Journal of
Information Technology & Software Engineering

Tesema and Tamirat, J Inform Tech Softw Eng 2017, 7:2
DOI: 10.4172/2165-7866.1000200

Research Article Open Access

Enhancing the Text Production and Assisting Disable Users in Developing Word Prediction and Completion in Afan Oromo

Workineh Tesema[1]* and Duresa Tamirat[2]

[1]Department of Information Science, Jimma University, Jimma, Ethiopia
[2]Department of Information Science, Medawolabu University, Robe, Ethiopia

Abstract

This work presents a word prediction and completion for disable users. The idea behind this work is to open a chance to interact with computer software and file editing for disable users in their mother tongue languages. Like normal persons, disable users are also needs to access technology in their life. In order to develop the model we have used unsupervised machine learning. The algorithm that used in this work was N-grams algorithms (Unigram, Bigram and Trigram) for auto completing a word by predicting a correct word in a sentence which saves time, reduces misspelling, keystrokes of typing and assisting disables. This work describes how we improve word entry information, through word prediction, as an assistive technology for people with motion impairment using the regular keyboard, to eliminate the overhead needed for the learning process. We also present evaluation metrics to compare different models being used in our work. The result argued that prediction yields an accuracy of 90% in unsupervised machine learning approach. This work particularly helps disable users who have poor spelling knowledge or printing press, institutions or government organizations, repetitive stress injuries to their (wrist, hand and arm) but it needs more further investigation for users who have visual problems.

Keywords: Afan oromo; Word prediction; Word completion; Text production; Disable users

Introduction

As a technology growing fast, Natural Language Processing (NLP) plays a great role in our day to day activity especially in relation to word prediction and auto completion. In Ethiopia, there are more than 80 languages are used as regional and few of it as federal communication. Many people's are using their handheld devices or computers to create different files by their own mother tongue. While writing the files, especially using local languages, it takes time and resource, hence most of file editors are using the English and other language. For example, Afan Oromo which has more than half of the population of a country used as their mother tongue and where there is long vowels and short vowels, needs assisting technology to edit files. In case of long vowels which is vowel repetition, it needs to prediction and completion to save time and resources of users, particularly for disables, poor knowledge of spelling [1].

The absences of these applications in the local languages bring a challenging problem in the society. In rural and urban areas where Hotels, Private Companies and Governmental Organizations are facing a problem while making their identity name, trademark or advertising their product. For example, instead of saying *Daabboo* [Bread], they are saying *Daboo* [Cooperation] with the absence of the letter 'b' which makes completely different spelling, sense, speech and form of the language. In word prediction text entry on mobile phone which is limited to only hand devices [2]. His work presents a word prediction approach based on context features and machine learning. As the result, it shows that the accuracy performance of his system 56.8%. However, in our case we have used different techniques of the n-gram (unigram, bigram, trigram) when there is lack trained data like Afan Oromo.

Additionally, the speed of typing of many secretaries (disable users in this case) when they write Afan Oromo texts is very low and misspelled word that create miscommunication between authors and readers. Hence, the single letter may changes the meaning of the word if misspelled. Furthermore lack of Afan Oromo word auto completion impacts non-native speakers from learning Afan Oromo language. Due to misspelling and low speed of typing, the new Afan Oromo speakers are ashamed from practicing the language. Therefore this study is undertaken to solve the disables problems by providing Afan Oromo word prediction and auto completion.

The developed system was only useful for disable peoples who have lack of fingers to typing and other problems at their hands. And also it can support normal users who want to type their files and want to save their time. As this work describes that targets people with physical disabilities and motion impairments like cerebral palsy, muscular dystrophy, spinal injuries, and other muscular deficiencies. One of the main difficulties faced by such people in interacting with computers is that their word entry is very slow, and the typing process can be tiring. Based on this, our system was cannot be helpful for other patients. And also this system was developed for computer users. However, it cannot be support android, IOS and other environments.

This study was conceptually developed prototype for Afan Oromo words that predict and auto complete words. To develop Afan Oromo word prediction for disabled users and others, we have used java software environment. The researcher used Java program as tool in order to build user interface and it is a platform independent language. As it shown in the result and discussion section, the GUI of this system was developed as it was very easy to use. Hence, the users are may be disabled the researcher consider the issues and make user friendly.

Consequently, the main objective of this work was to develop word prediction prototype for Afan Oromo specifically at word level. Afan Oromo uses Latin based script called Qubee and it has 26 basic characters. The method that used in this work was unsupervised machine learning; hence there is no standardized annotated corpus for training the machines. Generally, guessing the next character or

***Corresponding author:** Workineh Tesema, Department of Information Science, Jimma University, Jimma, 378, Ethiopia, Tel: +251910127829; E-mail: workineh.tesema@ju.edu.et

Received April 13, 2017; **Accepted** April 21, 2017; **Published** April 27, 2017

Citation: Tesema W, Tamirat D (2017) Enhancing the Text Production and Assisting Disable Users in Developing Word Prediction and Completion in Afan Oromo. J Inform Tech Softw Eng 7: 200. doi: 10.4172/2165-7866.1000200

word (or word prediction) is an essential subtask of NLP application, handwriting recognition, augmentative communication for the disabled, and spelling error detection. The motivation behind this work was to bring technology to disable peoples and allow the users to use word prediction present in their language.

Materials and Methods

This section presents the proposed method employed in this work. In order to develop word prediction model for Afan Oromo we followed various step process which involve: (a) text preprocessing which take input and corpus, tokenize to remove stop words and perform normalization. The other one is to (b) extract words to providing the clue about the predicted term using two techniques (Frequency and Recency). Based on the frequency of co-occurrences of the words most frequencies will be listed, to select the candidate won. On the other hand, the word was listed which are recently accessed, then select if it is the candidate words and correct to make complete the misspelled words. In order to guess the next word or term the users should press one or more the starting letter of the term. Hence we have used the most frequently occurring words in the corpus to predict.

This approach was statistically driven, as have been virtually all of the predictive models developed since then. Statistical methods generally suggest words based on:

1. Frequency, either in the relevant corpora or what the user has typed in the past; or
2. Recency, where suggested words are those the user has most recently typed. Such approaches reduce keystrokes and increase efficiency. Even with the best possible language models, these methods are limited by their ability to represent language statistically. In contrast, by using common sense knowledge to generate words that are semantically related to what is being typed, text can be accurately predicted where statistical methods fail.

Training data

To this end, the machine which is an unsupervised approach was trained on the free text corpus. As discussed on the above sections, after the machine trained on the corpus, the system preprocessed the corpus as it discussed. Then, simply the users can press the first letter or letters of his candidate word. According to the nature the algorithm, it will apply and show the words start with the entered letters. Assume that the user enter the first letter of a word, then the system was counted the frequent occurrence of words started with entered letter and make a rank at first. On the other hand, the system can show the recently predicted words as first predicted word. Finally, the system shows the list of predicted words according to their frequency and recency. Then the user can select the words, to make auto-complete and meaningful terms to edit the file.

Implementation tools

As it discussed on the above sections, to develop this system the algorithms were implemented in Java Net Beans IDE 8.02 version software which was open and run on the prepared corpus. The reason why the researcher used this tool is hence Java is free and help us to develop user interface. And also, Java is a general purpose and open source programming language. Moreover, it is optimized for software quality, developer productivity, program portability, and component integration. Lastly, the reason why java selected for this work was it is platform independent language which is after application developed we can run on different operating system of the users.

Results and Discussion

This section describes the result and discussion of the work. A prediction and auto-completion were systems which assist disabled users who have poor knowledge of spelling and physically injured in Afan Oromo. Consequently, an unsupervised machine learning method and N-gram algorithm were used in this work. The given corpus is a sequence of sentences, tries to predict the succeeding word. In the current context, no grammar model or parse trees are used in order to gauge the morphology of the words to be predicted. For every word to be predicted, the algorithm needs to scan through the entire training corpus (Figure 1).

This experiment describes that people with physical disabilities and motion impairments like wrist, hand, arm, fingers, cripple and other muscular deficiencies faced challenging problems. One of the main difficulties faced by such people in interacting with computers is that their word entry is very slow, and the typing process can be tiring [3].

Based on the experimental result, auto complete is a word completion task, so that the user types the first letter or letters of a word and the program provides one or more higher probable words. If the user intended to type is included in the list, the user can select it for example by using the keys (complete button (F7) in our case). If the word that the user want is not predicted, the user must type the next letter of the predicting word. At this time, the word choice(s) is altered so that the words provided begin with the same letters as those that have been selected or the word that the user wants appears it is completed. Word prediction technique predicts word by analyzing the previous word flow for auto completing a word with more accuracy by saving maximum keystroke of any users and reduces misspelling. N-gram language model is an important technique for predicting correct word to complete Afan Oromo word with more accuracy [4].

Many researchers have also tried to incorporate linguistic knowledge by employing other grammatical information like the parts-of-speech tags, parsing trees, root and stem of the words of the English language [5] in order to boost performance of the N-gram prediction. However, in this work, such higher level linguistic information which requires annotated data is not used. The only input and features are the sequence of words from the given corpus. As the experiment shows that in most cases, the N-gram model, with N equal to 1 or 2, seem to work the best word prediction for Afan Oromo.

Since the finding of this work, if the users enter three more letters the accuracy of prediction system is fewer lists of words; hence at Afan Oromo three or more extra letters can make one word. Now if any of these three or four words are not in the training corpus, then the frequency of the word will be zero that means that the word does not exist in our corpus. So in this statistical method if we want to consider these words, then we need a huge data corpus that must contain all the words of the language [3]. So there may arise the problems like many entries in the corpus are with zero frequency and the frequency of a word sequence will be very low [4].

As the experiment shows that a typical interface to a word prediction would be a list holding a number of relevant words picked out by the prediction. The list is then modified and pruned from incorrect alternatives as the user types more letters of the word. If the word prediction list contains the correct word the user can select it and go on to type the next word. For example, a user may have typed the word that has started by "a" and a word prediction presents the possible list shown in Figure 2.

As the experiment shows that, the average accuracy of test terms

was 56.2% for the machine learning approach. From the Figure 2 above shows the predicted word displayed in the provided graphical user interface. If the listed words, does not listed in the predicted list by the user must know as the word is not present in the prepared corpus, the user can still write the next character to the next word. Sometimes, hence the dataset collected from multi resources, there are spelling errors. Unfortunately, our system cannot recognize the problem of misspelling due it needs Afan Oromo spelling checker system. However, if the word is spelled correctly and available in the corpus the performance is very good.

Additionally, our work gives other alternatives for users; hence they are physically limited to writing. This makes our work an easy and user friendly, hence by only one button pressing, users can text product as his/her interest. For instance, to Move Up (F5), to Move Down (F6), to Complete alternative (F7) the work was helpful.

Based on Figure 2, the result shows that the predicted words are predicted based on their frequency occurrences which are ranked according to frequency in the corpus as it shown in the Table 1.

Assume that the user wanted to type the word "*karaa*". Simply the user can press the first spelling of the word on providing interface of the users. After that the system will list all the words started by "***k***". The system inside will count the words started by this spelling and rank it at first if it has a high frequency of occurrence in the corpus. The user can choose the word directly from the list, for example, with the mouse or by pressing the correct spelling. If the word is not in the suggestion list the user can type the next letter of the word, in the case "***k***" and then get a new list with suggestions. Often a space is inserted after the suggested word, which allows the user to continue typing immediately after the prediction. A possible and often used, the choice when implementing a word prediction could be to make the insertion of an alternative automatic when there is enough information to make the decision (Figure 3).

As the experiment shows that, the result apart from being for people with physical disabilities word prediction can also assist individuals with poor spelling to use a greater variety of words [6]. However, although this word prediction can develop writing skills regarding aspects such as word fluency, a variety of words and motivation of writing, the tests completed so far have been inadequate in scope and design [7].

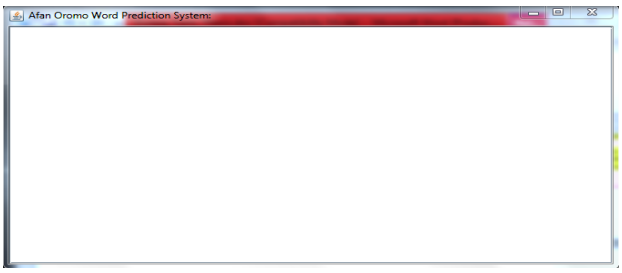

Figure 1: Text editor user interface for afan oromo.

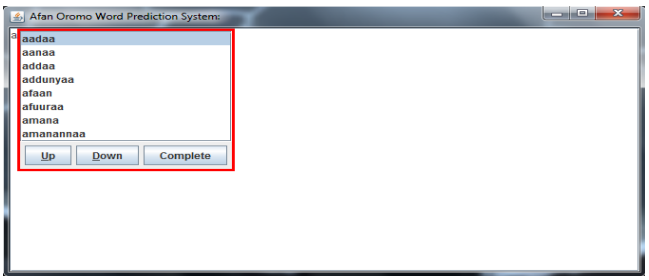

Figure 2: Assistive text production user interface.

Rank	Words
1	Aadaa
2	Afaan
3	Addaa
4	Addunyaa
5	Afuura
6	Ammana
7	Amanannaa

Table 1: Example of word prediction suggestion list.

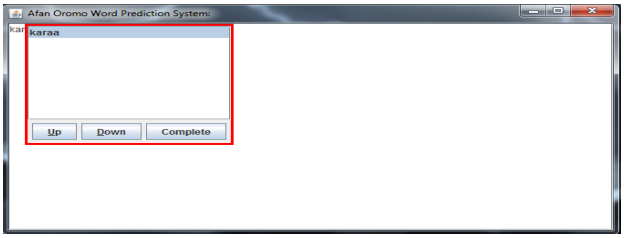

Figure 3: Sample example after three spelling pressed.

Based on the above experiment, the system shows that in Afan Oromo correctly predicted after the users enter two letters [8,9]. To be honest our system can work at one, two, three or four letters, even it was work at sequence of words to make a sentence, but the accurate performance of the system was not much surprising as the experiment argued, that is 83.84%, 61.4%, 54.8% of unigram, bigram and trigram respectively.

Evaluation methods

As a result revealed that word prediction can be a sequential process over time with an input stream of characters. The task is to predict the next character given a string representing the input history. In this work the first character of a string represents oldest input and the last character represents the newest input [10,11]. For example, given the string "*abaaboo*" a good guess for the next character would be 'b' since 'b' follows 'a' which is a prefix of 'b' in the input history. It is well established that there is a close relationship between the tasks of prediction and auto-completion [12].

$$\text{Accuracy} = \frac{\text{Total number of words predicted correctly}}{\text{Total number of words}} \times 100$$

It is not only important to predict the succeeding word theoretically, but sometimes it is possible to calculate the accuracy of the system which is the probability of several possible words is also important. In our case, the frequency of the words helps us to compute the possible probability of the words. The probability of the succeeding word is used in different applications [13]. This measure gives an indication about what is the probability assigned to the correct word as compared to what is the most likely word according to the algorithm (Table 2).

We have used the two metrics of the evaluation method which are precision and recall metrics. The evaluation component exists on its own in order to have the possibility to automatically evaluate the system [14]. It allows testing the predictions under different metrics and formally, its interface is similar to that of the GUI component (Table 3).

Accuracy of Algorithms			
Number of Words Predicted	**Unigram**	**Bigram**	**Trigram**
1	83.84%	61.4%	54.8%
2	75.4%	59.77%	40.7%
3	70.1%	54.9%	31.8%
4	67.9%	52.68%	29.53%
5	61.8%	50.6%	26.89%

Table 2: Accuracy of the model.

No	**Accuracy Performance**	
	Precision	**Recall**
Number of Correctly predicted	90%	73.34%

Table 3: Evaluation of the model.

Conclusion

The overall focus of this research is to investigate word prediction and auto-complete which addresses the problem of poor spelling and completion. Ideally, this can speed up and ease the user's typing of word production. This work is improving and enhancing textual information entry for disabled users has been investigated, with unsupervised approach and user interface proposed and implemented to facilitate and simplify text input for such people. In this work the unsupervised machine learning achieved an accuracy of 90%, 73.34% of precision and recall respectively.

References

1. Yarowsky D (1995) Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, M.A. pp: 189-196.
2. Gudisa T (2013) Design and implementation of predictive text entry method for Afan Oromo on mobile phone, Addis Ababa, Ethiopia.
3. Cockburn A, Siresena A (2003) evaluating mobile text entry with the fasta keypad, people and computers: British computer society conference on human computer interaction, Bath, England.
4. Masudul H, Tarek H, Mokhlesur R (2015). Automated word prediction in bangla language using stochastic language models, bangla. IJFCST 5: 67-75.
5. Wu D, Sui Z, Zhao J (1999) An information-based method for selecting feature types for word prediction. Proc Eurospeech.
6. Sutherland BM (2004) Predictive text entry in immersive environments. Proceedings of the IEEE Virtual Reality.
7. Even-Zohar Y, Roth D (2003) A classification of approach to word prediction. Proceedings of the 1st North American Chapter.
8. Debela T, Ermias A (2011) Designing a rule based stemmer for afaan oromo text. IJCL.
9. Gudisa T (2013) Design and implementation of predictive text entry method for afan oromo on mobile phone.
10. Muzeyn KB (2012) Development of stemming algorithm for Silt'e language text. Thesis faculty of informatics, Addis Ababa University, Addis Ababa.
11. Nancy I, Veronis J (1998). Word sense disambiguation, the state of the art. Computational Linguistics.
12. Tesema W (2016) Afan Oromo sense clustering in hierarchical and partitional techniques. Journal of Information Technology and Software Engineering.
13. Longkai, Li L, Houfeng W, Sun X (2014) Predicting chinese abbreviations with minimum semantic unit and global constraints. Empirical Methods in Natural Language Processing (EMNLP), pp: 1405-1414.
14. Tesema W (2015) Towards the sense disambiguation of afan oromo words using hybrid approach. Jimma University, Jimma, Ethiopia.