DEVELOPING A SPELL CHECKER FOR MYANMAR UNICODE SYSTEM

Soe Moe Aye, Khin Khin Lay
University of Computer Studies, Yangon
soemoeaye1033@gmail.com

ABSTRACT

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. Natural-language processing is a very attractive method of human-computer interaction. Spelling-checkers have become an integral part of most text processing software in NLP. The focus of the system is mainly on the non-word error or misspelling. Misspells are detected when the specific word does not belong to the language words domain. The duty of spell checker is to detect the errors and suggests the best similar words. In order to detect the errors, it is necessary to model the related knowledge of language words for the system. After error detection, it is necessary to suggest the similar words with their correction methods for the system. The purpose of the system is to detect all possible errors from the Myanmar Unicode text document and give the best suggestions to fix the misspelled words. To do this, this system uses Lexicon - for error detection and Edit Distance technique - for error correction.
Keywords: Spelling checking, error detection, levenshtein distance, real word errors

1. INTRODUCTION

Nowadays, having general access to Internet as a universal phenomenon, electronic texts development and using the text based query interfaces have made the existence of assisting tools for text manipulation. Spell checkers intended to cover a large part of own native language have a vast application zone. In some researches, Lexicons and their representations have been studied in details. Some researches have focused only on the Lexicons containing stems of words, and inflection rules and morphology have been utilized. Knowledge representation of words of a language is one of the significant issues in each system related to NLP. Spell Checking has a long history in Computer science and nowadays spell checking system is as an essential part for almost all application software. The word-error can belong to one of the two distinct categories, namely, non-word error and real-word error. Let a string of characters separated by spaces or punctuation marks be called a candidate string. [2]

A candidate string is a valid word if it has a meaning. Else, it is a non-word. By real word error we mean a valid but not the intended word in the sentences, thus making the sentence syntactically or semantically ill-formed or incorrect. In both cases, the problem is to detect the erroneous word and suggest the corrected alternatives or automatically replace it by the appropriate word. There are several issues to be addressed in the error correction problem. The first issue concerns the error patterns generated by different text generating media such as typewriter and computer keyboard, typesetting and machine printing, OCR system, speech recognizer output, and of course, handwriting. Usually, the error pattern of one media does not match with that of the order.

In the paper, the proposed system overview and possible spelling errors are presented in Section 3 and system architecture, levenshtein edit distance are illustrated in Section 4. In Section 5, experimental results are shown and conclusion, limitation and further extensions are included in Section 6.

2. RELATED WORK

Several approaches have been considered for detecting spell errors and correction these errors until now. Error detection is the challenging process in the natural language processing approach. Various techniques and approach are applied to perform spelling checking process.

"XUXEN: A Spelling Checker/Corrector forBasque Based on Two-Level Morphology" has been proposed by Agirre E., Alegria I., Arregi X.,Artola X., Diaz de Ilarraza A., Maritxalar M., Sarasola K.Informatika Fakultatea and Urkia M.U.Z.E.I. In this paper, they present the spelling checker with two level morphology techniques and Lexicon. This spelling checker will accept as correct any word which allows a correct standard morphological breakdown. When a word is not recognised by the checker, it is assumed to be a misspelling. "Reversed word dictionary and phonetically similar word grouping based spell-checker to Bangla text" has been shown by Bidyut Baran Chaudhuri, Computer Vision and Pattern Recognition Unit, India Statistical Institute. This paper demonstrates the spell check program using reserved word dictionary and n-grams approach. There are many approaches for spelling checking. In the proposed system, headwords is used to detect misspell word and Levenshtein edit distance algorithm is used to give suggestion words. [1] [3]

3. SYSTEM OVERVIEW

A spell checker has a set of routines for scanning text and extracting words and an algorithm for comparing the extracted words against a known list of correctly spelled words. The first step is to prepare a language-related lexicon and error pattern extracting. In next step, error patterns will be modeled, in order to detect the errors and proposing the suitable error suggestions. The system performs the process of spell checking for Myanmar Unicode text documents. The first portion of the spell checker is to detect the possible errors. The next step intends to suggest the best similar words (Suggestions) if it is misspelled. To do the former process, Lexicon, where includes the knowledge representation of words of a Myanmar language is used to detect the errors. For the later process, Levenshtein edit distance algorithm is used to generate the suggestions for the misspelled words.

3.1. Basic Syllable Structure

Burmese writing system is a syllabic writing system where a unit of writing can be denoted by a syllable, which consists of Vowels or Consonants and vowels. These scripts are Structure of Myanmar Syllables according to Unicode 4.0. The system is implemented by using Myanmar 2 font Unicode standard. [7]

Name	Encoding	Example
kinzi	<U+1004, U+1039>	
consonant	[U+1000–U+1021]	
subscript consonant	<U+1039, [U+1000–U+1019, U+101C, U+101E, U+1020, U+1021]>	
medial ya	<U+1039, U+101A>	
medial ra	<U+1039, U+101B>	
medial wa	<U+1039, U+101D>	
medial ha	<U+1039, U+101F>	
vowel sign e	U+1031	
vowel sign u, uu	[U+102F, U+1030]	
vowel sign i, ii, ai	[U+102D, U+102E, U+1032]	
vowel sign aa	U+102C	
anusvara	U+1036	
atha (killer)	<U+1039, U+200C>	
dot below	U+1037	
visarga	U+1038	

Figure1. Myanmar language encoding in Unicode4.

3.2. System Flow Chart

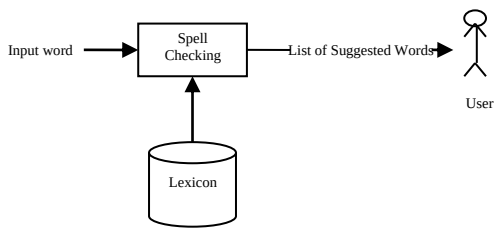

Figure 2. System flow chart

In the above figure, the input to the system is Myanmar Unicode text word. The spell checking program performs the each word to be matched with Lexicon. And then generate the list of Suggested words to the user using Levenshtein edit distance algorithm. [8]

Components of a Spellchecker

- Lexicon
- A bunch of algorithms for
- Lexicon lookup i.e. error detection
- Error correction
- Ranking of corrections.
- Morphological Preprocessing

In the system the isolated word error is detected using headwords in Lexicon. In linguistics, the lexicon of a language is its vocabulary, including its words and expressions. More formally, it is a language's inventory of lexemes. The lexicon includes the lexemes used to actualize words. Lexemes are formed according to morpho-syntactic rules and express sememes. A lemma in morphology is the canonical form of a lexeme. *Lexeme* refers to the set of all the forms that have the same meaning, and *lemma* refers to the particular form that is chosen by convention to represent the lexeme. In lexicography, this unit is usually also the citation form or headword by which it is indexed.

3.3 Error Patterns

The error pattern issue of each media concerns the relative abundance of insertion, deletion, substitution and transposition errors. In order to achieve the main goal of spell checking, which is error detection and correction; it is needed to store a proper integration between Lexicon and the structure of error pattern models.

3.3.1 Spelling errors

Spelling errors can generally be divided into two types:

Typographic errors occur when writer knows the correct spelling of the word but mistypes the word by mistake. These errors are mostly related to the keyboard. The causes of the spelling mistakes are Keyboard Adjacencies, Shift-key Characters, Phonetic Similarity, and Visual Similarity.

Cognitive errors (also called orthographic errors) occur when writer does not know or has forgotten the correct spelling of a word.

Substitution Error: Using a letter instead of the other

အမှားစကားလုံး - ကွါန်တော်

အမှန်စကားလုံး - ကျွန်တော်

Deletion Error: Unintended elimination of one or more letters

အမှားစကားလုံး - ကျောင်းသး

အမှန်စကားလုံး - ကျောင်းသား

Insertion Error: Unintended insertion of a letter in a word.

အမှားစကားလုံး - ကောလိပ်စ်

အမှန်စကားလုံး - ကောလိပ်

Transposition Error: Transposition of two adjacent letters

အမှားစကားလုံး - ဆမရာ

အမှန်စကားလုံး - ဆရာမ

3.4. Levenshtein Edit Distance(LED)

Levenshtein distance (LD) is a measure of the similarity between two strings, which refer to as the source string (s) and the target string (t). The distance is the number of deletions, insertions, or substitutions required to transform s into t. For example,

If s is "test" and t is "test", then LD(s,t) = 0, because no transformations are needed. The strings are already identical.

If s is "test" and t is "tent", then LD(s,t) = 1, because one substitution (change "s" to "n") is sufficient to transform s into t.

The greater the Levenshtein distance, the more different the strings are.[6]

4. SYSTEM ARCHITECTURE

Simple spell checkers operate on individual words by comparing each of them against the contents of a lexicon, possibly performing on the head words. If the word is not found it is considered to be an error, and an attempt may be made to suggest a

word that was likely to have been intended. One such suggestion algorithm is to list those words in the lexicon having a small Levenshtein distance from the original word. In the system, we use head words over 30000 words that are defined by Myanmar Natural Language Processing in Hlaing Campus.

4.1 The Algorithm

- Compute edit distance between erroneous word and all head words.
- Select those dictionary words whose edit distance is within a pre specified threshold value.
- Present these words as suggestions [5]

Table 1. Edit distance algorithm

Step	Description
1	Set n to be the length of s. Set m to be the length of t. If n = 0, return m and exit. If m = 0, return n and exit. Construct a matrix containing 0..m rows and 0..n columns.
2	Initialize the first row to 0..n. Initialize the first column to 0..m.
3	Examine each character of s (i from 1 to n).
4	Examine each character of t (j from 1 to m).
5	If s[i] equals t[j], the cost is 0. If s[i] doesn't equal t[j], the cost is 1.
6	Set cell d[i,j] of the matrix equal to the minimum of: a. The cell immediately above plus 1: d[i-1,j] + 1. b. The cell immediately to the left plus 1: d[i,j-1] + 1. c. The cell diagonally above and to the left plus the cost: d[i-1,j-1] + cost.
7	After the iteration steps (3, 4, 5, 6) are complete, the distance is found in cell d[n,m].

Table 2. Demonstration of without error

							်				

		ေ	ာ	ျ	ာ	င		း	သ	ာ	း
ေ	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ာ	**1**	0	1	2	3	4	5	6	7	8	9
ျ	**2**	1	0	1	2	3	4	5	6	7	8
ာ	**3**	2	1	0	1	2	3	4	5	6	7
c	**4**	3	2	1	0	1	2	3	4	5	6
်	**5**	4	3	2	1	0	1	2	3	4	5
း	**6**	5	4	3	2	1	0	1	2	3	4
သ	**7**	6	5	4	3	2	1	0	1	2	3
ာ	**8**	7	6	5	4	3	2	1	0	1	2
း	**9**	8	7	6	5	4	3	2	1	0	1
	10	9	8	7	6	5	4	3	2	1	**0**

There is no error/ distance between the two words "ကျောင်းသား" and " ကျောင်းသား " with distance value 0 of lower right corner.

Table 3. Deletion error with edit distance 1

		ေ	ာ	ျ	ာ	c	် း	း	သ	ာ	း
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ေ	**1**	0	1	2	3	4	5	6	7	8	9
ာ	**2**	1	0	1	2	3	4	5	6	7	8
ျ	**3**	2	1	0	1	2	3	4	5	6	7
ာ	**4**	3	2	1	0	1	2	3	4	5	6

c	**5**	4	3	2	1	0	1	2	3	4	5
်ှ	**6**	5	4	3	2	1	0	1	2	3	4
ုံ	**7**	6	5	4	3	2	1	0	1	2	3
သ	**8**	7	6	5	4	3	2	1	0	1	2
ုံ	**9**	8	7	6	5	4	3	2	1	1	**1**

Deletion error is described in table 3 with distance 1 between "ေကျာင်းသား" and "ေကျာင်းသး" in calculate in Levenshtein.

Table 4. Insertion error with distance

		ေ	ာ	ာ	ာ	ိုိ	ပ	ိုိ
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
ေ	**1**	0	1	2	3	4	5	6
ာ	**2**	1	0	1	2	3	4	5
ာ	**3**	2	1	0	1	2	3	4
ာ	**4**	3	2	1	0	1	2	3
ိ	**5**	4	3	2	1	0	1	2
ပ	**6**	5	4	3	2	1	0	1
်	**7**	6	5	4	3	2	1	0
စ	**8**	7	6	5	4	3	2	1
်	**9**	8	7	6	5	4	3	**2**

Insertion error is demonstrated in table 4 between "ေကာလိပ်" and "ေကာလိပ်စ်" with edit distance algorithm.

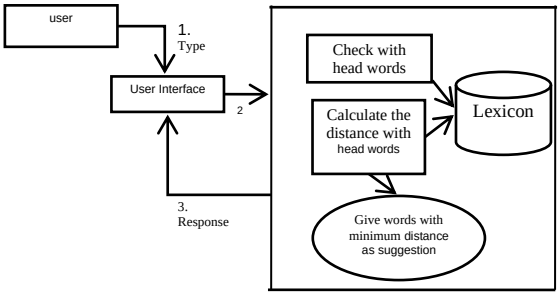

Figure 3. System architecture

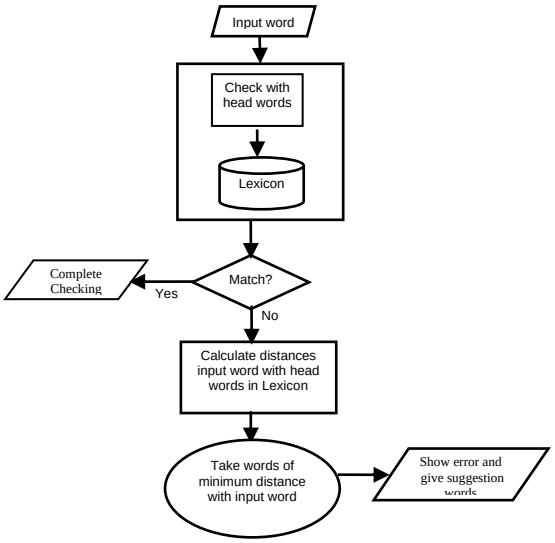

Figure 4. Data flow diagram of the system

5. EXPERIMENTAL RESULT

The user can type the input word to check its spelling. The system firstly tokenizes words and compare against in headwords in Lexicon. If input word is not matched with predefined head words, it will be an error and the user can see the suggested

words list with minimum edit distance value. The evaluation of spelling checker for isolated-word error detection is depend on Lexicon size and finding the accuracy for single error misspellings.
Lexicon size: Our lexicon contains over 30000 words. The accuracy for single error misspellings is fully implemented. In the system phonetic, typographical, OCR generated types of error are not considered to handle.

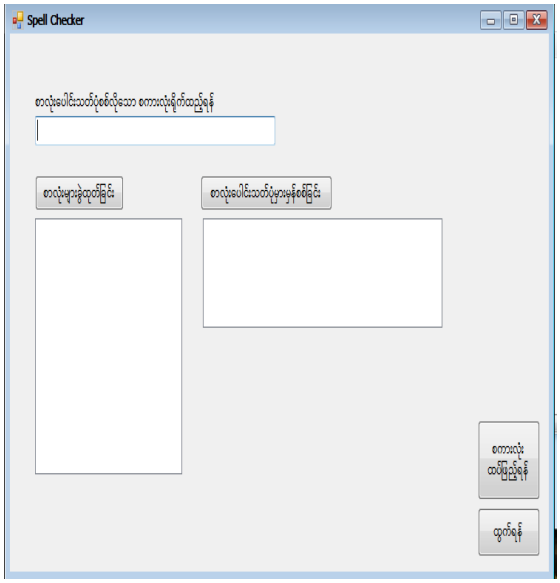

Figure 5. Main Interface System

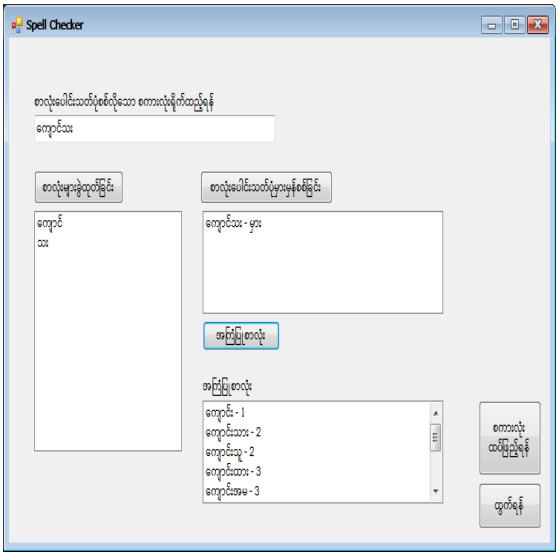

Figure 6. Check the input word with head words and show suggestions

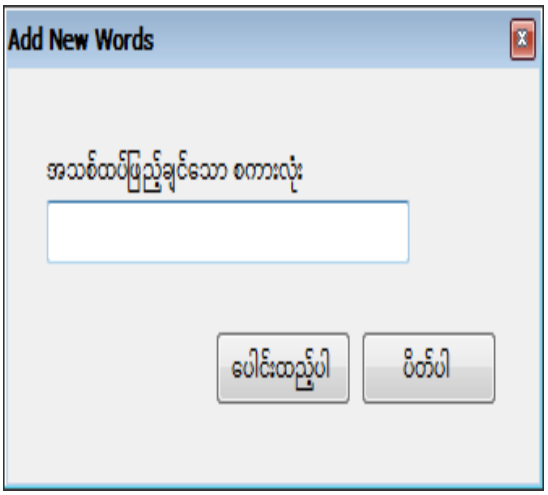

Figure 7. Add new words

The user can add new more words to the head words in lexicon for further processing as show in figure 7.

6. CONCLUSION

The spell checker for Myanmar Unicode language is implemented in the system. The spell checking process is an essential component of Natural

Language Processing (NLP). It is necessary to have a native language spell checker for Myanmar. It has a vast application zone and it can be used in other application areas such as desktop application and tools, OCR and information retrieval. The typographical errors is mainly considered in the system and it will be successful extends to the rest of the NLP processes for Myanmar Language.

6.1. Limitation

The system is only implemented on the head words (over 30000 words) defined by Myanmar natural language processing team. Myanmar 2 keyman software is used to type the spell-checked word. When we can face typing sequence error in predefined head words in Lexicon, the system cannot detect the dedicated error of input word. Again, phonetic similarities are not being considered in the system.

6.2. Further Extensions

Several natural language processing approaches and detection methods have been applied to check word from the headwords. To error is human and to correct is spellchecker.In this paper describes the Myanmar Spelling checker using Myanmar 2 Unicode font. It has a vast application zone and it can be used in other application areas to extend. By appropriately integrating techniques from each of these disciplines, useful new methods for detecting word or string can be developed.

REFERENCES

[1] Agirre E., Alegria I., Arregi X.,Artola X., Diaz de Ilarraza A., Maritxalar M., Sarasola K.Informatika Fakultatea and Urkia M.U.Z.E.I "XUXEN: A Spelling Checker/Corrector forBasque Based on Two-Level Morphology", 20080/20009, Donostla

[2] Barari Loghman , QasemiZadeh1 Behrang "CloniZER Spell Checker Adaptive, Language Independent Spell Checker", *Digital Clone, Speech and Language Department,* 6th floor, No 880, College cross, Enqelab Ave, Tehran, Iran

[3] B. B. Chaudhuri, “Reversed word dictionary and phonetically similar word grouping based spell-checker to Bangla text”, Proc. LESAL Workshop, Mumbai, 2001.

[4] http://www.merriampark.com/ld.htm

[5] http://www.codeproject.com/KB/recipes/Levenshtein.aspx

[6] http://www.mgilleland.com/ld/ldjavascripthtm

[7] http://www.myanmarnlp.net.mm

[8] http://www.nist.gov/dads/HTML/Levenshtein.html.

.