Integrated Model to Develop Grammar Checker for Afaan Oromo using Morphological Analysis: A Rule-based Approach

Jemal Abate[1]
Department of Information Science
Haramaya University, Haramaya, Ethiopia

Vijayshri Khedkar[2]*, Sonali Kothari Tidke[3]
Symbiosis Institute of Technology, Symbiosis International
(Deemed University), Pune, Maharashtra, India

***Abstract*—This study has implemented a rule-based approach on grammar checkers by integrating a spell-checker with a morphological analyzer to improve the Afaan Oromo grammar checker. A corpus containing about 300,000 words has been prepared to be used for spell-checker. About 300 grammar rules are constructed to detect the grammar error within the Afaan Oromo text and to suggest the possible grammar correction. The developed frameworks have experimented on the document having pairs of 100 correct and incorrect sentences. The experimental result for checking the spelling errors has scored 73% of recall, 76% precision, and 75% of F-measure. The score for suggesting the correct spelling is 78% of recall, 62% precision, and 70% precision F-measure while the evaluation result for detecting the grammar errors has 47% recall, 90% precision and 68% f-measure score. For suggesting the possible correct grammar on the detected error, the system has scored 61% recall, 71% precision and 66% f-measure. The overall performance of the developed system has a good performance. However, there is still a need to conduct further research to improve the Afaan Oromo grammar checker.**

Keywords*—Grammar checker; spell checker; part-of-speech tag; error detection; syntactic analysis; semantic analysis; morphological analyzer; NLP

I. Introduction

For communication, a natural language is a language used by humans. Natural language processing (NLP) is a field of study mainly concerned with the communications between human languages and the computer. During a written communication, people may make an error which can be a spelling error and/or a grammar error. A spelling error is an error if the given the word does not exist in the language's vocabulary, whereas a grammatical error occurs when the written sentence is not as per the grammar rule of a given language [1].

With globalization, many individuals work on National and International languages for everyday communications. But it is difficult for many to write various contents using correct spelling and grammar, whether professional communication or a regular discussion, which makes grammar checking one important tool in the word processing software. Grammar checking has significance for having a good flow of ideas, exactness and quality of the expressed written content [1]. Grammar checking is one of the activities in the natural language's written communication and NLP application. Grammar checking checks the grammatical errors in text and suggests possible corrections in many cases. This is one of the most frequently used tools in language engineering.

One of the main challenges in the application of grammar checking for natural language is it is language-specific. Due to this, the grammar checker that works for one language may not work for another language. Therefore, it's necessary to design a grammar checker for each language as per the grammatical rule of a particular language.

Afaan Oromo is one of the most widely spoken languages in east Africa, which accounts for about 40 million speakers in Ethiopia; it is also spoken in neighboring Ethiopian countries like Kenya, Somalia, Djibouti and Egypt [2].

Today, Afaan Oromo is an official working language of Oromia national, regional, and educational languages for primary school (1-8) students in the Oromia region. As a result, many works for official and/or personal purposes use Afaan Oromo for interpersonal communication. Most of the time, while preparing various reports, peoples are vulnerable to make a grammatically wrong statement unintentionally or unknowingly. Some languages, such as English, French, etc., have a tool to catch and make corrections on these problems. However, Afaan Oromo is one of the under-resourced languages. Due to this, there are no tools that can assist the users in catching and making corrections on the grammatically wrong statements. Therefore, it is required to develop and implement the Afaan Oromo grammar checker to eliminate these problems. Developing a grammar checker for Afaan Oromo is significant for non-writers and nontechnical peoples to have accurate and quality written content. This study is intended to design and develop an integrated model to develop a grammar checker for Afaan Oromo using morphological analysis – A Rule-Based Approach.

II. Grammar Checker Approaches

There are three widely used approaches by various researchers like Syntax-based approach, statistical-based approach and rule-based checking [3].

- Syntax-based approach: In this approach, text parsing is used by assigning a tree structure to the sentence. If the text parsing mismatches, the statement is labeled as a grammatically wrong sentence; otherwise, it is considered correct.

*Corresponding Author

- Statistical-based approach: In this approach, a corpus annotated with its part of the speech tag is prepared and used to build a sequence of parts of the speech tag list. If the frequency of the tagged text is less than with the already trained model, this text is considered incorrect otherwise it's it is considered correct.
- A rule-based approach: In this approach, a set of grammar rules has been manually prepared to match against a text. If the text matches the crafted rules, it is considered a correct otherwise wrong sentence.

III. Related Works

Various studies have been conducted on natural language processing for the Afaan Oromo language. For instance, [4] have designed a spell checker for non-word Afaan Oromo spell checker. This work is focused on checking misspelled words and use the proposed algorithm to observe if the right word is generated under suggestion. But authors are not considering the context of the statement in their work.

The author in [5] attempted to design a grammar checker for Afaan Oromo using a rule-based approach. The author constructed about 123 rules for the detection of the Afaan Oromo grammar error. The rules are constructed depending on the language rule of affixation. However, false alarms in the developed framework lead the system to become challenged to catch the grammar error correctly. The author in [5] has mentioned some of the reasons for the problem that occurred. These are some words root and affix are identified wrongly by the stammer, a wrongly assigned part-of-speech tag and incompleteness of the rules. However, there are many reasons for the problems with [5] framework in addition to the reasons mentioned. Rules ranging from 81-86 were constructed to catch the error of past perfect tense. But this system may falsely alarm the correct grammar as wrong due to incorrect rules. Let's have a look at the following example, which tests his rule. One of the rules which are crafted by [5] says, "If the subject of the sentence is third-person singular masculine, so the verb must end with the suffix -ee, and the sentence must end with ture."

Example: Callisee Bira dabruu yaalee ture. (He tried to pass in silence).

But the above rule may flag other correct statements as a grammatically wrong sentence. Because, in Afaan Oromo, if the subject of the sentence is third-person singular masculine, but the verb may end with the suffix other than -ee and also it's not necessary for the sentence to end with 'ture'."

Example: Callisee Bira dabruu yaalus hin milkoofne. (He tried to pass in silence, but he did not).

The author in [6] has developed the Afaan Oromo grammar checker by using a statistical approach. The author has implemented two statistical approaches: the token n-gram frequency whose probability of sequence is calculated and the tag n-gram frequency whose POS tag is assigned to all tokens, and the probability of the tag sequence is calculated for training and testing the checker. However, [6] didn't incorporate the checking of spelling errors in his work which is fundamental to the work of grammar checkers. The corpus used for training contains only 10000 words due to this POS tagger, and the Morphological analyzer has become inaccurate. Additionally, the article didn't highlight constructing a rule that will be used for training the model and improving the model performance. As a result of the factors mentioned, the work proposed in the paper has to be improved by incorporating spell checking along with a good morphological analyzer.

Even if few works have been done on grammar checkers for the Afaan Oromo language, there is still a gap in developing a good grammar corrector that can help to fix common spelling errors, instances of the passive voice, clunky and hard to understand language.

IV. Material and Methods

A. Material Used

Various development tools have been used to design the Afaan Oromo grammar checker. Python is one of the tools that were used to develop the prototype. Python is used because it's easy to work with, learn and adaptable scripting language, making it attractive for development [7]. HORN MORPHO 2.5 is used for POS tagging activity. It is a program that analyzes Amharic, AO, and Tigrinya words into their constituent morphemes (meaningful parts) and generates words, given a root or stem and a representation of the word's grammatical structure [8]. The proposed research has also used the Hunspell dictionary to store the required grammar correction rules and integrate the developed framework into Apache OpenOffice.

B. Design Approach

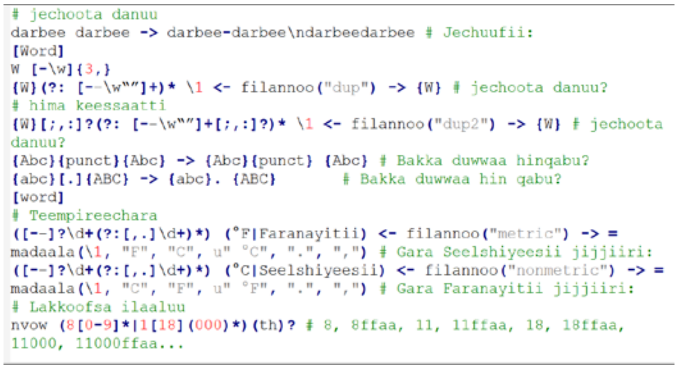
There are quite a lot of ways in which diverse approaches to grammar checking can be illustrated. This study intended to develop a grammar checker for Afaan Oromo using a rule-based approach by integrating with spell-checker and morphological analyzer. The dictionary containing about 300,000 unique words is constructed that can help to detect spelling errors. For grammar checking, a set of grammar rules has been constructed, which accounts for about 250 rules that check the errors in the grammar.

To perform the grammar checking, the developed system will find a word from the sentence. Then the tag of the word is checked against the developed grammar rules to detect various errors such as style problems, agreement in gender, number, word order, etc. If the sentence contains an error, the suggestion to the error will be displayed. The designed framework has two components, namely the spell-checker component and grammar checker component.

1) Spell-checker components: The spell-checker components have a set of words and affix rules of Afaan Oromo languages. This component helps to catch the spelling errors from the texts and suggests the possible correction to the wrong words depending on the morphological rules of the Afaan Oromo languages.

2) Grammar checker components: This module consists of a set of grammatical rules of the Afaan Oromo languages.

This component works by checking the arrangements of words in a given sentence. The crafted rules check the syntactic agreement features in the statement, i.e., the agreement in the gender (Masculine or Feminine), number (Singular or Plural), a person (1st/2nd/3rd person singular, 1st/2nd/3rd person plural) and proofreading (punctuation). In Fig. 1, some of the sample rules are constructed.



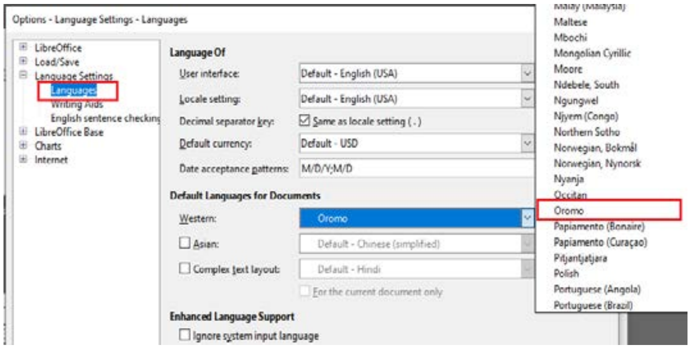
```
# jechoota danuu
darbee darbee -> darbee-darbee\ndarbeedarbee # Jechuufii:
[Word]
W [-\w]{3,}
{W}(?: [--\w""]+)* \1 <- filannoo("dup") -> {W} # jechoota danuu?
# hima keessaatti
{W}[;,:]?(?: [--\w""]+[;,:]?)* \1 <- filannoo("dup2") -> {W} # jechoota
danuu?
{Abc}{punct}{Abc} -> {Abc}{punct} {Abc} # Bakka duwwaa hinqabu?
{abc}[.]{ABC} -> {abc}. {ABC}         # Bakka duwwaa hin qabu?
[word]
# Teempireechara
([--]?\d+(?:[,.]\d+)*) (°F|Faranayitii) <- filannoo("metric") -> =
madaala(\1, "F", "C", u" °C", ".", ",") # Gara Seelshiyeesii jijjiiri:
([--]?\d+(?:[,.]\d+)*) (°C|Seelshiyeesii) <- filannoo("nonmetric") -> =
madaala(\1, "C", "F", u" °F", ".", ",") # Gara Faranayitii jijjiiri:
# Lakkoofsa ilaaluu
nvow (8[0-9]*|1[18](000)*)(th)? # 8, 8ffaa, 11, 11ffaa, 18, 18ffaa,
11000, 11000ffaa...
```

Fig. 1. Sample Rules.

V. Implementation and Evaluation

1) Implementation: The developed framework has been implemented using a Libreoffice text/document writer. LibreOffice is one of the document writer software supporting the integration of the Afaan Oromo spelling and grammar checker. Fig. 2 shows a list of extensions already in the Libreoffice software before installing the Afaan Oromo spelling and grammar checker.

As shown in Fig. 2, installing the developed framework as an extension of writing aids, there is no indicator of the Afaan Oromo language. However, after installing the Afaan Oromo spelling and grammar checker, the writer started to indicate an icon showing the language extension for grammar and spell-checking is installed. As shown in Fig. 3, the layout for the language extension is highlighted with a red color.



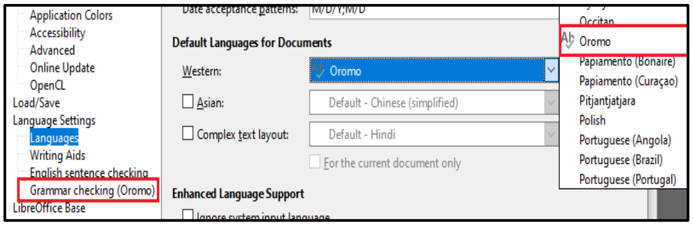
Options - Language Settings - Languages

Language Of
User interface: Default - English (USA)
Locale setting: Default - English (USA)
Decimal separator key: ☑ Same as locale setting (.)
Default currency: Default - USD
Date acceptance patterns: M/D/Y;M/D

Default Languages for Documents
Western: Oromo
☐ Asian: Default - Chinese (simplified)
☐ Complex text layout: Default - Hindi
☐ For the current document only

Enhanced Language Support
☐ Ignore system input language

Fig. 2. Layout before Installing the Afaan Oromo Spelling and Grammar Checker.



Date acceptance patterns: M/D/Y;M/D

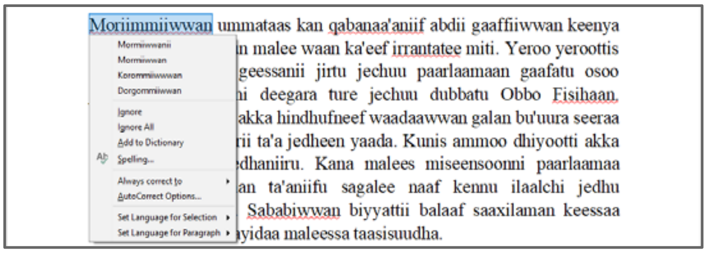
Default Languages for Documents
Western: Oromo
☐ Asian: Default - Chinese (simplified)
☐ Complex text layout: Default - Hindi
☐ For the current document only

Enhanced Language Support
☐ Ignore system input language

Fig. 3. Layout after Installing the Afaan Oromo Spelling and Grammar Checker.

The developed framework has been implemented to perform spelling checking, correction, grammar error detection and suggestion.

a) Spell-Checker Implementation: At this phase, the system is implemented to detect the wrongly spelled words and suggest possible corrections. Fig. 4 shows the implementation of the spell-checker.



Moriimmiiwwan ummataas kan qabanaa'aniif abdii gaaffiiwwan keenya
n malee waan ka'eef irrantatee miti. Yeroo yeroottis
geessanii jirtu jechuu paarlaamaan gaafatu osoo
hi deegara ture jechuu dubbatu Obbo Fisihaan.
akka hindhufneef waadaawwan galan bu'uura seeraa
rii ta'a jedheen yaada. Kunis ammoo dhiyootti akka
dhaniiru. Kana malees miseensoonni paarlaamaa
an ta'aniifu sagalee naaf kennu ilaalchi jedhu
Sababiwwan biyyattii balaaf saaxilaman keessaa
ayidaa maleessa taasisuudha.

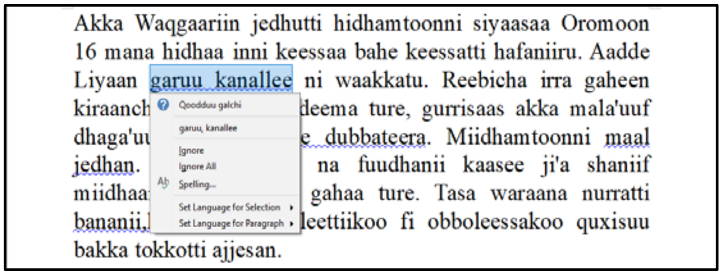
Fig. 4. Spell-Checker Implementation Sample Screenshot.

To indicate the wrongly spelled word in a sentence, all wrongly spelled words are underlined with a red flag. But if the word has no spelling error, no sign is displayed. As shown in Fig. 4, the wrongly written words are flagged with a red color underscore. When these wrongly written words are right-clicked, the possible suggestion is shown.

b) Grammar-Checker: This implementation was done to determine the wrongly written statements and extract the grammatically wrong statements within the text by the system. The prepared corpus (doc1 and doc2) is used alternatively for error detection and false alarm identification in this implementation. Fig. 5 is the screenshot of the implementation of detection and correction of the grammar of the proposed framework.

Fig. 5 shows how the system detects the grammar error and suggests the possible correction.

The statement that has a grammar error will be flagged with a blue underscore to make the user alerted and take proper action. The grammar suggestion will be shown when right-clicked on the flagged text. However, those statements written with a correct grammar rule will not be flagged.



Akka Waqgaariin jedhutti hidhamtoonni siyaasaa Oromoon
16 mana hidhaa inni keessaa bahe keessatti hafaniiru. Aadde
Liyaan garuu kanallee ni waakkatu. Reebicha irra gaheen
kiraanch deema ture, gurrisaas akka mala'uuf
dhaga'uu e dubbateera. Miidhamtoonni maal
jedhan. na fuudhanii kaasee ji'a shaniif
miidhaa gahaa ture. Tasa waraana nurratti
bananii,l leettiikoo fi obboleessakoo quxisuu
bakka tokkotti ajjesan.

Fig. 5. Detection and Correction of Grammar Errors.

2) Evaluation: The developed system is evaluated to check whether the objectives of the research are achieved or not. To evaluate the performance of the systems, two documents named doc1 and doc2 are prepared. These documents are used for testing and evaluation purposes, where each document contains about 500 sentences. All of the words and sentences within the first document (doc1) are grammatically correct, whereas the second document (doc2) contains the possible wrong spelling and grammar of doc1. The evaluation task is done in two phases, and the first phase evaluation is applied to test the detection and suggestion capability of the system for the spelling error within a sentence. The second phase is evaluated to check the system's performance for grammar error detection and correction suggestions made to the identified error.

a) Evaluation of Phase 1: The performance of the system for spelling checking has been evaluated at this phase. The document with 50 words has been prepared, and of these, 25 words were wrongly spelled while the rest 25 correctly spelled. The system is expected to suggest at least three correctly spelled words related to the detected spelling error.

b) Evaluation of Phase 2

- Grammar Error Detection: The framework is tested against doc1 to identify the grammatically wrong statements within the text, whereas doc2 is used to handle false alarms.
- Correction Suggestions: The correction suggestions evaluation has been done to evaluate the system's performance for suggesting the possible correct grammar of the detected errors. The system is tested by using the document containing the wrong grammar statements (doc1). Depending on the error detected, the number of correction suggestions may vary. However, the system is evaluated to suggest a minimum of two correct suggestions per error.

3) Confusion matrix: To demonstrate the performance of the proposed system, a confusion matrix has been used. The confusion matrix is a commonly used technique to describe the classifier's performance based on test data [9]. Table I shown below, describes the confusion matrix result for the system performance.

The performance measure shows the detailed description for the activities of the systems such as; checking the spelling error, suggesting the correction for the spelling error, grammar checking and correction are shown in the Fig. 6, 7, 8 and 9.

Evaluation matrix performance can also be explained using f-measure, precision and recall values. Following are the formulas used for these calculations.

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F - Measure = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$$

Table II is focused on the results of the evaluations.

TABLE I. PERFORMANCE MEASURE DESCRIPTION

S. No.	Evaluation Parameters	TP	FP	FN	TN
1	Have a Spelling Error	19	5	7	6
2	Spell-Checker Correction Suggestion	56	15	16	35
3	Have a Grammar Error	86	4	96	10
4	Predicted Correction Suggestion	122	28	78	50

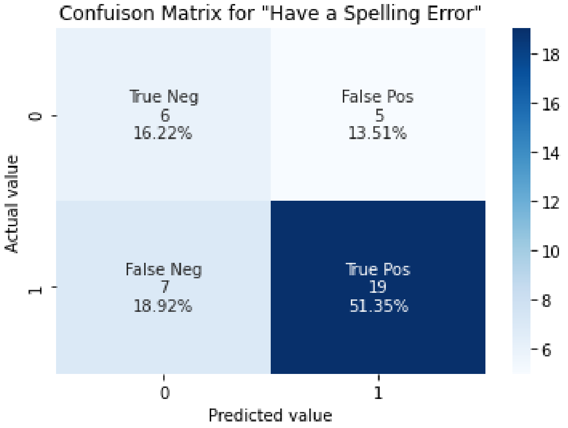


Fig. 6. Have a Spelling Error.

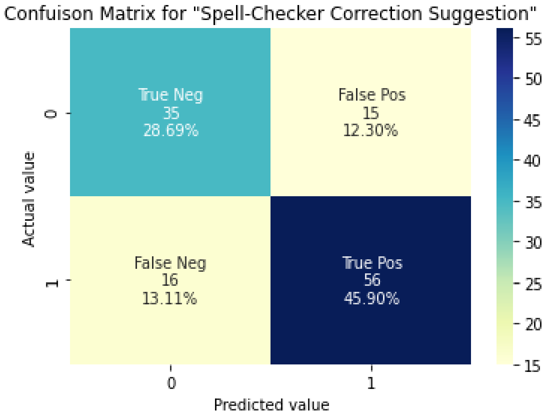


Fig. 7. Spell-Checker Correction Suggestion.

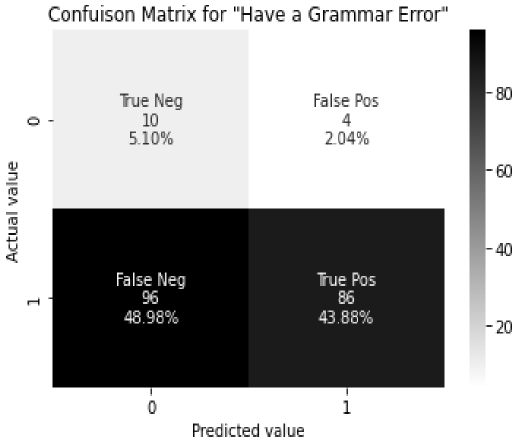


Fig. 8. Have a Grammar Error.

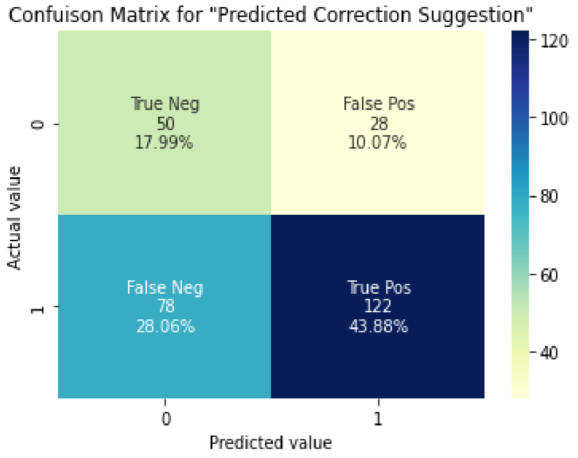


Fig. 9. Predicted Correction Suggestion.

TABLE II. PERFORMANCE EVALUATION MATRIX

S. No.	Parameters Used	Evaluation Result		
		Recall	Precision	F-Measure
1	Checking Spelling Errors	0.73	0.76	0.75
2	Suggesting Correction	0.78	0.62	0.70
3	Checking Grammar Errors	0.47	0.90	0.68
4	Suggesting Correct Grammar	0.61	0.71	0.66

VI. DISCUSSION

The developed system has been evaluated on four basic tasks: checking spelling errors within the text, suggesting the correct spelling for the detected error, checking grammar errors, and suggesting the correct grammar for the identified grammar errors. Accordingly, the confusion matrix result for checking the spelling error has scored 51.35% TP, 13.51% FP, 18.92% FN, and 16.22% TN, whereas spell-checker correction suggestion 45.90% TP, 12.30% FP, 13.11% FN and 28.69% TN score. But, for checking the grammar error and suggesting the possible correct grammar, the system has scored 43.88% TP, 2.04% FP, 48.98% FN and 5.10% TN result of confusion matrix measurement.

The performance evaluation result of the system, which is evaluated by the confusion matrix measurement, is described by using the f-measure, precision and recall values. So, for checking the spelling errors, the score is 73% recall, 76% precision and 75% f-measure. To suggest the correct spelling for the detected error, it has scored 78% recall, 62% precision and 70% f-measure. For checking grammar errors, the system has scored recall 47%, precision 90% and f-measure 68%, whereas, for the task of suggesting the correct grammar for identified error, it has 61% recall, 71% precision and 66% f-measure.

According to the experimental result presented, the framework has performed well for handling false alarm issues than error detection. The main reason that false alarm handling tasks scored higher than error detection tasks is that the written rules are primarily focused only on the common errors that mislead the meaning or make the statement meaningless. The evaluation performance for the task of grammar error detection within a statement scored lower performance. The problem with the task of error detection happened as a result of uncovered cases within the rule. Within a given statement, if the grammar problem is a semantic error, it will not be detected and/or corrected by checking only the sentence structures. To find and/or correct the semantic errors of the statement, it is must to analyze the semantic structure of the sentence. Therefore, it's possible to solve a challenge to detecting grammar errors due to the dynamic occurrences of the error.

VII. CONCLUSION

This study has developed an integrated model to develop a grammar checker for Afaan Oromo using a rule-based. Approach integrated a spell-checker along with a morphological analysis on Afaan Oromo grammar checker to improve its performance. Data has been collected from various sources such as Oromia Broadcasting Network, BBC Afaan Oromo and other sources used for designing and implementation purposes.

The experimentation was done in two phases: spell-checking & correction phase and grammar-error detection & correction phase. The system is evaluated on detecting wrong spelling and grammar from the text and the possible suggestion made to the error. As a result, the evaluation result for detecting grammar errors in a statement has a low score than other activities. One of the main reasons is due to the uncovered cases within the rule for detecting the error.

VIII. RECOMMENDATION

The designed prototype for improving the Afaan Oromo grammar checker has a promising result. Even if the model performance showed a good result, there is still work that needs to be done to improve the performance. The following are recommended as a future research direction.

- The grammar error can be syntactic, related to the structure of a sentence and/or semantic, which is related to the meaning of the sentences. But this work attempted to solve the grammars having only syntactic errors. Therefore, further research work has to be done on the Afaan Oromo Grammar checker that can check the semantic analysis of the sentences.
- Since it is difficult to catch all of the grammar errors using only a set of rules, it must implement another method. Therefore, machine learning methods combined with a rule-based approach may improve the performance of the Afaan Oromo grammar checker.

REFERENCES

[1] Anonymous., "the-major-importance-of-grammar-check-website," 01 February 2016. [Online]. Available: https://www.nounplus.net/blog/the-major-importance-of-grammar-check-website/. [Accessed 22 March 2020].

[2] Kualo, "Oromo language, alphabet and pronunciation," omniglot, 2010. [Online]. Available: http://www.omniglot.com/writing/oromo.htm. [Accessed 14 October 2018].

[3] S. D. Baviskar and S. S. Bahekar, "Comparative Study of Rule-Based Approach for Grammar Checker," *International Journal of Management, Technology And Engineering,* vol. IX, no. I, p. 1315, 2019.

[4] G. O. Ganfure and D. Dr. Midekso, "Design And Implementation Of Morphology Based Spell Checker," *International Journal of Scientific & Technology Research,* vol. 3, no. 12, 2014.

[5] D. Tesfaye, "A rule-based Afan Oromo Grammar Checker," *International Journal of Advanced Computer Science and Applications,* vol. 2, no. 8, 2011.

[6] Mideksa, "Statistical Afaan Oromo Grammar Checker," MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2015.

[7] Mainsheet, "What is Python?," javatpoint, 07 April 2011. [Online]. Available: http://www.javatpoint.com/what-is-python. [Accessed 18 October 2018].

[8] M. Gasser, "HORN MORPHO2.5 User’s Guide," Research group of human language technology and the democratization of information, 2012.

[9] K. Markham, "Simple guide to confusion matrix terminology," Data School, 25 March 2014. [Online]. Available: https://www.dataschool.io /simple-guide-to-confusion-matrix-terminology/. [Accessed 08 March 2021].