Myanmar Spelling Correction Based On N-grams Model

1st Thazin Win
University of Computer Studies,Yangon
Yangon, Myanmar
thazinwin@ucsy.edu.mm

2nd Win Pa Pa
University of Computer Studies,Yangon
Yangon, Myanmar
winpapa@ucsy.edu.mm

Abstract*—This paper describes Myanmar spelling correction intended for real-word errors and non-word errors. There are three main modules in this paper. They are error detection,candidates generation, error correction. Dictionary look up method is used for detecting errors, Levenshtein Distance Algorithm is used for generating candidates and N-grams model is used for correcting errors. There can be human-generated misspellings which can be distinguished into three groups (i) Typographic Errors(Non-word error) (ii)Phonetic Errors(Cognitive error) (iii) Context Errors(Real word errors). This spelling correction can solve all of these three misspellings problem and the main contribution of this paper is to solve the context errors using n-grams model in sentence level. Moreover, this spelling correction can solve the pali misspelling errors. Experimental results show that each of error types can be solved by this spelling correction. The general accuracy of all error types is greater than 85%.

***Keywords*—Levenshtein Distance Algorithm, Myanmar spelling correction, N-grams Model**

I. Introduction

Spell correction for Myanmar Language has more difficulties than that for English. Spell checking detects misspelled words in text. Spell correction gives advice after errors are detected. When only the former works, it is spell checking and when the latter is also involved, it is spell correction. The work in this paper will focus on both tasks. For English spell checking, many studies have been made and good results have been obtained. For Myanmar spell checking,it is still challenging work due to Myanmar writing system. Unlike English,word boundaries are not marked with spaces in Myanmar sentences. So, it is difficult to tokenize. Spelling errors can be classified into two main categories (i)non-word error and (ii)real-word error. The former is one in which the input word is definitely incorrect and cannot be found in dictionary. For example, using “fcrm” rather than “farm”. The latter is one in which the input word is found in the dictionary but is incorrectly used. For example, using “come form” than “come from”. The most common reasons for misspelled and misused errors are caused by phonetic similarity and typing error of Myanmar characters. In this work, first we study the details on Myanmar Language to identify the problem area of Myanmar spell errors and then we develop Myanmar Spell Correction. It consists of three phases: error detection,candidates generation and error correction. Myanmar commission language(MLC) dictionary is used to detect errors. Levenshtein Distance Algorithm is used for generating candidates and N-grams Model is applied for error correction.

This paper is organized as follows: Section describes the related work. Section 3 describes Types of Spelling Errors. Section 4 describes Myanmar Corpus. Section 5 describes Myanmar Spelling Correction. Section 6 describes Language Model. Section 7 describes Experiments. Section 8 describes Conclusion on this work.

II. Related Work

Like many other natural language processing tasks, spell checking is one the most important tasks. Many researches for English spell checking have been done for four decades. Even though other Asian spell checker researches have been done for two decades, Myanmar spell checker research is still challenging work. In this section, we discuss briefly some of related work.

Chinese Spelling Checker System that used N-gram Model and String Matching Algorithm was proposed in [6]. Context-sensitive Spelling Correction was proposed in [4] that use Google Web 1T 5-Gram Information. Myanmar Word Spelling Checking was proposed in [3] that use LevenShtein Distance Algorithm in word level. Myanmar Spell Checking was proposed in [5] that use Levenshtein Distance Algorithm and Naive Bayesian Classifier.

In this paper, we propose a Myanmar Spelling Correction for both non-word errors and real-word errors by applying Myanmar Text Corpus, Myanmar Language Commission(MLC) Dictionary, Levenshtein Distance Algorithm and N-gram Model.

III. Types of Spelling Errors

This spell correction aims to detect and correct spell errors in the text. There are three types of spelling errors as follow.

(i) Typographic Errors(Non-word errors)
(ii) Phonetic Errors(Cognitive errors) and
(iii) Context Errors (Real word errors)

i. Typographic Errors

These errors are mistakes made when typing something. These errors are unintentional errors that happen when we accidentally hit the wrong key on keyboard. For example, typing “teh” when we mean to type “the”. The examples of these misspellings errors are shown in table I.

ii. Phonetic Errors

These errors are usually highly recognizable the intended words because they can spell phonetically. When the writer substituted letters they believe sound correct into a word .The misspelling is pronounced the same as the intended word but the spelling is wrong .For example, typing ‘Furm’ when we

mean to type ‘Farm’). The examples of these phonetic errors are shown in table I.

iii. Context Errors

These errors turn intended word into another word of language. They can produce a real word error. These errors can be seen as a subset of phonetic errors. For example, typing ‘piece’ when we mean to type ‘peace’, where the misspelled word is pronounced the same as the intended word. The examples of these errors are also shown in table I.

TABLE I. TYPES OF MISSPELLINGS

Types of Misspellings	Misspelled Words	Correct Words
Typographic	ကျောင်းသာ	ကျောင်းသား
Typographic	ကြားပန်း	ကြာပန်း
Phonetic	ခြေလှန်း	ခြေလှမ်း
Phonetic	နေလှမ်း	နေလှန်း
Context	အချိန်မှီ	အချိန်မီ
Context	မီခို	မှီခို

IV. MYANMAR CORPUS

Corpus is a large and structured set of texts. Building of the text corpus is very helpful for the development of spell checking. In this work, a monolingual Myanmar text corpus is created manually to apply in Myanmar Spelling Correction. It contains various training sentences including stem words, compound words, derivative words, pali words and personal names. The corpus includes approximately 10,000 Myanmar segmented sentences. The sentences are segmented into word level and collected from news pages. These sentences are written by unicode system and are manually checked for spelling by using Myannmar Language Commission (MLC) Dictionary.

V. MYANMAR SPELLING CORRECTION

Myanmar word boundaries detection is the first step in spell correction because word boundaries are not marked with spaces. So, word segmentation is considered as pre-processing. The input sentences are segmented with spaces by using left to right segmentation as in [2]. And then each word is detected by using Myanmar Language Commission(MLC) Dictionary. Then candidates are generated for each misspelled word by using Levenshtein Distance Algorithm . Finally,mispelled words are corrected with the best candidates by using N-grams Language Model. The processing steps are shown in fig 1.

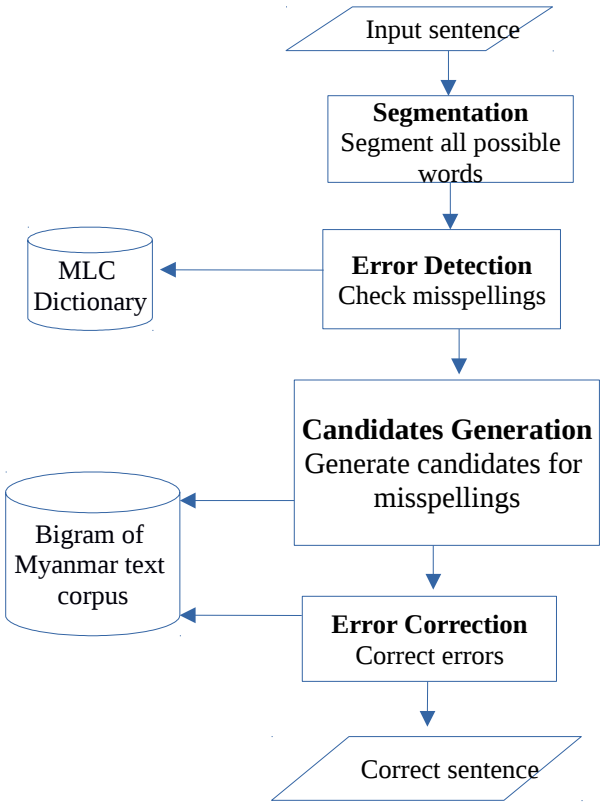

Fig. 1. Myanmar spelling correction

I. Word Segmentation

Word segmentation is considered as pre-processing in Myanmar Spelling Correction. The input sentence is segmented by using left to right segmentation as in [2]. Although the correct sentence can be correctly segmented,the sentence including misspelled words cannot be segmented correctly. Therefore,each segmented word is combined with the next word by using Myanmar corpus. Then, the distance between the combined word and the corpus word is calculated by using Levenshtein Distance Algorithm. If the distance is less than or equal 2 ,it may be a word and can be combined these words. If the distance is greater than 2, these words cannot be combined and can be segmented as each word.

II. Error Detection

Error Detection is the first step of spell correction. The spell correction checks whether segmented word from input sentence is in the MLC dictionary. If the segmented word exits in the dictionary, pass into next word. Unless the segmented one is in the dictionary, it is considered as a misspelled word.

III. Myanmar Dictionary

Myanmar Language Commission (MLC) dictionary is used to detect misspelled errors in correction system. It contains approximately 28,800 words. It includes stem words e.g. သွား, လာ, စား, compound words e.g. မှန်တင်ခုံ, ရေအိုး, derivative words e.g. ဒါရိုက်တာ, ဗက်တီးရီးယား, pali word e.g. မေတ္တာ, သိက္ခာ

IV. Candidates Generation

Candidates generation is the second step of spelling correction system. The system generates a list of possible candidates for every misspelled errors from error detection. The Levenshtein distance algorithm is used to generate candidates. The Levenshtein distance between two words is the minimum number of single-character edits (insertions,

deletions,reversion or substitutions) required to change misspelled word into the correct word. All the words that are at a Levenshtein distance of 2 or less than 2 from misspelled words are generated.

V. Levenshtein Distance Algorithm

The Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. Levenshtein distance may also be referred to as edit distance. There are other popular measures of edit distance such as Longest Common Subsequence (LCS) as in [8], Jaro-Winkler Distance as in [9]. The longest common subsequence (LCS) distance allows only insertion and deletion, not substitution. The Hamming distance allows only substitution, hence, it only applies to strings of the same length. The Jaro-Winkler distance allows only transposition. Among these algorithms, Levenshtein Distance Algorithm is the best algorithm for this spelling checking. This algorithm allows the transposition of two characters as an operation and produces the number of operations need to be transformed from one word to another. LD is a measure of the similarity between two strings, which we will refer to as the source string (s) and the target string (t). It is used in some spell checkers to operate Insert, Delete, Reverse and Substitute transformations. At the end, the bottom-right element of the array contains the answer. The resulted distance is the number of deletions, insertions, reversion or substitutions required to transform s into t.

VI. Error Correction

Error correction is the final step of spelling correction system . The system chooses the best candidate in candidates generation from second step and correct .N-grams model is used for error correction .N-gram model is now widely used in natural language processing such as speech recognition,machine translation and spell checking.

VI. Language Model for Error Correction

I. N-grams Model

N-gram model is a type of probabilistic language model for predicting the next item in a sequence. The item can be phonemes, syllables, letters or words according to the system. In this spell correction, a word can be considered as an item. An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" ; size 3 is a "trigram". As n-gram length increases, the amount of times we will see any given n-gram will decrease. And so, bigram is used for this system. Bigram probabilities are calculated by Laplace Smoothing method because it can avoid zero probability and increase a small positive number.

II. Bigram Model

The bigram model approximates the probability of a word given all the previous words by using only the conditional probability of one preceding word.

And so, bigram model is used to predict the conditional probability of the next word,thus the probability is calculated by the following equation:

$$P(W_n|W_{n-1}) = \frac{Count(Wn-1, Wn)+1}{Count(Wn-1)+V} \quad (1)$$

The probability of a word depends only on the previous word is also known as Markov assumption.

Markov models are the class of probabilistic models that assume that we can predict the probability of some future unit without looking too far in the past. Each step of Myanmar Spelling Correction is shown in fig 2, fig 3, fig 4.

Input:ယခုအခါဗိုလ်ချုပ်ပြတိုက်ကိုဖွင့်လှစ်ထားပြီဖြစ်သည်

Segment: ['ယခုအခါ', 'ဗိုလ်ချုပ်','ပြတိုက်', 'ကို', 'ဖွင့်လှစ်', 'ထား', 'ပြီ', 'ဖြစ်သည်']

Error Detection: ်ပြတိုက်

Candidate Generation: ['ပြတိုက်']

Error Correction: ယခုအခါ ဗိုလ်ချုပ် ပြတိုက် ကို ဖွင့်လှစ် ထား ပြီ ဖြစ်သည်

Fig. 2. Example of typographic error output by proposed system

Input: သမင်သည်ကြော့ကွင်းတွင်မိနေသည်

Segment: ['သမင်', 'သည်', 'ကြော့ကွင်း', 'တွင်', 'မိနေသည်', '']

Error Detection: ကြော့ကွင်း

Candidate Generation: ['ကျော့ကွင်း']

Error Correction: သမင် သည် ကျော့ကွင်း တွင် မိနေသည်

Fig. 3. Example of phonetic error output by proposed system

Input: မနက်စာပေါင်မုန့်ထောပတ်သုပ်ညစာအဖြစ်ချင်းသုတ်စားကြမည်

Segment: ['မနက်စာ', 'ပေါင်မုန့်', 'ထောပတ်သုပ်','ညစာ', 'အဖြစ်', 'ချင်းသုတ်', 'စား', 'ကြမည်', '']

Error Detection: ထောပတ်သုပ် ချင်းသုတ်

Candidate Generation : ['ထောပတ်သုတ်'] ['ချင်းသုပ်']

Error Correction : မနက်စာ ပေါင်မုန့် ထောပတ်သုတ် ညစာအဖြစ် ချင်းသုပ် စား ကြမည်

Fig. 4. Example of context error output by proposed system

VII. EXPERIMENTS

I. *Experiment Data Set*

In order to verify the validity of the system,some experiments are based on the monolingual Myanmar corpus. There are two types of sentences for experimental data. They are open sentences and closed sentences. The open sentences are the ones that do not include in the training corpus. The closed sentences are the ones that include in the training corpus.

1,000 sentences are tested to evaluate the experimental results of the system. The average number of words including in one sentence is 9 words. Experiment data set is shown in table II.

TABLE II. EXPERIMENT DATA SET

Types of Error	Open Sentence	Close Sentence	Total
All	350	650	1000
Typographic	100	150	250
Phonetic	200	400	600
Context	50	100	150

I. *Experimental Results*

Spell correction performance is evaluated in terms of accuracy,precision, recall and F1-score which are the common way to measure spelling checking system's performance .Accuracy is simply the ratio of correctly predicted observation to the total observations .Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations to the all observations in the system. F1-score is the mean of precision and recall. Experimental results are shown in table III.

TABLE III. SPELLING ERROR OF GENERATION RESULTS

Types of Error	Accuracy	Precision	Recall	F1-score
All	94.67%	98.38%	90.79%	94.43%
Typographic	92.07%	98.57%	85.19%	91.39%
Phonetic	95.73%	98.51%	92.74%	95.54%
Context	85.42%	91.67%	81.48%	86.28%

VIII. CONCLUSION

In this paper,we proposed spelling correction for Myanmar language. This spelling correction can check and correct Typographic errors, Phonetic errors, and Context errors and the performance of the system is measured in terms of accuracy, precision, recall, F1-score. A monolingual Myanmar Text Corpus is built for Myanmar Spelling Correction system. We applied Dictionary Look Up Method for error detection, Levenshtein Distance Algorithm for generating candidates and N-gram Model for error correction. The system can check and correct every sentence that has any number of errors and any types of errors. Moreover,Pali words can be handled by this spelling correction. Myanmar spelling correction gave satisfiable results to apply in Myanmar NLP applications. The average accuracy of all errors is 92.07% typographic errors, 95.73% phonetic errors, 85.42% context errors. Therefore, the performance of context errors is not better than that of typographic errors and phonetic errors. This is because there are only 10,000 training sentences in the corpus. If there are more than 10,000 training sentences in the corpus, the spelling correction can provide a better performance.

REFERENCES

[1] Myanmar Words Commonly Misspelled and Misused Book, Department of Myanmar Language Commission, Ministry of education, Union of President Myanmar July, 2003

[2] Win Pa Pa, Words Segmentation for Myanmar, University of Computer Studies, Yangon, 2008

[3] Nwe Zin Oo, Myanmar Words Spelling Checking Using Levenshtein Distance Algorithm, University of Computer Studies, Yangon, 2010

[4] Youssef Bassil and Mohammad Alwani, Context -sensitive Spelling Correction Using Google Web 1T 5-Gram Information

[5] Aye Myat Mon, International Journal of Science and Research (IJSR), 2013

[6] Jui Feng Yeh, Chinese Spelling Check based on N-gram Model and String Matching Algorithm, Department of Computer Science and Information Engineering, National Chia-YI University, 2017

[7] V.J. Hodge, J.Austin, A comparison of standard spell checking algorithms and a novel binary neural approach, Department of Computer Science, University of York , UK, 2003

[8] Asif Mohaimen, Charkraborty, Bangla OCR Post Processing Word Based Longest Common Subsequence Techique, Department of Computer Science and Engineering, 2017

[9] Hicham Gueddah, Abdellah Yousfi, Mostafa Belkasm, The filtered combination of the weighted edit distance and Jaro-Winkler distance to improve spell checking Arabic texts, University Mohammed V of Rabat, Morocco, 2015