Gadaa Journal/Barruulee Gadaa

Vol. 3 No. 1 January 2020
https:journals.ju.edu.et

e-ISSN: 2616-3985
p-ISSN: 2616-3977

A Bilingual Journal of the Institute of Oromo Studies (IOS) Jimma University

---

Full Length Research Paper

A Context Sensitive Text Writing Correction and Error Detection for Afaan Oromoo Words

Tirate Kumera*[1], Million Meshesha[2], and Workineh Tesema[3]

[1]Mettu University, Department of Management Information System
[2]Addis Ababa University, Department of Information Science
[3]Jimma University Institute of Technology, Department of Information Technology

Submission Date: November 23/2019
Acceptance Date: January 10/2020

Abstract

This article presents Afaan Oromoo context sensitive spelling checking using unstructured free text corpus. In the present paper, we describe a new and original approach for word error correction and detection which is a context based. Spell checking is ultimately important in Afaan Oromoo hence, Afaan Oromoo language is a morphologically rich. This application was developed to reduce the problem of misspelling at the typing time in Afaan Oromoo text writing. The purpose of the paper was to fix spell checking errors by developing spell checking application hence, such application is important for Latin script like Afaan Oromoo where there are long vowels in the words. The method used in this work was statistical method (2-gram) and unsupervised approach hence there is no annotated corpus for Afaan Oromoo. To find the number of the contexts in a free corpus, n-gram (bi-gram) was used and to find the similarity measure between the words levenshtein distance was used. The finding shows that non-word error and real-word error found in Afaan Oromoo while typing Afaan Oromoo words. The result shows that the performance of the system was surprising; however, the corpus suffered by data sparseness problem as cannot capture large vocabulary of words including proper names, abbreviations, special acronyms, hyphen, apostrophe, domain-specific terms, technical jargons, and terminologies. To the best of our knowledge, this work is the first of its kind for Afaan Oromoo .It argued that the accuracy of the system was 93.9%, this shows that the model is good but it needs further investigations particularly on proper names, abbreviation words, acronyms, hyphenated words, apostrophe words, domain-specific terms, technical jargons, and special acronyms using different algorithms.

* Corresponding author.

Key Terms: Spelling Correction, Context Sensitive, Afaan Oromoo Typing, Word Error Detection, Afaan Oromoo Word

Axareeraa

Barulleen qo'annaa kun kan dhiheessu akkaataa jechootni Afaan Oromoo akka galumsa isaaniitti kompiitarri kuusaa jechootaa fayyadamuun sirreessuu irratti kan xiyyeeffateedha.Qo'annoon kunis jechoota Afaan Oromoo barreessuufi dogoggora Soroorsuu/sirreessuufis jalqabaa fi haaraadha. Kompiitaraan jechoota sirreessanii barreessuun kaayyoo isa guddaandha.Sababiin isaas, Afaan Oromoon baay'ina dham-jechootaan sooressa (morphological rich) waanta'eefidha. Kaayyoon sombajjii kanaas rakkoo seeraan qubeessuu dadhabuu jechoota Afaan Oromoo hir'isuudha. Kaayyoon waraqaa kanaas sombajjii uume kana fayyadamuun rakkoo jechoota katabuu dadhabuu furuudha. Kunis, Afaan Oromoo sagaleen dubbachiiftotaa jechoota keessatti waan baay'atuufiidha. Maloonnifi al-goorizimiin qo'annoo kana keessatti tajaajilanmala staatistiksii (2-gram)fi kuusaa jechoota Oromoo guuraman, kana dura kan hinleenjineefi tajaajila biraaf kan hinoolleedha. Kuusaa jechoota kana keessa akkaataa galumsaa barbaaduuf 'n-gram' kan ta'e 'bi-gram'-fi walitti dhiyeenya galumsa isaanii beekuu fi immoo 'levenshtein' fayyadamneerra. Bu'aan qorannoo kanaa dogoggora jechaa kanta'ee fi dogoggora jechaa kan hintaane yeroo kataban kan mul'isuu dha. Rakkoo kana furuuf sombajjiin kun gahee guddaan qaba. Sombajjiin kun yeroo madaalamu bu'aa gammachiisaa mul'iseera. Sababni isaas, Afaan Oromoon kuusaa jechootaa qulqullina isaa eeggate waan hinqabneefii dha. Yeroo shaakalamu %93.9 gahumsa qulqullinaa qabaachuunisaa biraga'ameera. Haata'umalee, hanqinni sombajjii kanaa maqaa waantootaa, gabaajee, kottoonfachiisaa, jechat-ishoo, hudhaa fi kan kana fakkaatan irratti gahumsa gadaanaa waanqabuuf, qo'annoon itti aanu al-goorizimii jijjiiruun qo'annoo biraa gaggeessuun akka danda'amu argarsiisa.

Jechoota Ijoo: Dogoggora Jechaa, Galumsa Jechaa, Jechoota Afaan Oromoo, Jecha Barressuu Jecha Soroorsuu

---

Introduction

In computing, spell checker is the process of detecting and providing spelling suggestions for wrongly spelled words in a text when users type words. Fundamentally, a spell checker is a computer application that uses a corpus of words to perform spell checking. The bigger corpus is the higher mis-spelled error detection data set (Bassil and Alwani, 2012). The fact that spell checking is based on a free corpus, suffer from data sparseness problem because it cannot capture large vocabulary of words including proper names, hyphen, apostrophe (*hudhaa)*, domain-specific terms, technical jargons, special acronyms, and terminologies. As a result, it exhibits low error detection rate and often fails to catch major errors in the text.

A context based word error is an error that turns an intended word into another word of the language that can give different meanings. A correction of such error is easy if the erroneous word is not part of the language. However, the detection is harder if the erroneous word is part

of the language. For example, the mis-spelling of the intended word "*dhuuftan*" as "*dhuftan*". Such an error can be detected by examining the contexts in which the candidate word is used (Nemera, 2001).

Afaan Oromoo has no spell checkers to write Afaan Oromo words. Therefore, Oromoo spell correction and error detection is needed to solve mis-spelling problem (Desta and Mehta, 2018) and (Olani and Midekso, 2014). As stated in Oromoo G. Q. A. (1995) like a number of other African languages, Afaan Oromoo has a very rich morphology. Most of these Afaan Oromoo words have long vowels which increases the probability occurrence of misspelling word to a higher frequency.

To the best of our knowledge, this work is the first of its kind for Afaan Oromoo. Several researches were conducted on the languages such as English, Arabic, and Chinese. Fewer researches were conducted on Amharic language, but so far little has been done on Afaan Oromoo. Many researchers think that once spell checker developed for the other languages, it works for Afaan Oromoo too. However, this is not true for Afaan Oromoo due to its structures, punctuations (where the hyphen and apostrophe have implications) and the formation of syntax of the Afaan Oromoo was different from that of others (Oromoo G. Q. A., 1995). This study revealed that the best spelling error correction and detection in Afaan Oromoo.

As computers, smart phones and other technology gadgets are used in a daily practice for everyone, uses them to type and express their idea on social media and different applications to communicate with persons. Context sensitive spelling correction brought a huge advantage and accumulated rapidly because of the development of online media and social networks. Additionally, it helps to share knowledge between authors and readers. As a result, writing system must clearly transmit the knowledge without confusing the readers due to misspelling of the authors in his/her messages. In a written language communications poor spelling can lead to information gap, loss of information and misunderstanding of information which hampers the advancement of the language use in a modern technology for designed gadgets.

Related Work

The study conducted by Desta and Mehta (2018) on Afaan Oromoo rule based spelling checking showed that rule based spell checker used to spell correctly and its space and time complexity is less in comparison with other methods. In the same way, a dictionary look up method is used for non-word errors detection. However, this rule based spelling checker was not work for all Afaan Oromoo words as they were developed simple rule for spell checkup. Additionally, the performance of this rule based spell correction was not measured and simply justified the rule based spell checking method is better than the other methods. Finally, their finding didn't show how real word error occur and detect it unlike our study. But, this context sensitive word error correction and detection is appropriate to fix word error in Afaan Oromoo written in long vowels and easy to capture the contexts in which the misspelt words occurred.

Another study conducted on Afaan Oromoo morphology based spell checker by Olani and Midekso (2014) showed that unlike other languages, the spelling correction tool is not available

for Afaan Oromoo, the Cushitic language family spoken in Ethiopia. The intention of the work was to design and implement non-word Afaan Oromoo spell checker. A dictionary based morphological analysis which is morphology based spell checker) was used to design the system. Besides, to develop morphology based spell checker, the knowledge of the language morphology is used. However, contrary to our work, they argued that their methodology works for other languages showing similar morphology with Afaan Oromoo than contexts. Their study focuses only on non-word error detection and correction.

The aim of this study was to investigate misspelling detection and correction in Afaan Oromoo language in particular. For Afaan Oromoo there are no misspelling detection and correction. Unlike other studies, this study focused on the contextual use of that co-occurred word with misspelt words. Therefore, the novel idea behind this study was to detect misspelt words in Afaan Oromoo using contextual sensitivity. The other innovated idea of this study was that it focused on both real words error and non-word errors. Afaan Oromoo is under resourced and lacked standard corpus. For this study purpose we prepared corpus. Therefore, this study was basically to guide the contextual based real word error and non-word error detections and correction for further investigation.

Context Based Word Correction and Error Detection

In contrast to spell checkers, a syntax checker may possibly detect a context based word error. Error detection by a syntax checker is difficult if the word contains many special terms, symbols (hyphen), abbreviation, or acronyms whose syntactic contribution cannot be established without a complete understanding of the contexts. For a given occurrence of a word *w*, the structure of the neighborhood of the occurrence and the connection with other occurrences of *w* and their neighborhoods were examined statistically (Tesema, *et al*., 2016). To examine the word under investigation, the second word acts as a reference.

Unlike other languages, word error detection and correction in Afaan Oromoo was different as revealed in the finding of the study.The non-word spelling mistake is an error in which the sentence consists of a word that doesn't exist in the language. For instance, *Siifan ganamaan mana* ***cirree*** *seentee* ***cirree*** *nyaachuuf* here is a two error of words occurred in this sentence. These two words are Afaan Oromoo words when we look at the terms. However, contextually this sentence has two errors which are semantically different from the context in which it occurred. The bold term "***cirree"*** is not correct according to the contexts on the left and right hand side. In Afaan Oromoo, these errors are identified only by determining the contexts around misspelt word. This type of error can occur when the user is double typing the spelling 'r') from keyboard on the computer.

Moreover, high typo errors occur when writing Afaan Oromoo words. For example, "***Baga nagaan dhuuftan!"*** to say "welcome". But here due to the spelling error in the last word; its meaning and semantics will be changed into unintended one which meant in advertently "***It is good that you fart in peace"***. For more examples of typographic errors that might occur in Afaan Oromoo due to misspelling can be seen from table(table1) below.

Table 1.Sample of Typographic Errors

Afaan Oromoo Typographic Errors	
Dhuftan	*Dhuuftan*
Ciree	*Cirree*
Dirree	*Diree*
Ragaa	*Raagaa*
Daboo	*Daabboo*

In order to propose candidate word corrections Kernighan,*et al.,* (1990) make the simplifying assumption that the correct word will differ from the misspelt by a single insertion, deletion, substitution, or transposition. As Damerau's (1964) results showed, even though this assumption causes the algorithm to miss some corrections, it should handle most spelling errors in human typed text. The list of candidate words is generated from the typo by applying any single transformation which results in a word in a large online dictionary.

A confusion matrix can be computed by hand-coding a collection of spelling errors with the correct spelling and then counting the number of times different errors occurred (Grudin, 1983). Kernighan,*et al.,* (1990) used four confusion matrices, one for each type of single-error.

- del[x, y] contains the number of times in the training set that the characters xy in the correct word were typed as x.
- ins[x, y] contains the number of times in the training set that the character x in the correct word was typed as xy.
- sub[x, y] the number of times that x was typed as y.
- trans[x, y] the number of times that xy was typed as yx.

Typographic Errors

These errors are occurring when the correct spelling of the word is known but the word is wrongly typed mistakenly. These errors are mostly related to the keyboard, therefore; do not follow any linguistic criteria.

A. Single-Error

A study conducted by Damerau (1964) showed that 80% of the typographic errors were fallen into one of the followings four categories.

i. Single letter insertion; e.g. typing *dhuuftan* for *dhuftan*.

ii. Single letter deletion, e.g. typing *ciree* for cirree.
iii. Single letter substitution, e.g. typing *ciree* for *cilee*.
iv. Transposition of two adjacent letters, e.g. typing *mana* for *nama.*

B. Cognitive Error

Cognitive errors are produced when the correct spellings of the words are not known and/or lack of knowledge about correct spelling of the target language. In these types of error, the pronunciation of the misspelt word is intended as correct word. The user simply does not know how to spell and /or have misconception on how to spell the word; so, writes the word in an erroneous form. For instance, the users write "computer flies" instead of "computer files". In the spelling error, cognitive error happened due to pronunciation similarities between the erroneous word and the correct word (Christen, 2012). For instance, the user writes "ingenious" instead of "ingenuous" in the sound they formed are similar.

Spelling checkers are used in various applications, such as machine translation, search, information retrieval, etc. There are two main issues related to spell checker, these are error detection and error correction. Many techniques are available for detection and correction to develop upon the real word error and non-word error checking processes.

I. Non-Word Error

Non-word error is spelling error that is not found in the list of words in the dictionary (Tavast, *et al.,* 2012). In non-word error, a word may be wrongly typed because there is extra space, extra character, misspelt word, or other possibilities. These errors are easier to detect, because just comparing the words in a text with the entries in a dictionary will filter out the erroneous words. According to Lorraine (2008) 80% of misspelt words that are non-word error are the result of a single insertion, deletion, substitution or transposition of letters:

- ***Insertion:*** *Adding an extra letter, e.g., 'laekki' instead of 'lakki' which means no. An important special case is a repeated letter, e.g. 'deemmi' instead of 'deemi'to mean" go" in English context.*
- ***Deletion:*** *Missing a letter, e.g., 'tle' instead of 'tole'. An important special case is missing a repeated letter, e.g., 'eyeen' instead of 'eyyeen''.*
- ***Substitution:*** *Substituting one letter for another, e.g., 'ejjenni' instead of 'ejjenno'. The most common substitutions are incorrect vowels.*
- ***Transposition:*** *Swapping consecutive letters, e.g. 'jiaarchuu' instead of 'jiraachuu'.*

II. Real Word Error

Correcting real-word error is the most challenging task for a spell checking system. However, different scholars tried to tackle the problem of real word error. Rill and Moore (2000) suggested the use of a noisy channel to predict the actual correction for a real-word error. The dataset prepared for this study was around 100 million words of corpus and use n-gram statistics to correct real-word error. The idea centers on generating candidate spellings for every misspelt word by only applying simple edit operations such as insertion, deletion and substitution, and

then using *n*-gram statistics derived from a corpus by computing the probability of word (Sundby, 2009).

Method and Materials

This study dealt with a new context-sensitive spelling correction method for detecting and correcting non-word and real-word errors in Afaan Oromoo text documents. The statistics of the corpus consists of small volume of n-gram word sequences, extracted from different sources (tourism bureau, newspaper, cultural documents and websites). The reason why bigram (2-gram) is selected was small size of the collected corpus and has sparse problem. Basically, the proposed method comprises an error detector that detects misspellings, a candidate spellings generator based on a character 2-gram model that generates correction suggestions, and an error corrector that performs contextual error correction.

A number of methods have been developed for the detection of context based spelling errors. The Bayesian method (Golding, 1995) handles context-based spelling correction as a problem of ambiguity resolution. The ambiguity is modeled by confusion sets. The Bayesian method uses decision lists to choose the proper word from the confusion set. It also relies on classifiers for two types of features, context words and collocations. The method learns these features from a training corpus of correct word. The Bayesian method can be viewed as a way of applying such an algorithm to the spelling error correction problem; we pick the candidate word which is closest to the error in the sense of having the highest probability given the error.

The other algorithm used in this study was levenshtein distance to measure the similarity between contexts. It is the minimum number of edition operations necessary to transform one word into another. In this case instead of talking about the minimum edit distance between two contexts, we are talking about the maximum probability alignment of one context with another. It is a string matrix for measuring the difference between two sequences. It is a minimum number of single character edits required to change one word into the other (insertion, deletion, substitution and transposition).

Corpus Collection Procedures

In spelling correction and error detection the size of the corpus is (92 MB) were prepared as a training dataset. In this work, Afaan Oromoo free corpus is created manually to apply spell checking system as there is no standard corpus developed for Afaan Oromoo., Therefore, we collected and developed dataset from Oromia Broadcasting Network (OBN), Oromia Culture and Tourism office, and from different websites (using web crawler), and Voice of America (VOA) Afaan Oromoo section. This corpus contains different contents of disciplines, such as cultural, social, political, sports, and economics in order to avoid data scarcity and to prepare the rich dictionary as well as test the model (Han and Baldwin, 2011).

Training and Testing

The corpus was preprocessed (tokenized and normalized) and cleared from any kind of unnecessary errors which is invalid to represent Afaan Oromoo vocabulary words. The sample

of test dataset were 5, 818 words for testing the proto type. These words were raw terms and didn't not use for any researches purpose. The misspelt words that exist in the tested dataset were 100% checked and manually inputted to the system. Prepared words were stored in the form of the dictionary and accepted as corrected words and used for constructing bigram model to detect and correct errors. The dictionary prepared from the collected corpus was preprocessed by applying tokenization and normalization. Stored words were saved in the forms of text and used for cross check when the user inserted words. The architecture of the system as depicted in figure 1.

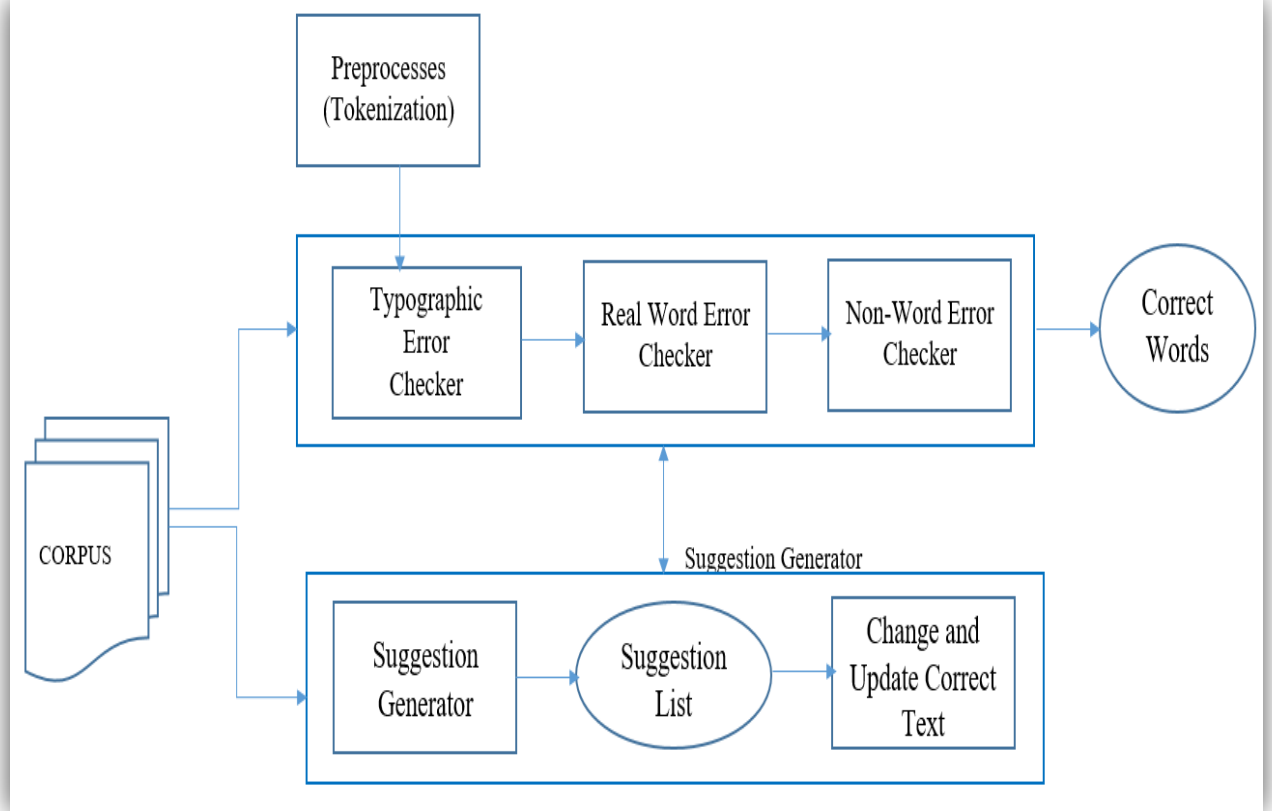

Figure 1. Architecture of the System

Implementation and Experiments

The experiment conducted on a set of text documents from different domains and containing misspellings, showed an outstanding spelling error correction rate and a drastic reduction of both non-word and real-word errors. Every word from the text is looked up in the speller lexicon. When a word is not in the corpus, it is detected as an error. In order to correct the error, a spell checker searches the corpus for words that resemble the erroneous word most. These words are then suggested to the user who chooses the word that was intended.

As mentioned in Figure 2, the system has text area which takes input from the users and input text is directly typed to the text area. Since, the model is interactive it waits for button "spelling check" click to detect the error. The error detection module is responsible to preprocess and

compares the inserted words with the dictionary and bigram model. For those words which could not be found in the list, the model accepts as the misspelt words, but for others the model leave as it is. As it mentioned above, error detection was executed by using dictionary look up and the bigram analysis.

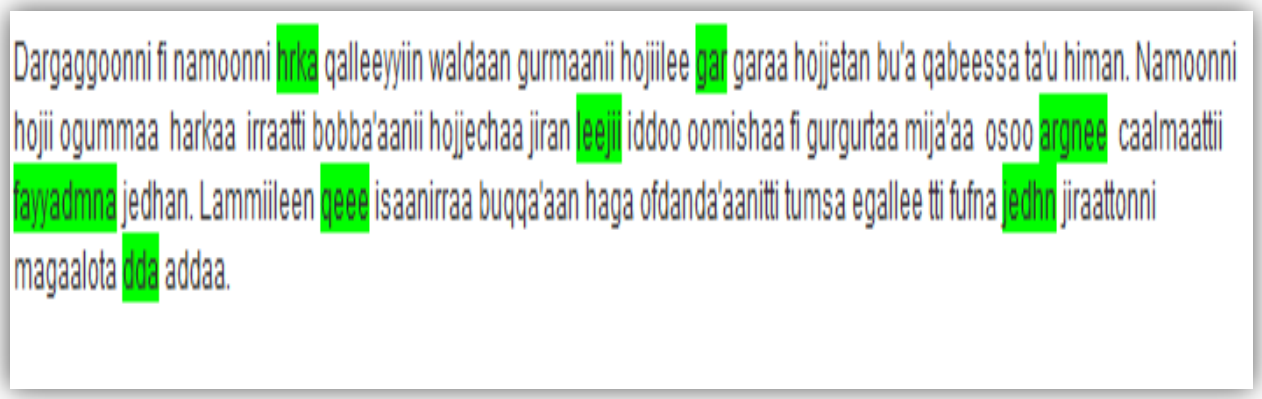

Figure2. Non-Word Error Detections

The model detects error at word level to identify non-word spelling errors that not found on the list of dictionary by dictionary look up method. In the user words, the word that does not exist in the dictionary is detected as misspelt words and highlighted by the color, otherwise the model accepts as corrected words.

On the other hand, the real word errors are detected under the consideration of bigram words sequences that comes together and sequence of bigram words does not exist in the bigram list, then it's detected as real word error. The input sentences were broken down into bigrams.The words were generated along with its probability information which is used to rank the candidate suggestion to correct the errors. If the bigram words found in the bigram model, there is no error which is considered as valid word. But if one of the word does not exist in the bigram word list it is considered as misspelt word and the model flagged by highlight with the magenta color and display misspelt words (see figure 3).

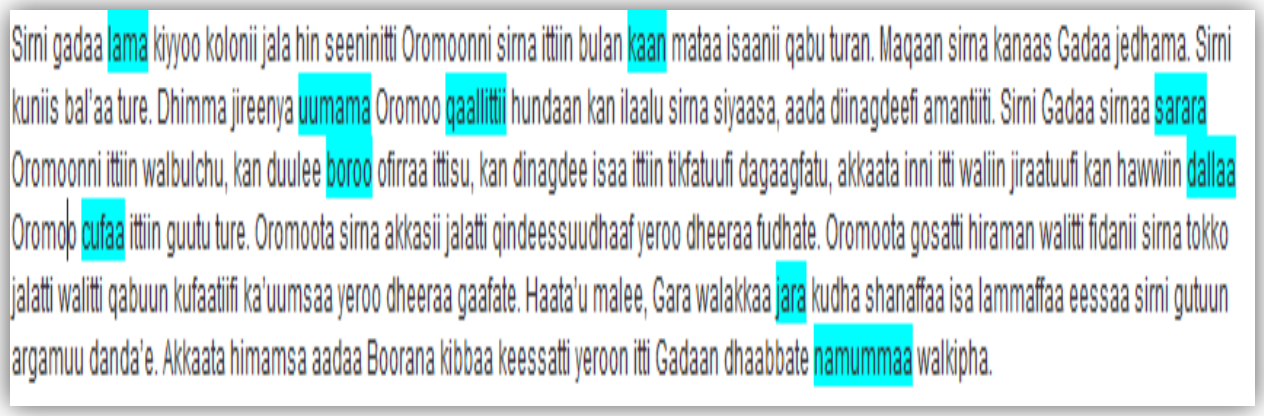

Figure3. Real-Word Error Detections

Error Corrections

To correct the misspelt words, the algorithm is providing a set of possible candidate corrections (see figure 4). After a word is flagged as wrongly spelt, possible set of suggestion is provided for the user. For non-word error levenshtein edit distance responsible to generate the suggestion for misspelt words and take minimum values and rank them accordingly.

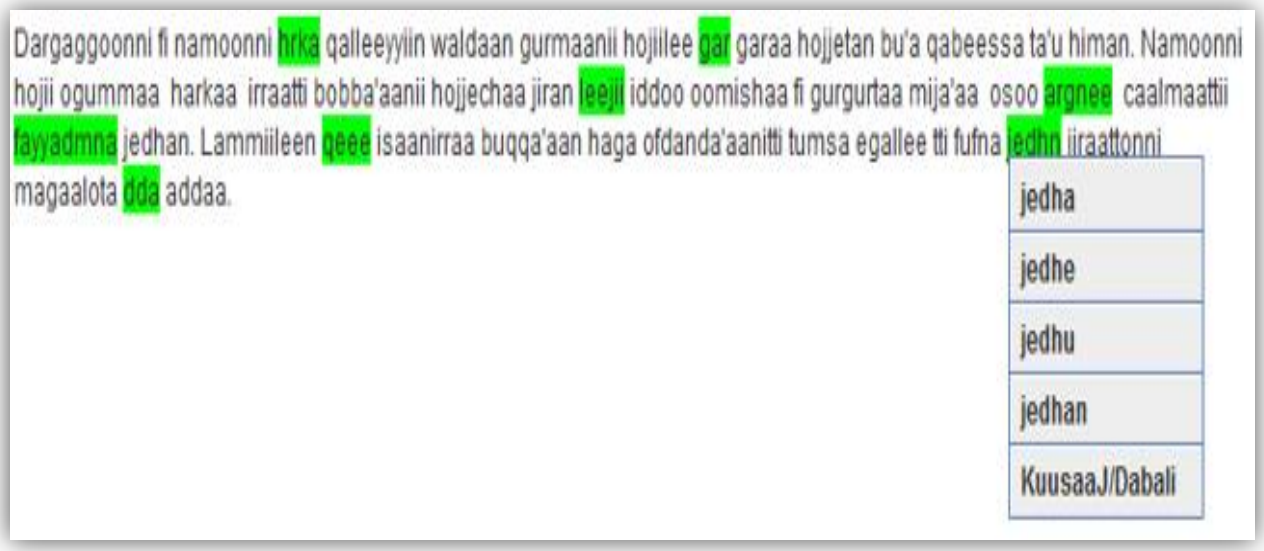

Figure4. Non-Word Error Correction

Correction module uses dictionary list to provide a spelling suggestion for each word error flagged as misspelt in the given inputs of words. The errors were corrected and modified through the suggested words that displayed in the pop up menu. The model provided a list of candidates depending on the value returned by computing the distance between them. Additionally, if the users are aware of the misspelt words, the model gives him or her a chance to include the misspelt words as corrected words into the dictionary.

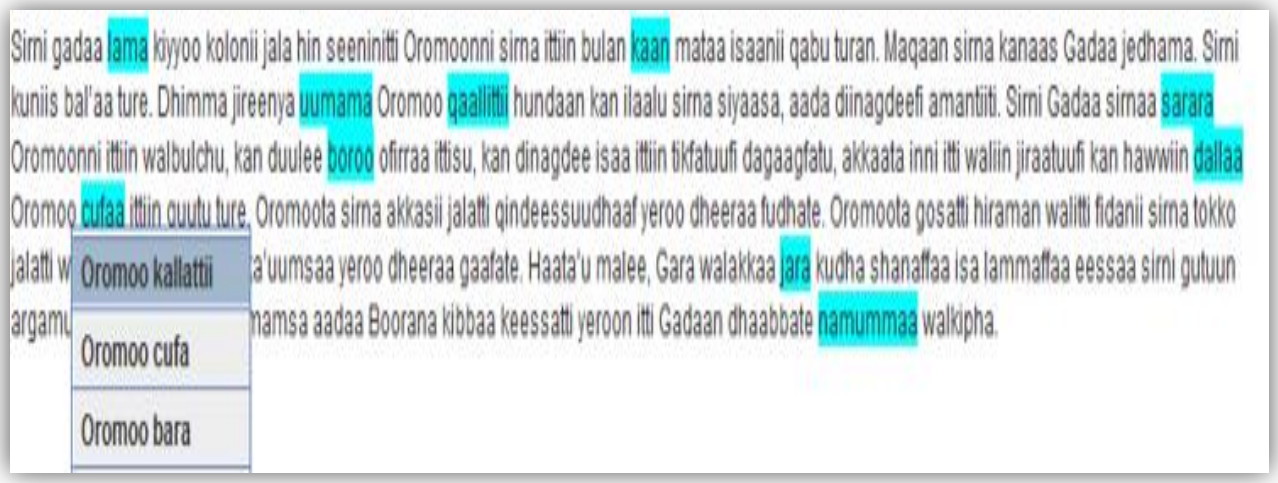

Figure5. Real-Word Error Correction

Add to Dictionary

Additionally, the model provides 'Add-to-Dictionary' method for the user if the user correctly knows the misspelled flagged by the model is correct word, flagged words accepted as corrected word by clicking on the 'Add-to-Dictionary', and the words automatically added to the dictionary.

On the other hand, real word errors were corrected using bigram probabilistic information. After misspelt words were flagged, the correction module provides candidate lists for real-word error from the corpus by computing the probability of the occurrence of words one after another. Then replace the invalid words by clicking each error word at any position and search the alternatives from the pop up menu.

Result and Discussion

The finding of this study revealed that the captured contexts had a great role to detect the misspelt words in Afaan Oromoo. The captured contexts were preceding and immediately following misspelt words. As the obtained result indicated the preceding contexts were highly important to detect Afaan Oromoo misspelt words particularly (Tesema, *et al*., 2016).

Contrary to Desta and Mehta (2018), this study confirmed that typing Afaan Oromoo words produced both real word and non-word errors. As it shown in Figure 5, real word errors were corrected and modified through the suggested words that displayed in the pop up menu. However, the result obtained by Desta and Mehta (2018) pointed out that real word errors was not found in Afaan Oromoo which is contrary to our work. In their study, they used rule based misspelt words correction focusing only on non-word error identification which cannot detect real word error. Thus, based on this result we conclude that rule based spell checkers in Afaan Oromoo has its own bottleneck to correct misspelt words in comparison with a context based misspelt words.

The finding of this study revealed that spelling errors exist in Afaan Oromoo. In the analysis the study found out that 2,921 words were misspelt. Of this 153 words were non-word errors while 89 words were found out to be real word errors. Thus, a further comprehensive study is required to come to a clear conclusion. However, it was indicative enough to realize that non-word error detection is the first step towards a truly professional spell checker development. The system achieved 93.9% performance accuracy for correction of word errors. The result indicated that the model effectively and efficiently detect and correct both non-word errors and real word errors. The coverage of analysis of the corrected words model was determined by recalling 93.7 % value. However, the model needs additional data to increase the value scored by the recall. Though all misspelt words correctly flagged were scored 100 values by the model according to the experiment conducted and presented (Guya, 2003).

As the study conducted in a syntax checker indicated Afaan Oromoo word error detection by a syntax checker is likely face difficult when the sentence contains many special terms such as symbols, abbreviations, hyphens, apostrophes, acronyms or conventions that their syntactic contribution cannot be established without a complete understanding of the contexts. In the same token, Afaan Oromoo use different symbols like hyphen to connect two words which makes it likely difficult to detect and correct errors in Afaan Oromoo. The finding of the study revealed that most words in Afaan Oromoo were typo-erroneous namely non-word and real word error (Gamta, 2005).

On the other hand, the performance of spell checker for real word error is 92.9% due to the size of the corpus in training text. This result showed that the corpus size used for training the model did not cover all the words in the testing data. Therefore, the sequence of words in the sentences depended on the bigram generated and needs more corpora to increase the efficiency and effectiveness of checking the spelling errors. The model coverage of corrected Afaan Oromoo words was determined by recall score of 92.7% from the sample set for training to check the model. This result showed that the sample test used for testing the prototype system needs improvement because the real word errors were marked based on proceeding words in the sentences. The misspelt words that exist in the tested data were 100% checked and the system was proved correct in detecting all errors not included in the bigram lists (Gragg and Gene, 2006).

The result of the current study support the using of N-gram probabilistic approach is affording optimistic result to correct the real word misspellings. The study conducted by (Ghotit, *et al*., 2011) also supported this result which showed around 93% of real word errors were corrected by using N-gram probabilistic method. Moreover, N-gram probabilistic approach is promising result as compared with restoring lexical Cohesion method, which focuses on semantic relatedness to correct the real word error as shown by (Bassil and Semaan, 2012). The recall scores of 50% have a precision of 20% to correct misspelt words. On the other hand, to correct non-word error Levenshtein provides expectant results as shown in Table 2. The study conducted by (Yitayal,*et al*., 2016) revealed that correcting non-word error by using Levenshtein edit distance score of 95.62% accuracy in Amharic language.

According to the evaluation of the spellchecker presented in Table 2, context based spell checker for Afaan Oromoo is optimistic system in order to solve the problem of the misspelt words, the system scores 93.9% accuracy to correct both non-word and real word errors.

Table 2: Evaluation of the System

Error	TP	TN	FN	R	P	F	Accuracy
Non-word	2921	153	157	0.948	1	0.97	0.95
Real-word	2897	89	245	0.92	1	0.96	0.929
Average	-	-	-	0.937	1	0.965	0.939

The Contribution of the Study

This study contributed significantly to the word errors detection and correction in Afaan Oromoo unlike English, Amharic and Tigrigna as Afaan Oromo has long vowels, these vowels length sometimes need to be typed twice using the same spelling repeatedly and this leads to wrong spellings. Therefore, the idea behind this study was to detect misspelt words in Afaan Oromoo using contextual sensitivity. Even if there is no Afaan Oromoo misspelt checker system or software, this study brought a context based misspelt words to add a quality to the study, unlike other researchers. In a previous Afaan Oromoo word error detection and correction study no one applied in neighbor words in a context. However, two studies tried Afaan Oromoo spelling correction based on rule and morphology different from our study. In fact, for the sake of this study, we used existing algorithms Bayesian and levenshtein edit distance like other researchers. For under resourced languages like Afaan Oromoo, the matter is not with an existing algorithm rather than the corpus used. Even using the existing algorithms, the result and performance of the system was sometimes very poor unlike our system. Afaan Oromoo is under resource and has no standard compiled corpus for this purpose, thus these researchers prepared the corpus for their own for this research. The other innovated idea of this study was focusing on word and non-word errors. Therefore, this study was basically helpful to guide and give some insights to future researchers to the context based word error researches to apply in solving the problems in their future studies.

Conclusion

This study revealed that Afaan Oromoo context based spell checking performs well. Thus we can conclude that Afaan Oromoo context sensitive spell checker is indispensable in the text correction process in Afaan Oromoo as it is rich morphologically. Specifically, the preceding contextual use of words was vitally important to detect Afaan Oromoo misspelt words. Afaan Oromoo word level errors detection is difficult at syntactic level when the sentences contain many special terms such as abbreviations, acronyms, hyphenated, and apostrophe words. In addition, the study found out that Afaan Oromoo text typing is produced both word and non-word level errors. N-gram probabilistic approach is affording optimistic result to correct the real

word misspellings. Accordingly, its performance of the system confirmed that the system scored 93.9% accuracy in detecting errors. Consequently, a developed system was optimal to detect and correct misspelt words in Afaan Oromoo. However, further research is needed using parallel algorithm and standard corpus so as to enhance the accuracy of word and non-word error detection and correction performances of the context based spell checker processes.

. Future Work

This study has its own limitations on abbreviation, acronyms, hyphenated words, and words with glottal stop (*hudhaa*). Even though the performance model of our system was effective and efficient, the developed corpus has a lot of noise, sparse and scarcity of corpus spell checkers processing problem as Afaan Oromoo is under-resourced in its corpus compilation. Therefore, with such kind of noisy data, the obtained result was enough as the study was conducted on under-resourced language with non-standard corpus. Further research might be conducted on abbreviations words, acronyms, hyphenated words, and words with glottal stop in Afaan Oromoo.

Acknowledgement

We would like to thank Jimma University, College of Natural Sciences for financial contribution to conduct this research. We also thank all participants in this study for their valuable data provision and cooperation.

References

Abera, N. (2001). Long vowels in Afan Oromoo: A generic approach, School of graduate studies, Addis Ababa University.

Bassil, Yand Semaan, P. (2012). ASR Context-Sensitive Error Correction Based on Microsoft N-Gram Dataset, *4*(1), PP: 34–42.

Brill, E., and Moore, R. C. (2000). An improved error model for noisy channel spelling Correction. In Proceedings of the 38th Annual Meeting on Association for Computational

Linguistics. (Nd). Association for Computational Linguistics.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. Communications of the Association for Computing Machinery, 7(3), PP: 171–176.

Debela T.(2011). A rule-based Afaan Oromooo Grammar Checker, International *Journal of Advanced Computer Science and Applications*, Vol. 2, No. 8.

Gaddisa Olani Ganfure and Dida Midekso. (2014). Design and Implementation of Morphology Based Spell Checker, International Journal of Scientific & Technology Research Vol-3, No-12.

Gamta, T. (2005).*Seera Afaan Oromooo*. Finfinnee, Boolee Press.

Gersten, R. M., Fuchs, L., Coyne, M. D., and Greenwood, C. R. (2005). Quality Indicators for Group Experimental and Quasi-Experimental Research in Special Education.

Golding, A. R. (1995). Bayesian hybrid method for context-sensitive spelling correction, Proceedings of the Third Workshop on Very Large Corpora, Cambridge, MA, PP: 39-53.

Gragg and Gene B.(2006) Oromoo of Wollega: Non-semetic languages of Ethiopia , East lansing, Michigan state University press.

Grudin, J. T. (1983). Error patterns in novice and skilled transcription typing. In Cooper, W. E. (Ed.), Cognitive Aspects of Skilled Typewriting, pp.121–139. Springer-Verlag, New York.

Oromoo G. Q. A.(1995).*Caasluga Afaan Oromoo*, Jildi I, Komishinii Aadaaf Turizmii Oromiyaa, Finfinnee, Ethiopia, PP: 105-220.

Guya T. (2003) CaasLuga Afaan Oromooo: Jildii-1, Gumii Qormaata Afaan Oromoootiin Komishinii "Aadaa fi Turizimii Oromiyaa", Finfinnee.

Han, B., & Baldwin, T. (2011). Lexical Normalisation of Short Text Messages : Makn Sens a # twitter, PP: 368–378.

Jurafsky, D., and Martin, J. H. (2009). Speech and language processing: An introduction to Natural language processing, computational linguistics and speech recognition. Prentice Hall series in artificial intelligence, PP: 1-1024.

Kernighan, M. D., Church, K. W., and Gale, W. A. (1990). A spelling correction program base on a noisy channel model. In COLING-90, Helsinki, Vol. II, PP: 205–211.

Lorraine H.(2008).Spell Checkers and Correctors: A Unified Treatment. University of Pretoria.

MegersaDestaJeldu, and Rutvik Mehta. (2018). Rule Based Afan Oromoo Analyzer for Spell Checker, International Journal of Advances in Electronics and Computer Science, Vol-5, No-7.

Nigusu, Y, Getachew M, and Teferi K (2016). Context based spell checker for Amharic.

Unpublished master's thesis, JimmaUniversity.

Youssef Bassil and Mohammad Alwani. (2012). Context-sensitive Spelling Correction Using Google Web 1T 5-Gram Information, Computer and Information Science Vol. 5, No. 3.

WorkinehTesema, DebelaTesfaye and TeferiKibebew. (2016). Towards the Sense Disambiguation of Afan Oromo Words Using Hybrid Approach (Unsupervised Machine Learning and Rule Based), *Ethiopian Journal of Education & Science*, Vol-12 No-1.

Workineh Tesema and Duresa T., (2017).Investigating Afan Oromo Language Structure and Developing Effective File Editing Tool as Plug-in into Ms Word to Support Text Entry and Input Methods. American Journal of Computer Science and Engineering Survey