IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 6, June 2020

Grammar Error Correction using Seq2Seq

Jannya V[1], Savyan PV[2]

M. Tech., Student, Computer Science and Engineering, GEC, Palakkad, India[1]

Assistant Professor, Computer Science and Engineering, GEC, Palakkad, India[2]

Abstract: Grammatical Error Correction (GEC) in English language is a challenging topic among the emerging works. GEC is a process of converting the erroneous sentences to a corrected sentence by using Sequence2Sequence (Seq2Seq) method. Usually the system focused on correcting the grammars based on the 20 rules in English language and it includes punctuation, grammatical and word choice errors. Deep learning method is used to work behind the system. Long Short-Term Memory (LSTM) Encoder - Decoder model is used in the conversion of incorrect sentence to a grammatically corrected sentence. This is a supervised learning system which includes incorrect and corrected sentences in the GEC dataset and thus gives better results.

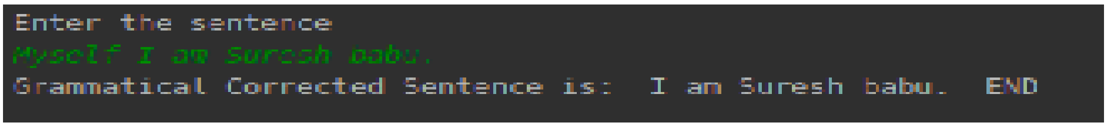
Keywords: GEC, Deep Learning, seq2seq, LSTM.

I. INTRODUCTION

English language is important in many areas especially in educational fields and competitive exams. In schools, colleges, competitive exams and even in interviews English language played an important role. Here is able to realize the importance of GEC system. Studying and listening for hours in a grammar lecturing makes our mind boring. If we are giving an incorrect sentence and getting its corresponding corrected sentence and finding what is the error? and understanding it by knowing the rules makes a student brilliant. GEC system is able play an important role in the field for the future generation, where they give more value to the electronic machines. So, each student and learner are able to study the grammar when they have time and when they are interested thus the GEC system help to guide them in a better manner.

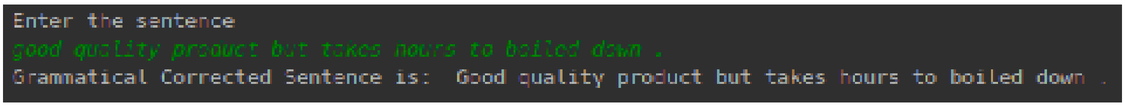
Recently, there are many grammar detection and correction online tools such as Grammarly, GingerIt, Language Tools etc. These tools focused only on some of the English rules and some of these tools makes wrong predictions. Thus, a user is not able get a correct output from these tools. GEC system is able to overcome all these tools by making correct predictions based on the punctuation, grammar and word choice errors. GEC system helps in auto completion of sentences by predicting more sentences from a single sentence.

Figure 1 shows one of the example in a GEC system. Incorrect sentence is entered to the system and corresponding corrected sentence is given as it is based on the word choice errors. Figure 2 describes one of the grammatical rules in English where the incorrect sentence starts with a lowercase and in its corresponding corrected sentence it corrects the first letter to uppercase based on the English grammar rules.



```
Enter the sentence
Myself I am Suresh babu.
Grammatical Corrected Sentence is:  I am Suresh babu.   END
```

Figure 1: Example of corrected sentences for corresponding incorrect sentences



```
Enter the sentence
good quality product but takes hours to boiled down .
Grammatical Corrected Sentence is:  Good quality product but takes hours to boiled down .
```

Figure 2: Example showing one of the rule in English Grammar

A. Applications

Application of Grammar Error Correction are:

- Automatic Grammar Correction System
- Automatic Spelling Correction System
- Knowledge Transfer

The section 2 includes various works done in the field of GEC. Section 3 includes system overview of the proposed GEC system and Section 4 is discussing about the model working behind the GEC system.

II. METHODS OF GRAMMAR ERROR CORRECTION SYSTEM

The section 2 deals with detailed note on related works done in the field of grammar error correction.

A. Methodology-1 Shamil Chollampatt et.al 2018

GEC system by Shamil Chollampatt et.al mainly focused on correcting the errors of learner's writings. Mainly in this system it depends on learner's writings as where they makes spurious corrections and fails to recognize the errors and these errors misleads the learner. Thus the correction of errors is done by using neural quality estimation of output sentences. This is trained in supervised manner where the instructors correct the writings and produces the output based on the quality score labels using the human annotated references. NUCLE dataset is used for neural quality estimation method to bring better results.

B. Methodology-2 Alla Rozovskaya et.al 2016

This approach of grammar error correction focused on machine learning classification and machine translation. Ma-chine translation method mainly helps in correcting complex mistakes whereas machine classification method trains the model without human annotated data and it have the ability to generalize the training of the model and is flexible to adjust the individual error types. Thus, the combination of machine translation and machine learning classification method is able to gain a better advantage in making the system more understandable and gives better results in correction of errors in grammar.

C. Methodology-3 Shashi Pal Singh et.al 2016

This GEC system proposed rule based grammar checking and is focused on the English language, as its an important one and is not easy to understand the meaning when the grammar mistakes takes place. The aim of the system is that to reduce the errors while speaking the language. This work mainly focused on the context and the tone of the speaker where the meaning of a sentence varies. Grammar checker is based on five rules in English grammar and its able to detect and correct the tense related mistakes. As it is rule based the model have many limitations and it produces wrong outputs when the rule breaks and this may misleads the user.

D. Methodology-4 Mariano Felice et.al 2013

This system is based on injecting the artificial errors for correcting the mistakes done by learner's and this system also focused on the English language. The aim of this GEC system is that it trains on artificially injected errors to improves the better precision at the expense of recall. This method have done the work in NUCLE and CoNLL 2013 dataset which contains sentences that is erroneous.

III. GRAMMAR ERROR CORRECTION MODEL

Grammar Error Correction (GEC) model is proposed by deep learning method, where seq2seq method with LSTM Encoder-Decoder model is used to translate incorrect sentence to corrected sentences. GEC dataset preparation is the most important task. GEC dataset is prepared with the help of WordNet lexical database. WordNet database gives multiple grammar corrected sentences for a single word and this helps throughout creating the GEC dataset. Figure 3 gives the overall idea of the proposed GEC system. Dataset includes incorrect sentences and its corresponding corrected sentences. Pre-processing step includes capitalizing first letter, adding periods and removing unwanted spaces in corrected sentences. Thus, the incorrect sentence is given to the encoder and then it is passed to the hidden layer by considering the context of sentences it is moved to decoder to generate the corresponding corrected sentences. This LSTM model leads to generate corrected sentences. This is the overall concept of proposed GEC system.

A. Seq2Seq Machine Translation Model

The LSTM Encoder-Decoder is a recurrent neural network designed to address seq2seq problem. In the proposed system the incorrect sentences have a different number of items and in corrected sentences it may vary so this can be solved by using the sequence to sequence model. Seq2Seq model helps to predict the corrected sentences for the corresponding incorrect sentences. The applications of Seq2Seq model includes machine translation, chatbot, learning to execute programs, question answering etc. Figure 4 shows a translation of incorrect sentence and after passing it through the seq2seq model it generates the corrected sentences.

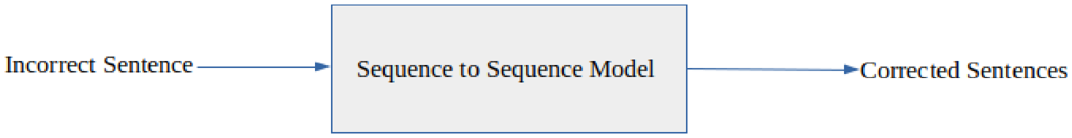


Figure 4: Seq2Seq Machine Translation

DOI 10.17148/IJARCCE.2020.9652

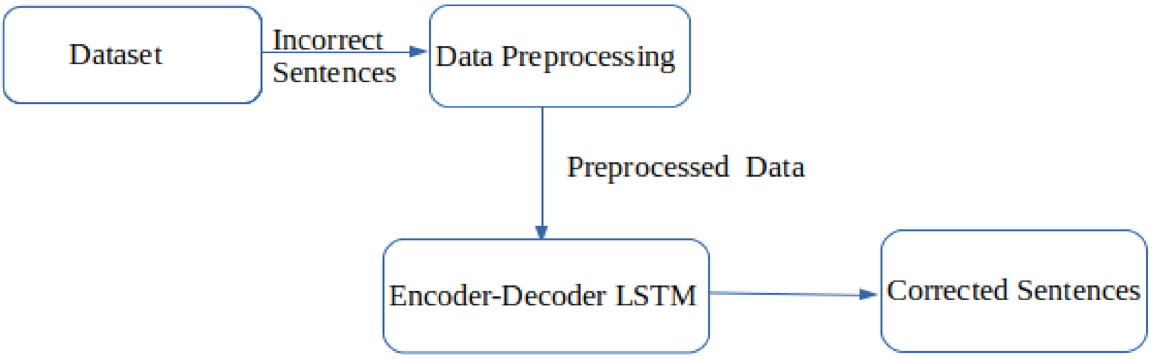


Figure 3: System Overview

IV. LSTM ENCODER -DECODER MODEL

The LSTM architecture consists of encoder and decoder model. Figure 5 shows that the incorrect sentence is given to the encoder model and by considering its length it pass the information to the hidden layers and the output of the encoder is discarded. From the hidden layers by considering the context of the sentence it is given as the input of the decoder model.

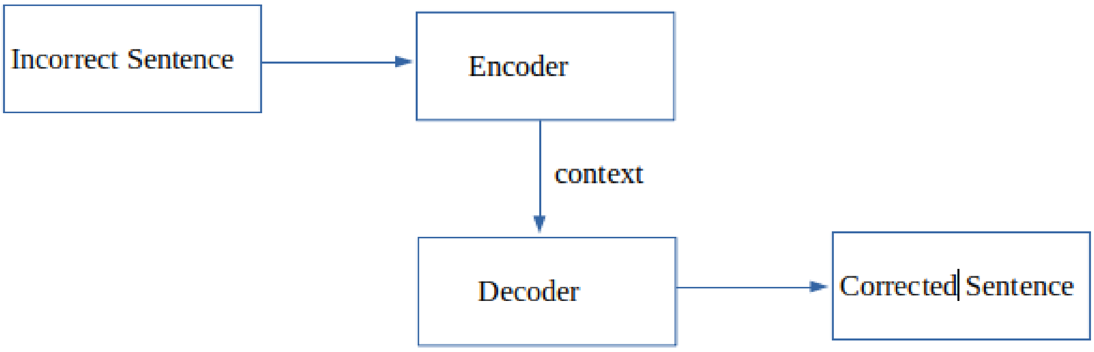


Figure 4: LSTM Model

The decoder model works separately for training and testing phase. The decoder model understands where a sentence begin and end based on the symbol that is given in the sentence. When the decoder find ’END’ the model understands that here the model ends the corrected sentences.

A. Training

Teacher Forcing method is used in the proposed model for training the LSTM model. Teacher Forcing works by considering the previous time step. By considering the start symbol ’START’ helps to understand that a sentence begins and ’END’ gives that the sentence is completed so without prediction by looking the previous step the training is able to done accurately. Inference algorithm is also considered in the proposed model for training the model by making idea about the starting and ending of the corrected sentences

V. CONCLUSION

The system is based on the supervised learning method where the dataset contains both incorrect and corrected sentences. This supervised learning method is implemented using deep learning neural networks, which includes LSTM model to generate the corrected sentences for the corresponding incorrect sentences. A large number of data is needed to improve the accuracy of the system. Tried to include majority of words in English vocabulary thus there are many words that are not included in the dataset thus it produces key error so try to include more words for better results. The dataset preparation takes time and it is needed to be correct and accurate then the model is able to produce better accuracy. This GEC model is simple, understandable and user friendly as every user are able to work on this GEC system and they can get the correct usage of grammar, especially it is more usable to the students in schools and colleges for getting the knowledge and understanding the meanings of every sentences. GEC system shows the incorrect sentences and its corresponding corrected sentences, thus it produces accurate sentences

In future, LSTM can be replaced with BiLSTM for better performance and to get an accurate result. Can be tried by increasing the data in the dataset for having accurate predictions.

DOI 10.17148/IJARCCE.2020.9652

ACKNOWLEDGMENT

First and foremost, I wish to express my wholehearted indebtedness to God Almighty for his gracious constant care and blessings showered over me for the successful completion of the project. Gratitude may be extended to my parents, teachers and my friends who supported me to complete the project.

REFERENCES

[1]. S. Chollampatt and H. T. Ng , “Neural quality estimation of grammatical error correction”, in Proceedings of the 2018 Conference on Emperical Methods in Natural Language Processing, pp.2528-2539, 2018.

[2]. A. Rozovskaya and D. Roth,”Grammatical error correction: Machine translation and classifiers”, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2205-2215, 2016.

[3]. S. P. Singh, A. Kumar, L. Singh, M. Bhargava K. Goyal, and B. Sharma, “Frequency based spell checking and rule based grammar checking”, in 2016 International Conference on Electrical, Electronics, and Optimization Techniques(ICEEOT), pp. 4435-4439, IEEE, 2016.

[4]. M. Felice and Z. Yuan, “Generating artificial errors for grammatical error correction”, in Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 116-126, 2014 .

[5]. K. Imamura, K. Saito, K. Sadamitsu, and H. Nishikawa, “Grammar error correction using pseudo-error sentences and domain adaptation”, in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Vol 2: Short Papers), pp. 383-392, 2012.

[6]. G. Sidorov, A. Gupta, M. Tozer, D. Catala, A. Catena, and S. Fuentes, “Rule- based system for automatic grammar correction using syntactic n-grams for english language learning(l2)”, in Proceedings of the Seventeeth Conference on Computational Natural Language Learning: Shared Task, pp. 96-101, 2013.

[7]. J. Lichtarge, C. Alberti, S. Kumar, N. Shazeer, N. Parmar, and S. Tong, “Corpora generation for grammatical error correction”, arXiv preprint arXiv:1904.05780, 2019.

[8]. N. Madi and H. S. Al-Khalifa, “A proposed arabic grammatical error detection tool based on deep learning”, Procedia computer science, vol. 142, pp. 352-355, 2018.

BIOGRAPHY

Jannya V earned her B. Tech in Computer Science and Engineering in MEA Engineering College, Perinthalmanna, India and is currently pursuing M. Tech in Computational Linguistics in Government Engineering College, Palakkad, Kerala, India. And completed my internship in Larsen and Turbo(L&T) Infotech, Bangalore.

DOI 10.17148/IJARCCE.2020.9652