Paper No: SC-03 **Smart Computing**

Grammatical error detection and correction model for Sinhala language sentences

H. M. U. Pabasara*
Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
pabaup@gmail.com

S. Jayalal
Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
shantha@kln.ac.lk

***Abstract*: As the national language of Sri Lanka, the greater part of the exercises at most of all the services are completed in Sinhala whereas it is imperative to guarantee the spelling and syntactic accuracy to convey the ideal significance from the perspective of automated materials with the unavailability of resources even though there are enough amount of available materials as hard copy and books. With the high multifaceted nature of the language, it sets aside extensive effort to physically edit the substance of a composed setting. The necessity to overcome this problem has risen numerous years back. But with the complexity of grammar rules in morphologically lavish Sinhala language, the accuracy of the grammar checkers developed so far has been contrastingly lower and thus, to overcome the issue a novel hybrid approach has been introduced. Spell checked Sinhala active sentences being pre-processed, separated nouns and verbs were analyzed with the help of a resourceful part-of- speech-tagger and a morphological analyzer and alongside the sentences were sent through a pattern recognition mechanism to identify its sentence pattern. Then a decision tree-based algorithm has been used to evaluate the verb with the "subject" and output feedback about the correctness of the sentence. To train this decision tree, a dataset consisting of 800 records which included information about 25 predefined grammar rules in Sinhala was used. Finally, the error correction was provided using a machine learning algorithm-based sentence guessing model for the three possible tenses. Conducted research results paved the way to identify the sentence pattern, grammar rules and finally, suggest corrections for identified incorrect grammatical sentences with an acceptable accuracy rate of 88.6 percent which concluded that the proposed hybrid approach was an accurate approach for detecting and correcting grammatical mistakes in Sinhala text.**

***Keywords*:** *Grammar checking, Hybrid approach, Part-of- speech-tagger*

I. INTRODUCTION

Sinhala is used as the first language by approximately 16 million Sinhalese people in Sri Lanka [1] [2], where it is one of the national and official languages, alongside the Tamil language. Communication and correspondence in a majority of state departments, private institutions and other offices are conducted mainly in the Sinhala language to cater to the needs of the end user. A large section of the general population has no clear idea on how to utilize the language accurately [3]. It is essential to utilize the language accurately to safeguard the planned significance where the normal significance can be distinctive because of the inaccuracy of grammar. When thinking about the composed setting, Sinhala is an intricate language that contains many spelling and syntax rules where the rightness of the formal composing thoroughly relies upon these well-characterized rules [4]. It is imperative to guarantee the spelling and syntactic accuracy to convey the ideal significance from the perspective of automated materials with the unavailability of resources even though there are enough amount of available materials as hard copy and books for the Sinhala language. On the other hand, with the high multifaceted nature of the language, it sets aside extensive effort to physically edit the substance of a composed setting. The necessity of an automated system to play out this assignment has risen for the Sinhala language numerous years back. Well-created grammar checking applications are accessible for dialects like English, Tamil and Chinese and plenty of many other languages [5]. But a morphologically rich language like Sinhala lacks precise grammar checking tools mainly because of the unavailability of enough resources [6]. Natural languages like Sinhala are not constrained by a specific syntax and have a wide range of vocabulary; hence, Sinhala grammar checkers must also have an extensive dictionary of most words with their complete meanings and part-of-speech usage [7].

Spell checking, on the other hand, is also one of the key areas to be considered when developing a grammar checker for a resource-poor language. Out of the many types of research that have been carried out in the context, the data-driven open-source approach proposed by Wasala et al. [8] [9], which claims to be able to test and correct spelling errors in the Sinhala sentences is one of the popular. Further attempts in the context like the approach proposed by Jayalatharachchi et al. [10] to obtain an alliance between two algorithms [8] [10] have been then further extended by Subhagya et al. [11] This implies that fair focus has been given for the development of a proper spell checker for the Sinhala language. Construction of relatively inexpensive spell checkers without deep linguistic knowledge for the Sinhala language has been carried out in different dialects with several approaches like n-gram statistics, data-driven approaches while considering the morphological richness of the language and unavailability of enough digital footprint for the language and thus still being able to identify and correct many of the language's common spelling errors with results showing promising output at a reasonable precision rate[8-11]. Availability of such accurate spell checkers for Sinhala language and to shrink the scope of this research study to a feasible level while considering the complexity of the Sinhala grammar rules and sentence structures, this study was conducted to address the knowledge gap in detecting and correcting grammatical mistakes in already spell checked Sinhala written text with the help of possibly available and publicly accessible useful resources while providing a novel

hybrid approach aligned with the machine learning algorithms.

II. BACKGROUND

A. Sinhala language

Sinhala is a language belonging to the globe-spanning language tree, Indo-Aryan and is the native language to approximately 16 million Sinhalese people in Sri Lanka [1] [2]. Also, the language is being used as an auxiliary language approximately by another 4 million people belonging to other ethnic groups living in Sri Lanka [12], where it is one of the national and official languages, along with Tamil. There is also a considerable composition of Sinhala speakers, writers, authors and people who are using the language for many other purposes all over the world. Nevertheless, the closest relative languages to Sinhala existing in the world currently are the Maldives and Dhivehi [13]. The writing system of the Sinhala language is a descendant of the ancient Indian Brahmi script [14] and is thus a member of the Aramaic script family [15] by extension.

B. Grammar checking tools

The natural language processing field is relatively new in the Asian language and plenty of apparatuses are yet to be created. One of these is an accurate grammar checker [16-17]. When all is said and done, rule-based and AI strategies have been utilized for creating sentence structure checkers [18]. Since the field of regular language preparing for Asian dialects is constrained, alongside the intricate syntax, it exhibits a few issues when building up a sentence structure checking framework for the Sinhala language.

Despite the fact that there are syntax checking devices accessible for different dialects, for example, English, the inaccessibility of a technique to check the sentence structure of Sinhala language at constant is as yet not pervasive. The unpredictability of the Sinhala syntax appeared to be the primary explanation behind this. Since a corpus-based vocabulary for the Sinhala language has been executed [19], it can very well be used to helpfully channel meaning and importance from the human language. Engineers and Data Scientists are eligible to prepare and structure complex processes such as relationship extraction, interpretation, grammar sentence structure checking, subject decomposition, etc while using NLP [20], pursuing an information-driven measurement-based methodology considerably at ease. A particular strategy to arrange the finish of the sentence should be actualized to break down a contribution by distinguishing each sentence. Furthermore, the framework should be executed such that it offers proposals for the end clients by recognizing the example of the given information.

III. RELATED WORK

The complex composition of a lot of spelling and grammar rules paves the way for the Sinhala language to be more complicated. All the more, it is difficult to implement all these rules according to a specific order nevertheless, perform all the grammar rules in one application [21].

Even with all these difficulties, there have been very few attempts in developing automated grammar checkers for the Sinhala language within recent decades. “Spell and Grammar Checking Tool for Sinhalese, අකුරු සෝදුව - සියඛස සුඛසක් කරනු රිසිසයෙනි”, developed by Abeyrathne et al. [4]is one such system. The fact that there are already existing systems to facilitate spell checking has been taken into consideration, and thus the system has been developed for providing spell-checking and grammar checking. Even though there are other existing spell checking systems such as spellchecker.net and “ මධුර”, to provide accurate spell checking functionality a spell checker has been implemented in a data-driven approach in “ සුඛස”, which is also already an existing system. A corpus-based Sinhala lexicon has been studied to establish the initial step of spellchecking in this study. A timely concept proposed for morphologically lavish languages like Sinhala, where the words extracted from the corpus have been labeled with parts-of-speech categories defined has been used when conducting the research. With the help of an approach based on n-gram statistics which is a well-known data-driven approach, checking and correcting spelling errors in Sinhala in the study has been conducted which is relatively inexpensive to construct without deep linguistic knowledge [4]. Another related project carried out by the developers: Akila Gunarathna, Pathmasri Ambegoda, Nirasha Katugamapala, Dhanushka Bandara with the guidance of the supervisors: Professor Gihan Dias, Dr. Sanath Jayasena from the Department of Computer Science and Engineering, Faculty of Engineering, University Of Moratuwa has been able to implement an open-source spelling and grammar checking tool for just simple three words sentence, with the name of the project as, “ මහරාවණා” [22]. Since the system has been developed using a rule-based approach, further improvements to that system have become more complicated and thus inevitable due to the heavy number of complex grammar rules, composite noun phrases changing their meaning according to the place of application in a sentence, variety of grammatical features embedded in the language structure, etc. in the Sinhala language [23]. But all these attempts for developing grammar checkers for the Sinhala language have been with the scope of checking grammatical correctness only in simple 3-words Sinhala sentences. Nonetheless, the researches have been only focused on using single approaches or a comparison of a few approaches separately. Also, usage of a higher amount of grammatical features has been at a minimum level in most of all the researches. Yet, the accuracy levels have also been in contrast revolved around lower values.

When considering the other languages similar to Sinhala language Tamil, is also considered as a resource-poor language closely related to the Sinhala language while acting the role of the other official language in Sri Lanka [24].

Nonetheless, because of the presence of larger Tamil speaker populations worldwide, especially in countries like India, a considerable amount of research and tools are available for NLP work related to the Tamil language [25]. Consequently, it is, therefore, reasonable to recognize that Sinhala and Tamil NLP endeavors help assist one another. Additionally, relativity along with alternative mechanisms have been helpful to bridge other languages like Japanese with the Sinhala language [26] [27]. Similarly, in research disbursed for the Kannada language, agreement of noun and verb phrase in those sentences has been sculptured [28]. Noun phrases have been categorized into three subcategories as adjective-noun, noun, and pronoun, but only the gender and the number have been considered as grammatical features of the language. The resemblance of a similar scenario in the

Sinhala language; verb agreement with the subject of the sentences in number and gender. Hence, the verb suffix is extracted to examine masculinity, femininity, and plurality of the verb. In this research along with Python programming language context-free grammar (CFG) has been used to implement the grammar rules [21]. Another related research that has been carried out by Sagar et al. [29] elaborates on the generation of a context-free grammar (CFG) for simple sentences in the Kannada language [29]. Here, both a Top-Down Parser and a Bottom-Up Parser have been used and compared their reliability for the grammar checking process. The main two conflicts have been identified by the authors occurring during the usage of Bottom-Up parser; Shift-Reduce and Reduce-Reduce. Therefore the Top-Down parser has been implicitly selected as the more suitable parser to be used. A Computational Grammar for Bengali has been implemented by Khan and Khan using the Head-Driven Phrase Structure Grammar (HPSG) formalism [30].

A process synchronic linguistics for a language may be a possibly helpful resource for polishing off varied language process activities like testing synchronic linguistics, machine translation, and question respondent. Similar to the situation in most South Indian languages [26] [31], among alternative grammatical options, Sinhala may be an extremely incurved language with 3 gender forms and 2 range forms [32]. To develop a computational model, it is required to understand the language well. Research has been conducted to describe the event of a feature-based CFG for non-trivial sentences in Sinhala [21]. Originally, ensuing synchronic linguistics covers a big set of Sinhala as represented by very popular synchronic linguistics hard copies [23] [32]. To manufacture the suitable, dissect tree(s) related to sentences input to the system, a parser has been conjointly implemented victimizing the NLTK toolkit during this research. Generally, the synchronic linguistics conjointly detects then rejects ungrammatical sentences during the conduct. 200 sample sentences have been acquired via Sinhala grammar textbooks of primary grades for testing purposes in the study. 10 sentence structures have been selected and have been accustomed to style the synchronic linguistics. The accuracy of the process has been limited to 60% according to test subjects. Implementation of a computational model of grammar for the Sinhala language has been considered in another research carried out by Hettige and Karunananda [33]. Usage of morphological and grammar analysis of Sinhala has been thought of during this research study to shapely employ Context-Free descriptive linguistics and a Finite State electrical device (FST) that also handles solely straightforward sentences that contain eight constituents [33]. Computational models for grammatical error detection have been evolved from simple rule-based models to deep learning models [34]. Statistical models have shown to perform well in many natural language understanding tasks [35] [36]. Applying statistics requires a clear understanding of the language and the language structure [37]. However, this requires a large amount of knowledge about the language, which can be hard to achieve on languages that are not popular, or which have complex structures like Sinhala [38].

The Sinhala language has a limited digital footprint when compared to other popular languages. Therefore, using Sinhala for natural language processing tasks are harder [35]. Natural language processing tasks for the Sinhala language has been conducted using many techniques [39]. The rule-based approach was a very common approach used in simple Sinhala natural language tasks [40]. The statistical models show the potential in achieving better natural language models for Sinhala. With technological improvement, most recent developments in NLP related researches have been, using machine learning algorithms [41-43]. These models have been used in many industrial language tasks since the models are better. However, statistical models fail to capture all features of language models [37]. Machine learning algorithms do not require any feature selections but are capable of achieving higher accurate results compared to statistical methods and rule-based models [41] [43-44]. Machine learning algorithms have been applied as a comparison technique for grammatical error detection in very few researches so far [4]. A machine learning algorithm-based model, however, requires a larger dataset to learn. Therefore, languages like Sinhala with smaller digital data presence fail in producing better results while only using a machine learning algorithm-based approach [41]. Thus, the necessity of a novel approach arises to earn more accurate and efficient results in the context of grammar mistake detection and suggestion of corrections for a morphologically lavish language like Sinhala.

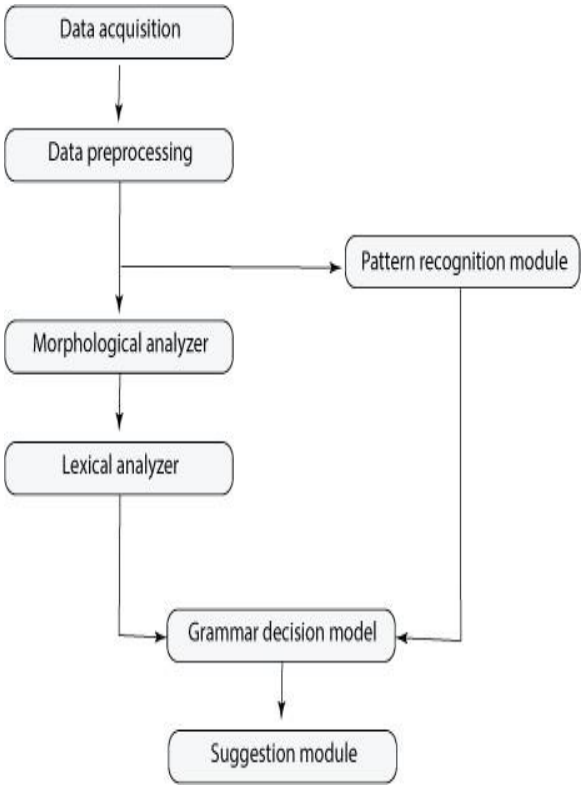

Fig. 1. Proposed and implemented a methodology for the study

IV. METHODOLOGY

The commencement via a thorough literature review of the research area deprived that applying only one approach has been unable to output with more accurate results and thus, this study has been carried out using a novel hybrid approach while combining traditional rule-based approach along with the machine learning algorithm-based approach, for grammatical error detection and correction in Sinhala text to identify the possibility of having a high potential of achieving comparative results. Since a proper lexical analyzer has not been implemented yet for the Sinhala language at present, engaging with only a rule-based approach could be inappropriate when considering the lack of possible resources, unlike other resource-available languages. Thus when examining most appropriate approaches for detecting and correcting grammatical mistakes for morphologically lavish languages like Sinhala, the severe complexity of the language, difficulty of defining a fixed set of grammar rules due to containing a lot of spelling and grammar rules and availability of enough resources for implementing such approaches were taken into consideration in this research.

The study has focused on detecting the grammatical mistakes in active sentences in Sinhala text where sentences in written text format were checked sentence-wise and word-wise respectively and the nouns and verbs were separately analyzed with the help of a resourceful part-of-speech (pos) tagger [7] and a morphological analyzer [45]. To uplift the performance of the grammar checker, both the pos tagger and the morphological analyzer used in this study have been further modified accordingly to upgrade the error detection and correction mechanism to uplift the available Sinhala

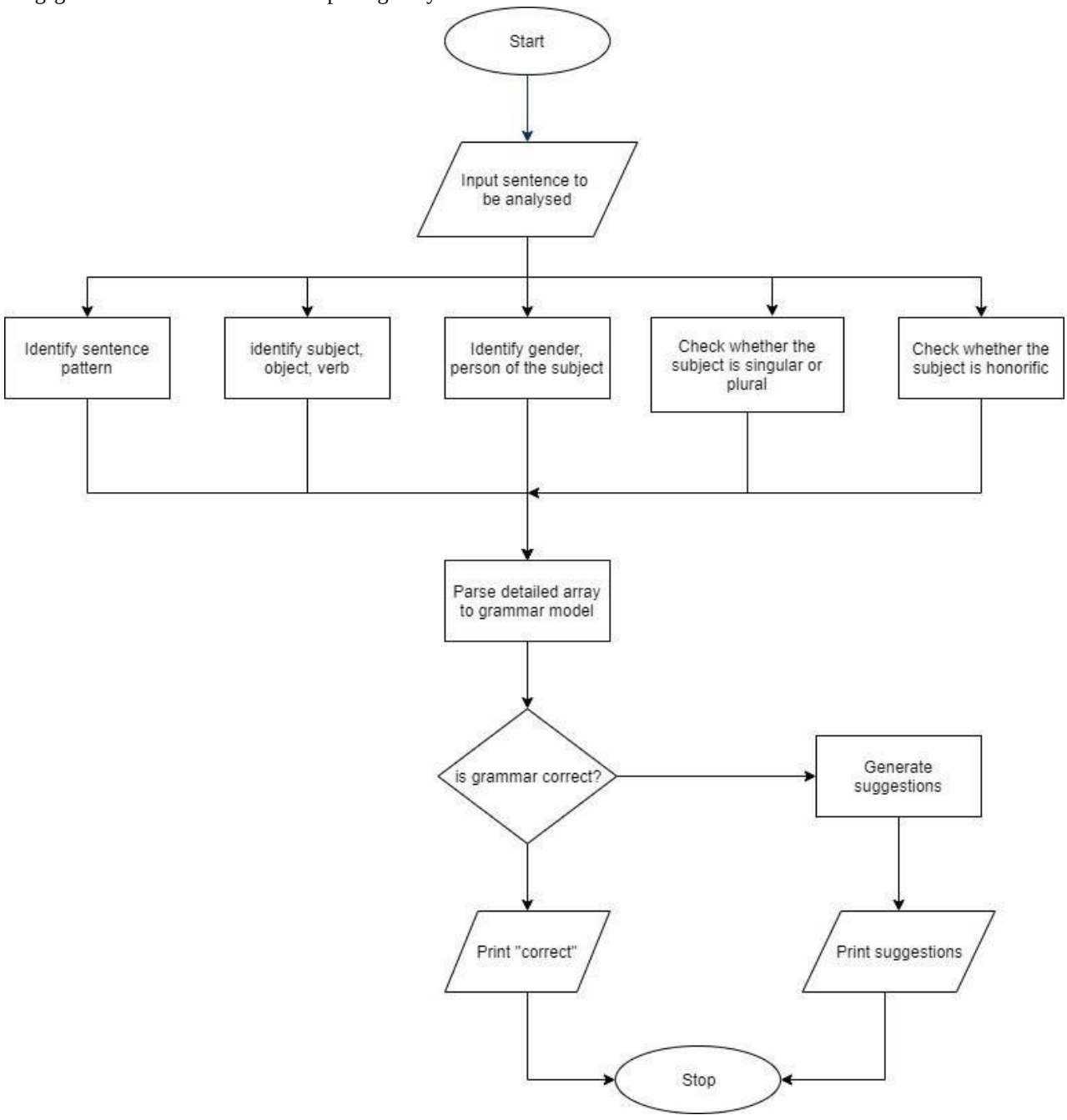

Fig. 2. The algorithm used for grammar mistake detection and suggestion of corrections

digital footprint to a higher level. Parallel to this phase, the given sentence was then sent through a pattern recognition mechanism that was developed to identify the sentence pattern of the given sentence.

Finally, the system which comprised of a hybrid novel approach combining a system developed using traditional rule-based approach and decision-tree; a machine learning algorithm-based model was able to detect the correctness of the given sentence and provide corrections for erroneous grammatical sentences. Specifically, here the consideration has only been on already spell checked sentences [46] for the detection for grammatical mistakes which helped compress the scope to a feasible level and provide more accuracy for the study. Fig.1 depicts the main approach which has been implemented in this study, and it has been further elaborated below.

A. Data acquisition

A data set of around 1050 Sinhala active sentences with no spelling errors without a specific number of word-range were gathered from the web using online Sinhala newspapers via Python scripting and they were categorized according to their tenses and grammatical person. Spelling check has already been done before with the help of an existing accurate spell checker for the Sinhala language; "සුබස | අකුරු විනිස පිරික්සුම" [46]. This data set was then thoroughly validated with the help of a group of Sinhala language experts from different dialects and a data set containing around 800 sentences with about 25 pre-defined sets of grammar rules and about 32 sentence patterns were finally separated according to their grammatical feature and those were recorded and used as the training dataset.

For further testing purposes, the remaining 250 sentences were used and they were then pre-processed using NLTK toolkit and Python programming language and a data set containing around 180 correct grammatical sentences and around 60 incorrect grammatical sentences were classified while eliminating 10 unidentified sentences due to lack of proper sentence structure in those sentences.

B. Data pre-processing

Separation of sentences in the paragraph, separation of words from spaces in a sentence, removing unwanted characters from the sentences were been carried out to pre-process the data input into the system.

C. Morphological and lexical analysis

Then the identification of subject and verb through the already pre-processed sentences were been carried out using an existing pos tag system [7] and the tag of each word was identified and stored separately with the corresponding word for further reference. A comprehensive, multi-level pos tag set for Sinhala was being used here which has been developed by the Center for National Language Processing, University of Moratuwa while considering its accuracy level [7]. This pos tagger has been further modified in this study, to increase the performance of the grammar checker for example with newer sub-tags for the gender of the common nouns used in the available pos tagger. Then a morphological analyzer [46] was used to separate the root and the suffix from the identified verb and separated root and suffix of the verb was further carried over to the next phase with the corresponding noun acquired from the above lexical analysis. Due to lack of access to most of the accurate morphological analyzers for Sinhala [16], on the notion that the above mentioned morphological analyzers work, another existing considerably accurate morphological analyzer which is commonly known as "polyglot" [45] has been used with further modifications to increase the performance of the grammar checker accordingly in this study which is a natural language pipeline that supports the Sinhala language in this study. Further modifications were necessary for the morphological analyzer because even though it provided accurate noun breakdowns, it provided several suffixes for the verb giving ambiguity when selecting the correct suffix to be concatenated to the verb root for a given verb in a correct grammatical sentence.

D. Pattern recognition module

Alongside the above process been carried out, the given sentences were then passed through a pattern-recognition mechanism that has been developed using a regular expression based algorithm, to identify the sentence pattern of the sentence input into the system. Around 32 different sentence structures have been identified and used in the pattern-recognition mechanism to detect grammatical mistakes in a variety of sentence patterns to expand the error detection mechanism for a higher possible range of sentence patterns.

E. Grammar decision model

Grammar rule-based model developed for identifying the grammar rules and sentence pattern-recognition mechanism for identification of sentence structures developed in this research that was used for the training purpose of the model was created with the thorough referencing through many Sinhala hard copies and literacy books, and then the produced data set was cross-checked with the help of a panel of experts consisting Sinhala language experts from several divisions including Sinhala teachers from different schools, language experts from different dialects. The finalized data set was then further directed to a set of language experts from the Department of Sinhala, Faculty of Humanities at the University of Kelaniya. All these processes were carried out with the comprehensive usage of the Delphi method [47] to acquire an accurate and non-bias data set to accomplish the tasks carried out in this research study.

F. Novel hybrid approach combining rule-based and machine learning algorithms (Suggestion module)

To elaborate in detail, the general flow once a sentence was entered into the system has been, first, each word was assigned with a tag by the pos tagger.

Eg: - In the sentence, "මම ගමට ගියෙමි", it worked as below.

```
{
"sentence": "කමල් සමග මම පාඩම් කරම්",
"is_correct": false,
"suggesions": {
   "present": "කමල් සමග මම පාඩම් කරමු",
   "future": "කමල් සමග මම පාඩම් කරන්නෙමු",
   "past": "කමල් සමග මම පාඩම් කළෙමු"
    }
}
```

Fig. 3. Machine-learning algorithm and grammar rule-based corrected sentence suggestion

- Pronoun -> මම - PRP
- Verb -> ගියෙමි - VFM
- Common Noun -> ගමට - NNC

The array which consisted above details were then passed through the pattern-recognition mechanism to identify its sentence pattern. According to the identified sentence pattern, lexical analyzed data and morphologically analyzed verb, the grammatical error detection were then carried out with the help of the developed grammar-rule data set using a decision-tree classification machine learning algorithm. This has been used to evaluate the verb with the "subject" and output feedback about the correctness of the sentence. Fig.2.shows the flow chart for the algorithm established in the study. The system has considered around 25 predefined sets of grammar rules in this phase while considering the performance as a key aspect. Accordingly, for grammatically incorrect sentences, the system has predicted suggestions using the decision-tree classification-based machine learning algorithm [43] for all three tenses. Fig.4.shows a sample data set that has been used for the decision tree algorithm in this study. Then the suffix-engine would try to concatenate given suffix and root of the verb and create a meaningful sentence to the user while finally providing the user with a correctly spelled and grammatically sound sentence.

V. Results and Discussion

	subject	tense	person	gender	animate	number	verb_root	is_correct
2	මම	අතීත	උත්තම	නොමැත	ප්‍රාණවාචී	ඒක	එමි	1
3	මම	අතීත	උත්තම	නොමැත	ප්‍රාණවාචී	ඒක	ඒය	0
4	මම	අතීත	උත්තම	නොමැත	ප්‍රාණවාචී	ඒක	එමූ	0
5	අපි	වර්තමාන	උත්තම	නොමැත	ප්‍රාණවාචී	බහු	අමූ	1
6	අපි	වර්තමාන	උත්තම	නොමැත	ප්‍රාණවාචී	බහු	අමි	0

Fig. 4. Sample data set used for the decision tree algorithm in the study

The data set consisted of around 800 records that were used to train the decision tree in the grammar model in the study was acquired with the thorough referencing through many Sinhala hard copies and literacy books [23] [32] [48-49], and then cross-checked with the help of a panel of experts consisting Sinhala language experts from different dialects including Sinhala teachers of different schools and finalized it by further directing it to a panel of language experts with the comprehensive usage of the Delphi method [47] to acquire an accurate and non-bias data set. This data set included data about 25 predefined grammar rules in Sinhala. It consisted variety of grammatical features including person, tense, gender, number, honorific feature, animate feature, etc. This data set also comprised of a broad range of subjects from most common pronouns to deterministic pronouns and even extended to several uncommon proper nouns as well. This trained model then gave us an accuracy score of 88.6%.

A system identified incorrect grammar sentences were automatically passed to the decision-tree classification-based machine learning sentence correction algorithm to suggest correct sentences. It then suggested a suitable verb root and then added it to a matching verb suffix and gave the output as shown in Fig.3.

The data set used for further testing the developed system consisted of around 240 sentences with a non-bias set of sentences with both correct and incorrect grammar. To test the performance of the classification model a confusion matrix was used and the results obtained have been elaborated in Fig.5 below.

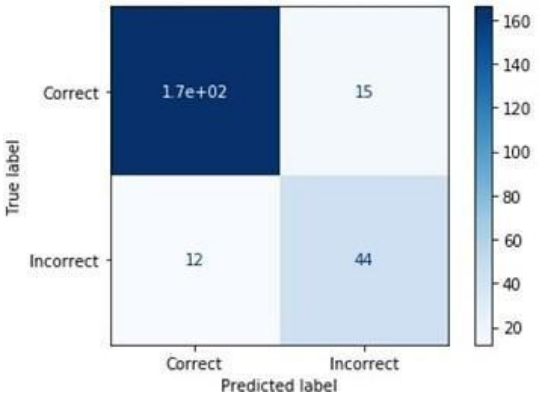

Fig. 5. Confusion matrix obtained to analyze the performance of the developed model

During the testing phase system gave the following results.

Input correct sentences = 181, System identify only 166 sentences

Input incorrect sentences = 56, System identify only 44 sentences.

- Accuracy = [(166+44)/237]* 100% = 88.6%
- Recall = [166/ (166+15)]*100% = 91.71%
- Precision = [166/ (166+12)]*100% = 93.25%

F-measure=[(2*0.9171*0.9325)/ (0.9171+0.9325)]*100%

= 92.47%

According to the above-calculated results that were obtained using the confusion matrix, the model has given higher recall, higher precision with a considerably high amount of accuracy for the given test cases. Nonetheless, F-measure which calculates the harmonic mean instead of arithmetic mean has precisely output with an accurate result using both the precision and recall values. Hence, a considerable amount of accurate results were achieved through the proposed and implemented a novel hybrid approach with machine learning techniques.

Most of the researches that have been carried out so far in the context of grammatical error detection for the Sinhala language have been only using 3-words sentences and have only covered the scope of error detection for selected features in the Sinhala language like number, gender, tense and person where features like volition has been omitted in most of all the scenarios. One important aspect that has been achieved in this study was the elimination of the limitation of grammar checking for only 3-words sentences. This has been upgraded to more than 3-words sentences in this study. Nonetheless, in this research as special accomplishments, sentences consisting of post-positions, conjunctions like, "ද, හා, නි, සහ", grammatical features like volition, deterministic pronouns like, "අයෙක්, කෙනෙක්, ඇතැමෙක්" and few proper nouns were taken into consideration. To elaborate more on the

study, given the sentence, "මම ගෙදර කමු", here the word gives a meaning of "I eat the house", which is semantically incorrect. But, the algorithm suggested in this study could escape this problem and the suggestion of meaningful sentences was also achieved as a secondary output of the study. So, this algorithm could be used to suggest a correct verb in any given word processing software.

Even though the main intention of this study was to propose and develop a grammatical error detection and correction mechanism using a novel hybrid approach, this research study was able to contribute to further developing and modifying the digital footprint for the Sinhala language as well. While accomplishing the task of grammar detection and correction, necessary changes and modifications were done accordingly for the existing pos tagger [7] and the morphological analyzer [45] to contribute to future researches that would be carried out on resource-poor Sinhala language.

VI. Conclusion

This research has been carried out to introduce a hybrid approach for identifying grammatical errors and suggesting corrections for the Sinhala written sentences and to identify whether there is a possibility of having a high potential of achieving comparative results compared to other relative approaches. With the gained results through the study, it was able to identify the sentence pattern in the given sentence and then identify its grammar rules. Further, using machine learning algorithms it was able to suggest corrections for the sentences identified as incorrect grammatical sentences with an acceptable accuracy rate. Therefore with all the achieved results from this study, it can be concluded that the hybrid approach is an efficient and accurate approach for detecting and correcting grammatical mistakes in written text formats for the Sinhala language. Further, the research work carried out in the context confides that the proposed novel hybrid approach could be possibly extended towards other emerging technological aspects as well with better promising results.

VII. Future Work

As suggestions for future work, the dataset must be developed and optimized further to identify more errors mostly in third person format and also the sentence pattern-recognition model could be further optimized to identify more sentence patterns to increase the performance of the established system. For this purpose, using a Deep leaning based extension for the used rule-based pattern recognition algorithm might be a success. To identify the gender of a given third-person noun, the rule-based approach was used in this study. But as third-person format consists of huge variety and number of words, using a separate dictionary for this is also not a practical solution. Therefore, a strong recommendation is suggested to focus more on overcoming this issue in future work. The existing morphological analyzer was used after a few modifications to overcome the ambiguity issue occurred when separating some verb-roots. Hence, a further modification for the existing morphological analyzer or development of a proper and accurate morphological analyzer is necessary for the future to uplift the Sinhala language-based research.

Also for future implementations, it can be possibly recommended to use emerging technologies like neural network systems and try out statistical models like mathematical model text classification to compare the accuracy of the hybrid system implemented in this research study which was developed by combining rule-based and machine learning algorithms to improve the development of Sinhala language-based researches in the future.

Acknowledgment

The authors would like to extend their heartfelt gratitude and acknowledgment to all the language experts, Sinhala teachers from different schools and especially, the Sinhala Department in the University of Kelaniya for their immense support and encouragement they gave throughout the development phase of the data sets.

References

[1] Statistics.gov.lk. (2017). Census of population and housing 2011. [on - line] Available at: http://www.statistics.gov.lk/PopHouSat/CPH2011/ in- dex.php?fileName=pop42gp=Activitiestpl=3 [Accessed 6 Jun. 2019].

[2] Ethnologue. (2017). Sinhala. [online] Available at: https://www.ethnologue.com/language/sin [Accessed 16 Jul. 2019].

[3] Anon, (2020). [online] Available at: https://groundviews.org/2011/10/24/reflections-on-issues-of-language- in-sri-lanka-power-exclusion-and-inclusion/ [Accessed 13 Nov. 2019].

[4] L. Abeyrathne, S. Edirisinghe, R. Premachandra, A. Warsha, N. De Silva and S. Thelijjagoda, "Spell and grammar checking tool for Sinhalese: අකුරු සෝදුව - සියඛස සුඛසක් කරනු රිසිසයෙනි", 2018.

[5] I. Wijesiri, M. Gallage, B. Gunathilaka, M. Lakjeewa, D. Wimalasuriya, G. Dias, R. Paranavithana, and N. De Silva, "Building a wordnet for Sinhala," in Proceedings of the Seventh Global WordNet Conference, 2014, pp. 100–108.

[6] B. Hettige, A.S. Karunananda, "A morphological analyzer to enable English to Sinhala machine translation," in Proceedings of the 2nd International Conference on Information and Automation (ICIA2006), Colombo, Sri Lanka.

[7] S. Fernando, S. Ranathunga, S. Jayasena, G. Dias, "Comprehensive part-of-speech tag set and SVM based POS tagger for Sinhala," In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016) (pp. 173-182) (2016, December).

[8] A. Wasala, R. Weerasinghe, R. Pushpananda, C. Liyanage, and E. Jayalatharachchi, "A data-driven approach to checking and correcting spelling errors in Sinhala," Int. J. Adv. ICT Emerg. Reg, vol. 3, no. 01, 2010.

[9] R. A. Wasala, R. Weerasinghe, R. Pushpananda, C. Liyanage and E. Jayalatharachchi, "An open-source data driven spell checker for Sinhala," International Journal on Advances in ICT for Emerging Regions (ICTer), 3(1), pp.11–24, 2011.

[10] E. Jayalatharachchi, A. Wasala, and R. Weerasinghe, "Data -driven spell checking: the synergy of two algorithms for spelling error detection and correction," in International Conference on Advances in ICT for Emerging Regions (ICTer2012). IEEE, 2012, pp. 7–13.

[11] L. G. B. Subhagya, L. Ranathunga, W. H. A. Nimasha, B. R. Jayawickrama, and K. L. Mahaliyanaarchchi, "Data-driven approach to Sinhala spellchecker and correction," in 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE, 2018, pp. 01–06.

[12] Statistics.gov.lk. (2013). [online] Available at: http://www.statistics.gov.lk/PopHouSat/PDF/Population/p9p11%20S pe aking.pdf [Accessed 3 Jun. 2019].

[13] N. Mohamed, (2005). Note on the early history of the Maldives. Archipel, Vol. 70(1), pp.7-14.

[14] P.T. Daniels and W. Bright, "The world's writing systems," Oxford University Press on Demand, 1996.

[15] R. Salomon, "Indian epigraphy: a guide to the study of inscriptions in Sanskrit, Prakrit, and the other Indo-Aryan languages," Oxford University Press, 1998.

[16] N. De Silva, "Sinhala text classification: observations from the perspective of a resource poor language," 2015.

[17] V. Verma and S. Sharma, "Comparative analysis of grammar checkers of various Asian languages," International Journal of Computer Sciences and Engineering, Vol 6(10), pp.697-700, 2018.

[18] A.Vernon, "Computerized grammar checkers 2000: capabilities, limitations, and pedagogical possibilities," Computers and Composition, Vol 17(3), pp.329-349, 2000.

[19] A.R.Weerasinghe, Dulip Herath, Viraj Welgama, "A corpus-based Sinhala lexicon," In Proceedings of the 7th Workshop on Asian Language Resources, Singapore, Aug 2009.

[20] R. Sangal, S. Bendre and U. Singh, (2003). Recent advances in natural language processing. Mysore: Central Institute of Indian Languages.

[21] C. Liyanage, R. Pushpananda, D.L. Herath and R. Weerasinghe, "A computational grammar of Sinhala," in International Conference on Intelligent Text Processing and Computational Linguistics, pp. 188–200, Springer, 2012.

[22] Findbestopensource.com. Maharavana - Spelling and grammar checker for Sinhala. [online] Available at: https://www.findbestopensource.com/product/maharavana [Accessed 21 May 2019].

[23] J. Gair and J. Paolillo, " Sinhala," Munchen: LINCOM Europa, 1997

[24] J. Gair and B. Lust, "Studies in South Asian linguistics," New York: Oxford University Press, 1998.

[25] N. Saharia, S. Sharma and J. Kalita, "Stemming resource-poor Indian languages", ACM Transactions on Asian Language Information Processing, Vol 13(3), pp.1-26, 2014.

[26] A. Herath, Y. Hyodo, Y. Kawada, T. Ikeda, and S. Herath, "A practical machine translation system from Japanese to modern Sinhalese", Gifu University, pp. 153–162, 1994.

[27] A.B. Kanduboda, "The role of animacy in determining noun phrase cases in the Sinhalese and Japanese languages," Science of words, vol. 24, pp. 5–20, 2011.

[28] B.M. Sagar, G. Shobha, and Ramakanth Kumar, "Solving the noun phrase and verb phrase agreement in Kannada sentences," International Journal of Computer Theory and Engineering, Vol. 1, No. 3, August 2009.

[29] B.M. Sagar, G. Shobha and Ramakanth Kumar, "Context Free Grammar (CFG) analysis for simple Kannada sentences," in Proceedings of the International Conference [ACCTA-2010] on Special Issue of IJCCT, Vol. 1 Issue 2, 3, 4., 2010.

[30] Naira Khan and Mumit Khan, "Developing a computational grammar for Bengali using the HPSG formalism," in Proceedings of the 9th International Conference on Computer and Information Technology (ICCIT '06), 2006.

[31] A. Grant, "Studies in South Asian linguistics: Sinhala and other South Asian languages (review)," Language, Vol 77(3), pp.639-639, 2001.

[32] D. Chandralal, "Sinhala," Amsterdam, the Netherlands: John Benjamins Pub. Co, 2010.

[33] B. Hettige, A. Karunananda, "Computational model of grammar for English to Sinhala machine translation," in Proceedings of the International Conference on Advances in ICT for Emerging Regions, 2011.

[34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-444, 2015.

[35] K. Senevirathne, N. Attanayake, A. Dhananjanie, W. Weragoda, A. Nugaliyadde, and S. Thelijjagoda, "Conditional random fields based named entity recognition for Sinhala," in IEEE 10th International Conference on Industrial and Information Systems (ICIIS), 2015, pp. 302- 307.

[36] E. Cambria and B. White, "Jumping NLP curves: a review of natural language processing research [review article]," IEEE Computational Intelligence Magazine, vol. 9, pp. 48-57, 2014.

[37] C. D. Manning, C. D. Manning, and H. Schu¨ tze, "Foundations of statistical natural language processing," MIT Press, 1999.

[38] J. Jayakody, T. Gamlath, W. Lasantha, K. Premachandra, A. Nugaliyadde, and Y. Mallawarachchi, "Mahoshadha," the Sinhala tagged corpus based question answering system," in Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1, 2016, pp. 313-322.

[39] S. Herath, S. Ishizaki, T. Ikeda, Y. Anzai and H. Aiso, "Machine processing of Sinhala natural language: a step toward intelligent systems," Cybernetics and Systems, 22(3), pp.331-348, 1991.

[40] P. Antony, "Machine translation approaches and survey for Indian languages," International Journal of Computational Linguistics Chinese Language Processing, Volume 18, Number 1, March 2013, vol. 18, 2013.

[41] N. Madi, R. Al-Matham and H. Al-Khalifa, "Grammar checking and relation extraction in text: approaches, techniques and open challenges," Data Technologies and Applications, 53(3), pp.373-394, 2019.

[42] J. Hirschberg and C. D. Manning, "Advances in natural language processing," Science, vol. 349, pp. 261-266, 2015.

[43] "Machine learning algorithm for learning Natural Languages," International Journal of Managerial Studies and Research, Vol 4(2), 2016.

[44] J. Weston, S. Chopra, and A. Bordes, "Memory networks," arXiv preprint arXiv: 1410.3916, 2014.

[45] Polyglot.readthedocs.io. (2019). Welcome to polyglot's documentation! Polyglot 16.07.04 documentation. [online] Available at: https://polyglot.readthedocs.io/en/latest/index.html [Accessed 17 Sep. 2019].

[46] Speller.subasa.lk.(2013) [online] Available at: http://speller.subasa.lk/spellerweb.py [Accessed 14 Nov. 2019].

[47] W. Weaver, "The Delphi method," Syracuse, N.Y.: Educational Policy Research Center, Syracuse University Research Corp, 1970.

[48] K. Jayathilake, "Nuthana Sinhala Vyakaranaye Mul Potha," Pradeepa Publications, 34/34, Lawyers' Office Complex, Colombo 12, (1991)

[49] A. M. Gunasekera, "A Comprehensive Grammar of the Sinhalese Language," New Delhi, India: AES Reprint, 1986.