[1] **Paulo Miguel A. Cano**
[2] **Juhl Mayne L. Catiggay**
[3] **John Michael M. Florida**
[4] **John Mathew R. Pua**
[5] **Mary Jane C. Samonte**

Natural Language Processing of Grammar Checker Tools for Academic Writing: A Systematic Literature Review

Abstract: Grammar correction is crucial for essential documents such as research, business articles, formal papers, but with language evolving every century, some rules for grammar correction are added or changed. There have been created grammar correction online, which has helped most students create academic writing. Some tools focus on correcting the paper by following specific rules; some run an algorithm based on what type of document you are writing. In this paper, we reviewed documents published from 2017 up to 2021 related to grammar correction and Grammar error detection, with numerous Natural Language Processing and models provided by past researchers, which may aid this paper towards a solution on creating a grammar correction tool.

Keywords: Computer Systems Organization, Real-Time Systems, Real-Time Languages, Deep Learning, Grammar, Grammar Correction, Grammar Error Detection, English Language

I. INTRODUCTION

The use of Grammar checking tools is to improve the sentence structure used in the document. It can also adjust to the tone of the user based on how the writer wants it to appeal to the readers. Like known grammar checking tools like Grammarly, they provide suggestions and feedback based on how the writer writes and compare it to other data using Natural Language Processing. Research stated that the English language has a set of rules that defines the structure of the sentence. The ability to write a paper with minimal to no grammatical errors is a skill requiring experience.

Linguistics guides us to better communicate with people around us. It has two major components, which are grammar and language. Grammar is a set of systematic rules wherein language can be composed or expressed. On the other hand, language in linguistics refers to sounds or gestures that people create that is often accompanied by grammar rules wherein it has five main domains: Pragmatics, Syntax (sentence structure), Morphology (structure of words), Semantics (meaning or definition), and Phonology (the sounds of language). In deep learning and natural language processing, AI researchers should address each of these components separately to achieve accurate error recognition and grammar improvement results.

With the help of Natural Language Processing, detecting grammatical errors would be possible through algorithm-based sentence structure and identifying grammar errors through a classification model such as neural networks, decision trees, genetic algorithms, or support vector machines.

II. METHODOLOGY

The Online Databases used for this study are Hindawi (Hindawi.com), IEEE Xplore Digital Library, ScienceDirect (Elsevier), ACM DL, Google Scholar (scholar.Google.com)

2.1 Study Filtering

There are 40 papers gathered based on the general topic of Grammar and Natural Language Processing from the selection process. The group chose 20 documents among the 40 papers collected according to their relevance to the study. The list of papers evaluation includes:

1. If the research paper discusses grammar and language filtering / re-arrangement?

[1] School of Information Technology, Mapúa University, Manila, Philippines. paulocano1316@gmail.com
[2] School of Information Technology, Mapúa University, Manila, Philippines. jmlcatiggay@gmail.com
[3] School of Information Technology, Mapúa University, Manila, Philippines. johnmichaelflorida@gmail.com
[4] School of Information Technology, Mapúa University, Manila, Philippines. johnpua.jmp@gmail.com
[5] School of Information Technology, Mapúa University, Manila, Philippines. mjcsamonte@yahoo.com

2. Does the article discuss further Grammar construction or language fluency?
3. If the filtered research gives numerical results upon conducting the investigation?

The summarized research are as follows:

Research Title	Source of Study	Deep Learning Techniques Used
Adaptive Spelling Error Correction Models for Learner English [1]	ScienceDirect	Natural Language Processing, Noisy Channel Model, Word Embedding-Based Method
Context-Sensitive Malicious Spelling Error Correction [2]	IEEE	Deep Learning
Deep Learning for Natural Language Parsing [3]	IEEE	Natural Language Processing, Deep Learning
Using Very Deep Convolutional Neural Networks To Automatically Detect Plagiarized Spoken Response [4]	ACM DL	Neural Networks Natural Language Processing, (Unsupervised Language Model)
The Creation and Evaluation Of a Grammar Pattern List For the Most Frequent Academic Verbs [5]	ScienceDirect	Natural Language Processing
Accelerating Mixed Methods Research With Natural Language Processing of Big Text Data [6]	Sage Journals	Natural Language Processing, Computer Assisted Qualitative Data Analysis (CAQDAS) Software, Machine Learning & Data Mining
Applying Natural Language Processing Tools to a Student Academic Writing Corpus: How Large are Disciplinary Differences Across Science and Engineering Fields [7]	Google Scholar	Tool for the Automatic Analysis of Lexical Sophistication (TAALES), Tool for the Automatic Analysis of Cohesion (TAACO)
Assessing Students' Use of Evidence and Organization in Response-To-Text Writing: Using Natural Language Processing for Rubric-Based Automated Scoring [8]	Springer Link	Natural Language Processing
Context-Aware Text Matching Algorithm for Korean Peninsula Language Knowledge Base Based On Density Clustering [9]	Hindawi	Text Matching Algorithm, Ah–Corasick algorithm
English Grammar Detection Based on LSTM-CRF Machine Learning Model [10]	Hindawi	LSTM-CRF Machine Learning Model, Neural Network Algorithm
English Grammar Discrimination Training Network Model and Search Filtering [11]	Hindawi	Natural Language Processing,
English Grammar Error Correction Algorithm Based on Classification Model [12]	Hindawi	Feature Extraction Model, Classification Model, Neural Networks Natural Language Processing, Machine Learning
English Grammar Error Detection Using Recurrent Neural Networks [13]	Hindawi	Neural networks, Natural Language Processing
How has Science Education Changed over the last 100 years? An analysis using Natural Language Processing [14]	Wiley	Natural Language Processing, Machine Learning, Latent Dirichlet Allocation (LDA Model)
Microclassroom Design Based on English Embedded Grammar Compensation Teaching [15]	Hindawi	Model based on the Education systems
Research on Business English Translation Architecture Based	Hindawi	Speech Recognition, Neural Networks

on Artificial Intelligence Speech Recognition and Edge Computing [16]		Edge Computing Tech.
Research on Intelligent Calibration of English Long Sentence Translation Based on Corpus [17]	Hindawi	Semantic Ontology Model of English Long Sentences. Bilingual Evaluation Understudy (BLEU)
Resolving "orphaned" non-specific structures using machine learning and natural language processing Methods [18]	NIH	Natural Language Processing, Machine Learning, Support Vector Machine (SVM)
Text Filtering through Multi-Pattern Matching: A Case Study of Wu–Manber–Uy on the Language of Uyghur [19]	MDPI	Natural Language Processing, Machine Learning,
Understanding EFL Linguistic Models through Relationship between Natural Language Processing and Artificial Intelligence Applications [20]	SSRN	Natural Language Processing, Neural Networks, Machine Learning

III. Results And Discussions

This paper aims to identify the topics and methods used to create a Grammar Correction tool; as follows (1) What are the algorithm and tools used to identify grammar errors and word suggestions?; (2) What deep learning techniques do researchers use?; (3) the datasets used in previous studies. The selected studies are clustered by publication year, with the year 2019 having the highest number are listed; the year 2017 (3 papers with the reference number of [1, 7, 8]); the year 2018 (1 paper with the reference number of [18]); the year 2019 (4 documents with the reference number of [2, 3, 4, 19]); the year 2020 (1 paper with the reference number of [5]); and year 2021 (10 documents with the reference number of [6, 9,10,11, 12, 13, 14, 15, 16, 17]).

3.1 Analyzing Grammar Correction

With the evolution of language, there are challenges of which students encounter when writing an academic paper. With little to moderate experience with writing papers, there will be grammatical errors that students may register. A related study conducted by Hindawi [10] determined that the evolution of language is preset in AI development. These grammatical changes have been anticipated, and much of these changes will be incorporated in the Tool by deep learning neural network structure.

3.2 Algorithms and Tools Used

As mentioned earlier, the range of the publication year of the collected research papers is from 2017 to 2021. Most of these studies utilized and curated timely and efficient algorithms, which the researchers used to improve language analysis and grammar-related deep learning and natural language processing methods. Language as the study of grammar and linguistics has different domains and rules that need different approaches to identify errors and clarity in semantics, syntax, pragmatics, spelling, morphology, and even plagiarism.

The list of tools that the gathered papers used are the Word2Vec deep learning tool, TAALES or Tool for the Automatic Assessment of Lexical Sophistication [7], and Scikit-learn [4]. 8 out of the 20 research papers used the Word2Vec tool as a deep learning tool (8 documents with the reference number of [1, 2, 4, 6, 9, 11, 19, 20]). The Word2Vec tool has been widely used for natural language processing since 2013. Word2Vec is a particular deep learning tool that produces something called word embedding. To explain further, Word2Vec associates a word into a vector or an array of numbers, and it's pretty intuitive to know, hence the tool's product name. Researchers also have a set of properties assigned to a dataset of labels, placing it into a graph or regression analysis to see the output and results.

The compilation of the research papers has varying algorithms from each other, where each algorithm is based on its use and purpose. Most algorithms are pattern matching or searching algorithms based on a corpus or dictionary: the Wu-Manber Algorithm[19] (hashing technique) and Ah–Corasick algorithm [9] (string-searching). Furthermore, other algorithms used are the VF2 algorithm [11] (graph-matching algorithm), backpropagation algorithm[16] (quickly calculate derivatives in a graph), LDA[14] (The Latent Dirichlet Allocation extracts definition from a word) algorithm, and the others are generic algorithms that the researchers created and proposed.

3.3 Language Error Detection

To create a language error detection algorithm, it is needed to use AI because of the wide variety of combinations of words there is in the English language. To detect language errors, it is necessary to divide them into two parts: spelling and grammar. Spelling exists so that words can be differentiated from one another, while grammar is essential to understand the usage and the context of the way the words are used. In order to correctly detect proper grammar usage, it is necessary for the spelling to be correct, hence the algorithm must also check for the spelling of the said text if it is in the English dictionary. The two work hand-in-hand when it comes to making language error detection tools.

The first step is to use the bag-of-words model to extract the text from the academic papers[1]. The boW is used because of its simplicity and flexibility when it comes to obtaining features from documents. The input used can be a learner corpus. Corpus is used to study many documents to produce tools such as spelling and grammar checkers. The target academic papers can also be added to increase the training data. The way of how BoW will work is that it first divides the document into sentences, then the sentences into words. This is when it can then move on to spelling error detection.

After the words are processed correctly, the process moves to spelling detection. Word vectors in NLP are used to analyze characters' relationships between words to contextualize them and find the intended word[2]. This is more appropriate as in most academic papers, and there is a context behind the usage of words instead of the noisy channel model.

A study done by Liqin Wu and Meisen Pan[10] used radial basis function neural networks to detect errors in the grammar used. As neural networks act as the brain in recognizing patterns, it is needed to acknowledge correct grammar patterns used in English. As the semantics of the English language is complicated, it is required to make further adjustments to the algorithm. Hence a radial basis function is added to improve its training.

IV. Proposed Grammar Checking System

Shanchun Zhou and Wei Liu[12]; many algorithms were based on a classification model because most trust it. After all, it automatically recognizes errors and corrects the grammar errors in English text written by nonnative language learners. *Figure 1.1* is from Shanchun Zhou and Wei Liu's[12] research; it shows the Algorithm process of English grammar error correction based on classification mode.

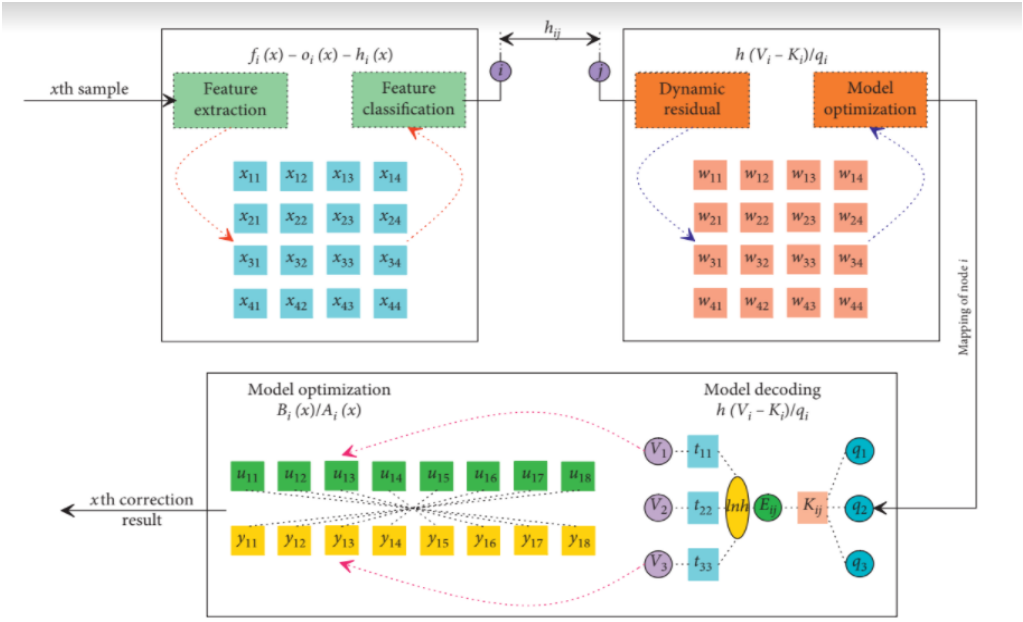

Figure 1.1

The proposed system aims for an efficient way of detecting grammar errors. To test the system, most research and projects try their algorithm based on gathered extensive data of corpus. A grammar checking system will pave its purpose to efficiently detect a user's grammar error and suggest an enhancement to its prior error. It will also highlight the text errors in grammar, and the users will have an option to change it for the betterment of their writings. The proposed system is created for academic-related work, but people can use it for work-related or personal agendas.

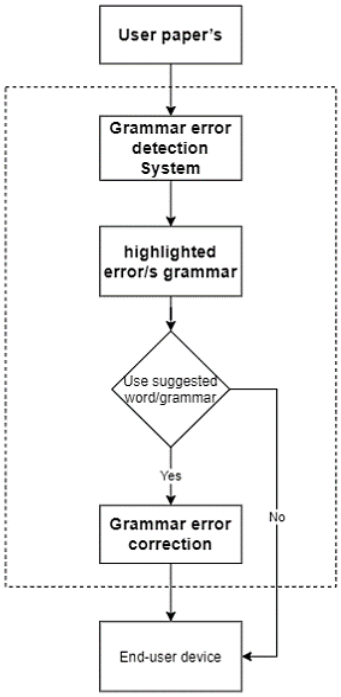

Figure 1.2

V. Implication And Conclusion

This paper has succeeded in determining the features applied through deep learning in previous studies and how Grammar Correction tools effectively detect errors that most students write in Academic Writing. The results have led us to formulate a Grammar Correction System, which became the basis for the model we've developed, the Grammar Correction model. The research was conducted by narrowing down results from 40 research papers and filtering them based on questions we researchers formulated. As a result, we screened 20 studies out of 40 gathered studies. The research we gathered had shown significance towards grammar error detection algorithms which focus on a set of rules used in sentence creation.

VI. Limitation And Future Research

The research limitations will serve as a guide for future researchers on how to approach a similar study. It is recommended for future researchers to conduct more extensive data of corpus as a test subject to provide more accurate data for the system. Apply one more improvement from the model, such as text analysis, so that the system can suggest a whole new, improved grammar. It is for the user to decide if they want to use the enhanced grammar. This study also indicates an enhancement on the system model, such as providing the user's learning material to avoid future errors like grammar structure.

References

[1] Ryo Nagata, Hiroya Takamura and Graham Neubig 2017. Adaptive Spelling Error Correction Models for Learner English. 474-483 (2017), 36-44. DOI: https://doi.org/10.1016/j.procs.2017.08.065

[2] Hongyu Gong, Yuchen Li, Suma Bhat, and Pramod Viswanath. 2019. Context-Sensitive Malicious Spelling Error Correction. (2019. DOI:https://doi.org/10.1145/3308558.3313431

[3] S. Jaf and C. Calder, "Deep Learning for Natural Language Parsing," in IEEE Access, vol. 7, pp. 131363-131373, 2019, DOI: 10.1109/ACCESS.2019.2939687.

[4] Evanini, Keelan & Wang, Xinhao. (2014). Automatic detection of plagiarized spoken responses. 22-27. DOI: 10.3115/v1/W14-1803.

[5] Hong Ma, Manman Qian, The creation and evaluation of a grammar pattern list for the most frequent academic verbs, Volume 58 (2020) DOI: https://doi.org/10.1016/j.esp.2020.01.002.

[6] Chang T, DeJonckheere M, Vydiswaran VGV, Li J, Buis LR, Guetterman TC. Accelerating Mixed Methods Research With Natural Language Processing of Big Text Data. Journal of Mixed Methods Research. 2021;15(3):398-412. doi:10.1177/15586898211021196

[7] Scott C, David R, Kristopher K, Ute R. Applying Natural Language Processing Tools to a Student Academic Writing Corpus: How Large are Disciplinary Differences Across Science and Engineering Fields?. 48 -81 (2017) "Applying Natural Language Processing Tools to a Student Academic Writing" by Scott A. Crossley, David R. Russell et al. (iastate.edu)

[8] Rahimi, Z., Litman, D., Correnti, R. *et al.* Assessing Students' Use of Evidence and Organization in Response-to-Text Writing: Using Natural Language Processing for Rubric-Based Automated Scoring. *Int J Artif Intell Educ* 27, 694–728 (2017). https://doi.org/10.1007/s40593-017-0143-2

[9] Khan Fazlullah, Xiang Li, ZongXun Li. Context-Aware Text Matching Algorithm for Korean Peninsula Language Knowledge Base Based on Density Clustering (2021) DOI: https://doi.org/10.1155/2021/5775146

[10] Ahmed Syed Hassan, Wu Liqin. English Grammar Detection Based on LSTM-CRF Machine Learning Model (2021) DOI: https://doi.org/10.1155/2021/8545686

[11] Wang Wei, Zhao Juan English Grammar Discrimination Training Network Model and Search Filtering (2021). DOI: https://doi.org/10.1155/2021/5528682

[12] Wang Wei, Zhou Shanchun, Liu Wei. English Grammar Error Correction Algorithm Based on Classification Model (2021) DOI: https://doi.org/10.1155/2021/6687337

[13] Jan Mian Ahmad, He Zhenhui. English Grammar Error Detection Using Recurrent Neural Networks (2021) DOI: https://doi.org/10.1155/2021/7058723

[14] Odden, Tor Ole B. Marin, Alessandro. Rudolph, John L. How has Science Education changed over the last 100 years? An analysis using natural language processing (2021) DOI: https://doi.org/10.1002/sce.21623

[15] Tsai, Sang-Bing. Yin, Yiqun Microclassroom Design Based on English Embedded Grammar Compensation Teaching (2021) DOI: https://doi.org/10.1155/2021/6528058

[16] Chen, Chi-Hua. Xu, Yunwei. Research on Business English Translation Architecture Based on Artificial Intelligence Speech Recognition and Edge Computing (2021) DOI: https://doi.org/10.1155/2021/5518868

[17] Su, Jian. Qiu, Haimei. Research on Intelligent Calibration of English Long Sentence Translation Based on Corpus (2021) DOI: https://doi.org/10.1155/2021/5365915

[18] Xu D, Chong SS, Rodenhausen T, Cui H. Resolving "orphaned" non-specific structures using machine learning and natural language processing methods. Biodivers Data J. 2018 August 10;(6):e26659. DOI: 10.3897/BDJ.6.e26659. PMID: 30393454; PMCID: PMC6207837.

[19] Tohti T, Huang J, Hamdulla A, Tan X. Text Filtering through Multi-Pattern Matching: A Case Study of Wu–Manber–Uy on the Language of Uyghur. *Information*. 2019; 10(8):246. https://doi.org/10.3390/info10080246

[20] Salim Keezhatta, Muhammed, Understanding EFL Linguistic Models through Relationship between Natural Language Processing and Artificial Intelligence Applications (January 1, 2020). DOI: http://dx.doi.org/10.2139/ssrn.3512620