N-Gram-Based Spelling Checker for Myanmar Noun Words

Naing Lin Oo
University of Computer Studies, Yangon, Myanmar
nainglinoo84@gmail.com

Abstract

Myanmar spelling checker system is the important task in one of NLP application such as information retrieval and in most other NLP applications. This system is based on N-Gram approach for Myanmar Spelling Checker. The process is a fundamental task in document processing allowing the automated handling of documents in electronic form. One difficulty in handling some classes of documents is the presence of different kinds of textual error, such as spelling. This system must work reliably and effectively on all input and thus must tolerate some level of these kinds of problems. The aim of the paper is to understand the basic concept of N-Gram approach and to study how to check Myanmar Noun words spelling checkers with data mining approach by using N-Gram-Based approach and then to obtain accurate and relevant result with reduce time consuming. This system can be checked efficiently for Myanmar Noun words spelling.

Keywords: Myanmar Noun words, N-Garms approach, textual error.

1. Introduction

With the increasingly widespread use of computers and the Internet in Myanmar, large amounts of information in Myanmar languages are using various word processing software packages. Therefore, Myanmar spelling checker system is becoming an urgent need in the Myanmar context [1].

Documents are generated with various word processing software packages [2] and are subjected to various kinds of automatic scrutiny e.g.; as well as to manual editing and revision. Many other documents, however, do not have the benefit of this kind scrutiny and thus may contain significant numbers of errors of various kinds. So, the system describes an N-Gram-Based approach that is tolerant of textual error such as spelling checker [3, 4]. This system will work very well for Myanmar language spelling checking.

A wide number of languages are spoken by human beings in the world and most of the people prefer to have information in their own language. A person who is not particularly familiar with English should able to check with his native language.

Although, there are so many have spelling checker system [5] none of these can't fully support for checking with Myanmar language font. For these reasons, this system will develop to assist checking for Myanmar spelling [6, 7]. The users given query will be checked through the Myanmar Unicode documents according to N-Gram approach.

The rest of this paper is organized as follows. Section 2 introduces background knowledge for Myanmar language. Section 3 describes N-Grams based theory. Section 4 describes the N-grams applications area. Section 5 presents overview and implementation of the system. Section 6 describes the evaluation and analysis of system. Finally, the paper is concluded in Section 7.

2. Background Knowledge for Myanmar Language

Myanmar Language (formerly known as Burmese), is a member of Sino-Tibetan language family, according to the international language family trees. The Myanmar writing system derives from a Brahmi-related script borrowed from South India in about the eight century for Mon language. This system employs a letter to represent each syllables and consists of 33 symbols for consonants, 11 vowels symbol and various symbols to represent vowel sounds, tone marks, specified symbols and punctuation marks. It is the official language of Myanmar, where 32 million people speak it as their first language. Some people in China and India also speak Burmese [7].

Burmese language is a syllabic writing system that differs from English and many other western languages which are based on alphabetic. In addition, Burmese Script is written from left to right and there are no space between words, although informal writing often contains spaces after each phrase. ASCII based fonts have been used for the Myanmar language data processing before Unicode. So, there were a lot of implementations and there is no standard among font encoding. After Unicode was invented, it is an international standard for all language. Myanmar Unicode was approved starting from Unicode 3.x. The range for Myanmar Unicode is from (U+1000 to U+109F). In Unicode 4.x, Unicode consortium defined standards for Myanmar Unicode encoding standards. Nowadays, Myanmar Unicode fonts are widely used in Myanmar Language web pages and documents.

3. N-Grams Based Theory

An N-Grams are sequences of characters or words extracted from a text. N-grams can be divided into two categories: 1) ***Character based*** and 2) ***Word based.***

3.1. Character based N-Grams

A *character N-grams* is a set of *n* consecutive characters extracted from a word. The main motivation behind this approach is that similar words will have a high proportion of N-grams in common. The system also append blanks to the beginning and ending of string in order to help with matching beginning-of-word and ending-of-word situations (the system will uses the underscore character ("_") to represent blanks).

Typical values for *n* are 2, 3 or 4 and so on. These correspond to the use of *bi-grams, tri-grams* and *quad-grams* respectively. For example, the word **"ဆန်း"** would be composed of the following N-grams.

* *Bi-Grams* : _ဆ ၊ ဆန ၊ န် ၊ ်း ၊ း_
* *Tri-Grams* : _ဆန ၊ ဆန် ၊ န်း ၊ ်း _ ၊ း_ _
* *Quad-Grams* : _ဆန် ၊ ဆန်း ၊ န်း _ ၊ ်း _ _ ၊ း_ _ _

In general a string of length *k,* padded with blanks, will have *k+1 bi-grams, k+1 tri-grams, k+1 quad-grams* and so on [3, 4]. *Character based N-grams* are generally used in measuring the similarity of character string. If the system counts N-grams that are common to two strings, it gets a measure of their similarity that is resistant to a wide variety of textual errors. Spellchecker, Stemming, OCR error correction are some of the applications which use *character based N-grams.*

3.2. Word based N-grams

Word N-grams are sequence of *n* consecu-tive words extracted from text [4]. *Word level N-gram* models are quite robust for modeling language statistically as well as for information retrieval without much dependency on language [7].

4. N-grams Applications Area

N-grams approach have been used in many applications area such as Speech recognition, Spelling collection, Handwriting recognition and Information retrieval are some major areas where " N-grams" based statistical language modeling can play an important role.

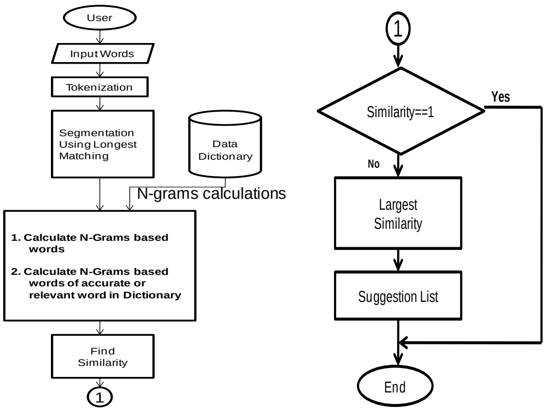
Character "N-gram" matching for computing a string similarity measure is widely used technique in information retrieval, stemming, spelling and error correction, text compression, language identification and text search and retrieval. The N-gram based similarity between two string is measured by **Dice's Coefficient** [4]. Consider the word **"ဆန်း"**; whose *tri-grams* are: _ဆန ၊ ဆန် ၊ န်း ၊ ်း _ ၊ း_ _ .

To measure the similarity between the words **"ဆန်း"** *and* **"ဆန်"**, the system can uses **Dice's Coefficient** in the following way. First, find all the *tri-grams* from the word **"ဆန်"** are _ဆန ၊ ဆန် ၊ န်_ ၊ _ _ and so on. The number of unique *tri-grams* in the word **ဆန်း** is 5 and in the word **ဆန်** is 4. There are 2 common *tri-grams* in both the words. Then similarity of these words are calculated with **2C/(A+B)** according to **Dice's Coefficient**, where A and B are the number of unique *tri-grams* in the pair of words; C is the number of common *tri-grams* between the pair.

5. Overview and Implementation of the System

5.1. Overview of the System

This system is based on N-Grams approach to check for Myanmar Noun words spelling. Therefore, users initially give Myanmar Noun words to the system. And then, the system can be make tokenization for input words and make segmentation from segmented syllables into set of word by syllable level longest matching algorithm. After these process, calculation for words spelling by using N-Grams-based approach. If user input has textual error, the system will gives suggestion lists to the user to obtain accurate and relevant result. Figure 1. shows the overview of the system.



User
Input Words
Tokenization
Segmentation Using Longest Matching
Data Dictionary
N-grams calculations
1. Calculate N-Grams based words
2. Calculate N-Grams based words of accurate or relevant word in Dictionary
Find Similarity
1
1
Similarity==1
Yes
No
Largest Similarity
Suggestion List
End

Figure 1. The System Overview

5.2. Implementation of the System

5.2.1. Tokenization and word Segmentation

This system is composed with two phases; tokenization and segmentation. It uses longest string matching for both.

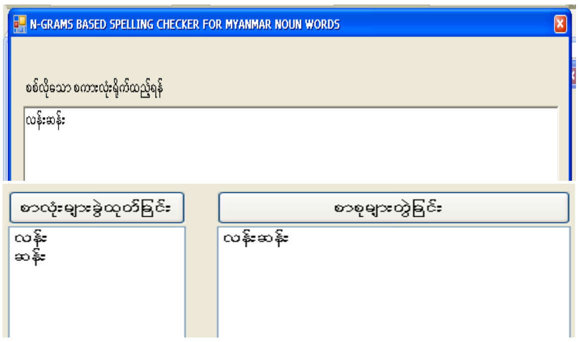
In tokenization process, a syllable is a basic sound unit or a sound. A word can be made up of one or more syllables. For Myanmar text tokenization, the system uses syllable level longest matching algorithm show in the Figure 2.

```
sub syllabification {
Load the set of syllables from syllable-file
Load the words to be processed
Store all syllables of length j in N_j where j=10…1
for-each word do
    length ← length of the words
     pos   ← 0
     while (length >0)do
           for j=10..1 do
                 for-each syllables in N_j do
                      if string-match word (pos,pos+j) with syllable
                                        Syllable found . Mark syllable
                                        pos ← pos+j
                                        length ← length-j
                      End if
                 End for
           End for
     End while
   Print syllabified string
 End for
 }
```

Figure 2. Longest matching pseudo code

For a trial, of the system; users input Myanmar Noun words to the system.

The system can be processed tokenization for input words, the system will tokenizes the string of characters into single syllable (word) [such as လန်း+ဆန်း] by using longest string matching algorithm from syllable file for Myanmar text tokenization. And then, word segmentation process can be made after tokenization process. Segmentation process is to merge the segmented syllables into set of word [such as လန်းဆန်း] by using above algorithm.


N-GRAMS BASED SPELLING CHECKER FOR MYANMAR NOUN WORDS
စစ်လိုသော စကားလုံးရိုက်ထည့်ရန်
စာလုံးများခွဲထုတ်ခြင်း
စာစုများတွဲခြင်း

လန်းဆန်း

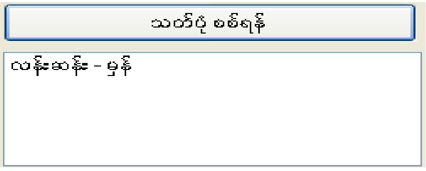
လန်း
ဆန်း

လန်းဆန်း

Figure 3. Two examples of syllable tokenization and word segmentation.

5.2.2. Calculation of N-Grams for Words Spelling

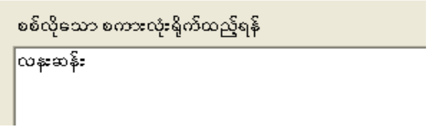
N-Grams calculation process, have two categories 1) Calculate based on N-Grams from input words. 2) Calculate based on N-Grams of accurate or relevant word in dictionary. And then, the system finds their similarity by using ***Dice's Coefficient*** formula. If the value of their similarity is equal to one, the user input words is correct.



လန်းဆန်း - မှန်

Figure 4. Check Spelling

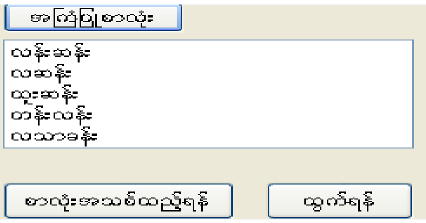
If the value of their similarity is not equal to one, user input words is not correct.



လနးဆန်း

Figure 5. Situation of input words error

But, the process finds their largest similarity and then it gives suggestion list to the users to obtain accurate and relevant result show in Figure 6.



လန်းဆန်း
လဆန်း
ထူးဆန်း
တန်းလန်း
လသာခန်း

Figure 6. Suggestion Lists

In this way, the spelling checker system check the input completely and correctly. Some of the valid words [such as လန်းလန်းဆန်းဆန်း ၊ ထူးထူးဆန်းဆန်း ၊ ထူးထူးခြားခြား ၊ လန်းဆန်းစိုပြေ and so on] are not listed in dictionary although these words are usually used in every day life of the real world. If the system checks this spelling, it will gives the spelling is error for user. Therefore, if user wants to check this spelling this system has been able to add new words for users.

6. Evaluation and Analysis of System

Four kinds of nouns are used for analysis: These are,

1. Common Nouns
2. Proper Nouns
3. Abstract Nouns
4. Collective Nouns

6.1. Common Nouns

A Common Noun is a name which is used for all things of the same kind or class. For example, the name 'teacher' (ဆရာ), 'student' (ကျောင်းသား), 'girl' (မိန်ကလေး),'sea'(ပင်လယ်), and so on. It is a name common to all .

6.2. Proper Nouns

A Proper Noun is the name of one particular person or thing. For example, ‘Maung Win’ (မောင်ဝင်း), is a boy’s own special name. It does not belong to all boys. So the name ‘Maung Win’ (မောင်ဝင်း), is called a Proper Noun. In the same way, the name ‘city’ (မြို့), is a Common Noun but the name (မြိတ်မြို့), which is the special name of a ‘city’ is a Proper Noun.

6.3. Abstract Nouns

An Abstract Noun is the name of a quality , action, or condition, that is, something that it cannot see or touch. It is the name of a thing that it can only think of or feel. For example, the thing called ‘music’ (တေးဂီတ), ‘pity’ (ဂရုဏာ), is something that it cannot see or touch. The name ‘music’, and ’pity’ are therefore an Abstract Noun.

6.4. Collective Nouns

A Collective Noun is the name of a number (or collection) of persons or things taken together as one thing, such as, ‘family’ (မိသားစု),‘crowd’ (လူစုလူဝေး) , ‘class’ (အတန်း), ‘army’ (စစ်တပ်), etc.

Therefore, the system has tested with the number of four kinds of nouns words spelling which contain more than 35000 words in dictionary. Four types of Noun are checked by Myanmar spelling checker system. To evaluate the system, 100 words of each type of Nouns are put randomly. These words are tested with correct spelling and incorrect spelling in different typing. The system can check correctly 95.5% of Common Nouns words, 50% of Proper Nouns words, 75.99% of Abstract Nouns words and 55% of Collective Nouns words.

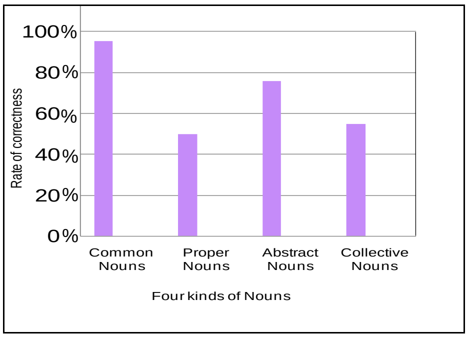


Figure 7. Percentages of Correctness and Nouns types

7. Conclusion

The N-grams-based matching provides the derived words from its very nature: since every string is decomposed into small parts, leaving the remainder intact. N-grams approach have been used in many applications area such as Speech recognition, Spelling collection, Handwriting recognition and Information retrieval are some major areas where “ N-grams” based statistical language modeling can play an important role. This system is implemented a framework for a textual error of Myanmar spelling checker based on N-grams model.

This system can be applied efficiently and effectively in N-grams-based Myanmar Noun words Spelling Checker. This system will be add to check sentences, multi words or even paragraphs for user requirements at the future.

References

[1] A.Zamora, “Automatic Detection and Correction of Spelling Errors in A Large Database”, Journal of the American Society for Information Science.

[2] Hla Hla Htay and Kavi Narayana Murthy, “Myanmar Word Segmentation using Syllable level Longest Matching”, Department of Computer and Information Sciences, University of Hyderabad, India.

[3] John M.Trenkle and William B. Cavnar, “N-Gram-Based Text Categorization”, Environment Research Institute of Michigan P.O.Box 134001 Ann Arbor MI 48113-4001.

[4] P Majumder, M Mitra and B.B. Chaudhuri, “N-gram: a language independent approach to IR and NLP”, Computer vision and pattern recognition Unit. Indian Statistical Institude, Kolkata.

[5] R.C. Angell, G.E. Freund, and P, Willette, “Automatic Spelling Correction Using Trigram Similarity Measure”, Inf,Proc. Mgt.18,255,1983.

[6] Thin Thin Naing, “Noun and Pronouns”, THALUN Bookstore, Yangon.

[7] Thin Zar Win and Khin Marlar Tun, “N-gram based Vector Space Approach for Myanmar Web Searching”, University of Computer Studies, Yangon.