RNN [6]

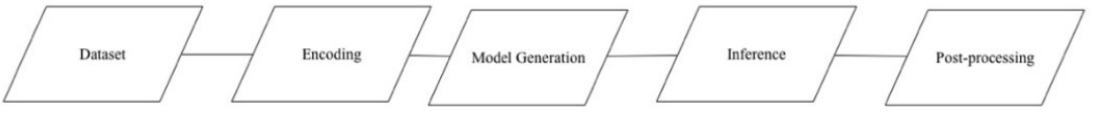

Figure 4: The workflow of RNN in English grammar correction (Photo/Picture credit: Original).

3. Results and Discussion

3.1. Results

3.1.1. The Result of Feedback Filtering Module

Figure 5 showed by Chen [2] in his essay demonstrates that the accuracy of error detection in English grammar based on deep learning is higher than other method. And in Table 1, an experiment conducted by Zhang [7], the sentence above is a provided refined version by user and the below one is by the system. Through calculation of feedback filtering module, the confusion degree of user is 171.673 while that of the system's modification is 228.743. Thus, the result turns out to filter that suggestion by user, which achieve the expect of its function.

DOI: 10.54254/2753-7048/35/20232156

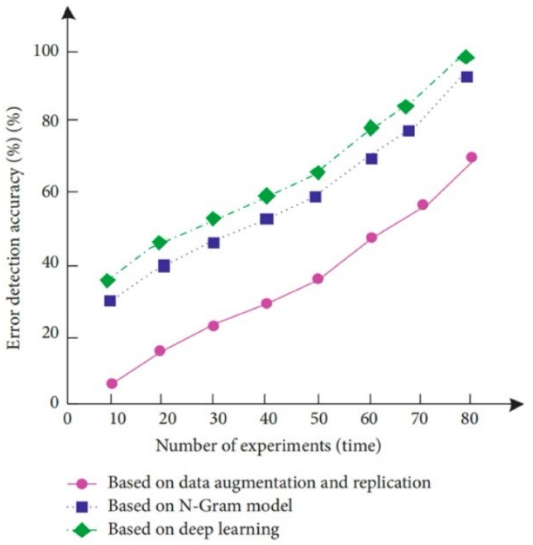

Figure 5: Comparison results of English grammar error detection [2]

Table 1: Filter calculation results

	User's modification	*System's modification*
logprob	-44.684	-42.4683
ppl	134.366	171.859
ppll	171.673	228.743

3.1.2. The Result of Recurrent Neural Network

The RNN sequence model is powerful in fitting text tasks, and it has a prominent advantage in processing long texts owing to its memory. In Table 2, RNN is being compared with several other methods for grammatical error detection and correction in English composition in the experiment. Variants being evaluated contains precision, recall and F-0.5 score:

$$F-0.5\ score = (1+0.5^2)\times\left(\frac{precision\times recall}{0.5^2\times precision+recall}\right) \quad (1)$$

Statistics show that the overall ability of RNN is rather good in those methods.

Table 2: RNN comparing with other methods

Method	Precision (%)	Recall (%)	F-0.5 score(%)
SVM [8]	76	71	74
Decision Tree [9]	79	74	76
RNN [10]	83	76	81
LSTM [11]	85	77	81

3.2. Discussion

The current studies of English grammar correction have already used various kinds of methods and models to enhance the efficiency and precision of the correction. By developing more comprehensive databases and adopting more powerful models, the application is able to provide strong function with better user experience.

However, there still exists gap for researchers to fill in the future studies. The feature extraction method depends largely on accurate recognition of part-of-speech in the given text while the hidden

DOI: 10.54254/2753-7048/35/20232156

meaning of sentences may sometimes be contextualized making it challenging for computer to perceive. The feedback filtering module also accounts on dependable feedback from authentic authority, otherwise the grammar rules input will cause grammatical disorder instead of providing a good correction. Similarly, RNN has its shortage too as in Table 2. Long Short-Term Memory (LSTM) shows advantages over RNN in statistics. LSTM is put forward in order to solve the gradient vanishing of RNN and bears the capability to memorize long-term data which RNN has problem with. Therefore, future development of English grammar correction is worth being expected and can be focused on directions such as contextual understanding and multilingual grammar correction.

4. Conclusion

This paper has listed several valid methods in machine learning field which can be adopted in English grammar correction and illustrated results of employing these methods to testify their availability in real circumstances. An overall analysis and synthesis of current existed research studies has been conducted in this paper in how machine learning can apply to English grammar correction. Feature extraction method, feedback filtering module and RNN are discussed for further method selection in English grammar correction applications. Conclusions are reached that RNN, as a classic neural network model, still successfully boost the efficiency of grammar correction, while it has drawbacks that can be covered by other newer technologies. Decisions should be cautiously made when choosing the better model for larger benefits according to different use conditions. Pragmatic grammar, which primarily serves as a communication tool in everyday language use, offers users considerable freedom and flexibility. Its effectiveness is determined by its ability to ensure mutual understanding among all participants in a communication group. Consequently, the challenges for future research lie in addressing questions such as how Artificial Intelligence can differentiate between formal grammar and pragmatic grammar and how to ensure that the data input for AI learning are valid examples of formal grammar. These questions await further investigation and resolution.

References

[1] *Shanchun, Z., Wei, L. (2021) English Grammar Error Correction Algorithm Based on Classification Model. COMPLEXITY.*

[2] *Chen, H. (2021) Design and Application of English Grammar Error Correction System Based on Deep Learning. Security and Communication Networks.*

[3] *Zhu, J., et al. (2021) Machine learning-based grammar error detection method in English composition, Scientific Programming, vol. 2021, pp. 1–10.*

[4] *Khalid, S., Khalil, T., & Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. In 2014 science and information conference (pp. 372-378). IEEE.*

[5] *Çayir, A., Yenidoğan, I., & Dağ, H. (2018, September). Feature extraction based on deep learning for some traditional machine learning methods. In 2018 3rd International conference on computer science and engineering (UBMK) (pp. 494-497). IEEE.*

[6] *Naveenkumar, M., & Kaliappan, V. K. (2019, November). Audio based Emotion Detection and Recognizing Tool Using Mel Frequency based Cepstral Coefficient. In Journal of Physics: Conference Series (Vol. 1362, No. 1, p. 012063). IOP Publishing.*

[7] *Zhang, A. (2022). Analysis of the Application of Feedback Filtering and Seq2Seq Model in English Grammar. Wireless Communications and Mobile Computing, Volume 2022, Article ID 9530379, 8 pages. Retrieved from https://doi.org/10.1155/2022/9530379*

[8] *Joachims, T. (1998). Making large-scale SVM learning practical (No. 1998, 28). Technical report.*

[9] *Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6), 275-285.*

[10] *Medsker, L. R., & Jain, L. C. (2001). Recurrent neural networks. Design and Applications, 5(64-67), 2.*

[11] *Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270.*