Myanmar (Burmese) String Similarity Measures based on Phoneme Similarity

Khaing Hsu Wai, Ye Kyaw Thu, Hnin Aye Thant, Swe Zin Moe and Thepchai Supnithi

***Abstract*— String similarity measurement is useful for a wide range of applications. It performs an important role in machine learning, information retrieval, natural language processing, error encoding, and bioinformatics. Measuring string similarity is also a basic and fundamental operation of data science, important for data cleaning and integration. Applications such as spell checking, duplicate finding, searching similar words, and retrieving tasks use string similarity. Moreover, Grapheme-to-Phoneme (G2P) conversion is the necessary task of predicting the pronunciation of a word given its graphemic or written form. In this study, string similarity metrics have been calculated for Burmese (Myanmar language) based on phoneme similarity and phonetic similarity. Similarity distance is measured between the datasets and query words, both of which are converted with G2P model and with the phonetic encoding mapping tables. As previous string similarity approaches are not suitable for fuzzy string matching of tonal-based Burmese, measuring string similarity based on phoneme similarity and phonetic mapping approaches are proposed in this study.**

***Index Terms*—Myanmar character, Burmese, String similarity metrics, Phonetic Similarity, Grapheme-to-Phoneme (G2P), Ripple Down Rules-Based (RDR)**

I. Introduction

MEASURING string similarity is widely studied in natural language processing (NLP). String similarity metrics help to find similar words according to a given query. NLP applications such as text-to-speech, machine translation, spell checking, and information retrieval calculate string similarity metrics to find how similar the strings are. It is a fundamental operation in many applications of machine learning. Languages are interesting, and each language has its own features and writing systems. In the literature, several approaches have been proposed for string similarity. Most of them are character-based metrics and associated with English or European languages. For Burmese (language in Myanmar), we need to consider new approaches together with the existing string similarity metrics. Burmese is a tonal-based language and also a very rich language [21]. It has 33 consonants, and the consonants are combined with vowels and medials to form syllables. In Burmese, not only one character can form a word (e.g., "ကဝ", dance in English) but also one syllable can form a word (e.g., "ကြိုက်", like) in English). Additionally, there are many phonetically similar sounds of characters and words in Burmese. Grapheme-to-Phoneme (G2P) conversion is about predicting the pronunciation of words given only the spelling. G2P conversion models are also very important for NLP, automatic speech recognition (ASR) and text-to-speech (TTS) developments. Most of the G2P conversions are supervised learning approaches where we have to clean the annotated data and perform some data preprocessing steps.

In our experiment, phonetic mapping and sound mapping have proposed and have applied G2P mapping to convert the strings. We introduced a new approach based on the idea of Soundex, the best-known phonetic encoding algorithm, for retrieving phonetically similar words by calculating the string similarity distance. We have collected two datasets: one dataset contains the confusion pairs of words with real spelling mistakes, and another is a manually developed word similarity dataset. We evaluated six measures (cosine distance, Damerau-Levenshtein distance, Hamming distance, Jaccard distance, Jaro-Winkler distance, and Levenshtein distance) on two datasets, with and without the proposed mappings. According to our results, all three mappings outperformed the existing approaches for retrieving Myanmar words with similar pronunciations.

II. Related Work

To the best of our knowledge, there is only one proposal that measured phonetic similarities of Myanmar Internationalized Domain Names (IDNs) [1]. To retrieve phonetically similar Myanmar IDNs, IPA (International Phonetic Alphabet)-Soundex functions were used for matching character values based on their phonetic similarities of Burmese. The normalized similarity method is capable of measuring similarity not only in a single language, but also in a cross-language comparison [2].

The Myanmar characters ultimately descend from a Brahmic script, either Kadamba or Pallava [4]. Likewise, most of the major Indian languages such as Devanagari (e.g., Hindi, Marathi, Nepali), Bengali (Bengali and Assamese), Gurmukhi (Punjabi), Gujarati, Oriya, Tamil, Telugu, Kannada, and Malayalam use scripts that are

Khaing Hsu Wai, Hnin Aye Thant and Swe Zin Moe are with Faculty of Information Science, University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar.

Ye Kyaw Thu and Thepchai Supnithi with National Electronics and Computer Technology Center (NECTEC), Thailand.

Corresponding Authors: khainghsuwai@utycc.edu.mm and yktnlp@gmail.com

Manuscript received December 21, 2019; accepted March 6, 2020; revised April 22, 2020; published online April 30, 2020.

derived from the ancient Brahmi script. They have approximately the same arrangement of the alphabet, are highly phonetic in nature, and a computational phonetic model was proposed for them [3]. It mainly consists of a model of phonology (including some orthographic features) based on a common alphabet of these scripts, numerical values assigned to these features, a stepped distance function (SDF), and an algorithm for aligning strings of feature vectors. The SDF is used to calculate the phonetic and orthographic similarity of two letters.

For grapheme-to-phoneme conversion, Myanmar pronunciation patterns are discussed with examples [5]. As a basis for pronunciation mapping, the Myanmar Language Commission (MLC) Pronunciation Dictionary is used to convert grapheme to phoneme [6]. However, it is necessary to extend the dictionary with foreign pronunciations. Some of the mappings are needed to modify to ensure consistency of syllable order and to facilitate mapping the syllables to the IPA. The main difference between the mapping used in the MLC dictionary and G2P mapping is that G2P mapping produces sequences of phonemes in the same order as they are spoken.

The relative performance of different machine learning techniques on Myanmar G2P conversion is also discussed in [7] for the conversion of grapheme to phoneme words. Seven G2P conversion approaches are evaluated on a manually tagged Myanmar phoneme dictionary.

III. String Similarity Metrics

String similarity determines how similar two strings are. Various studies on string similarity have been carried out for different languages. In the literature, many methods to measure the similarity between strings have been proposed. Each method has its own features useful for NLP. Most similarity metrics are used to reduce minor typing or spelling errors in words or syllables in pronunciation. Based on the properties of operations, string similarity metrics can be divided into several groups.

Edit distance-based metrics estimate the number of operations needed to transform one string to another. A higher number of operations mean less similarity between the two strings.

For token-based methods, the expected input is a set of tokens rather than complete strings. The purpose is to find similar tokens in both sets. A higher number of similar tokens mean more similarity between the sets. A string can be transformed into a set of tokens by splitting it using a delimiter.

In sequence-based methods, the similarity is a factor of common substrings between the two strings. The algorithms try to find the longest sequence that is present in both strings. The more of these sequences found, the higher is the similarity score.

A. Levenshtein Distance

The Levenshtein distance [8], also known as edit distance, returns the minimum number of edit operations in terms of the number of deletions, insertions, or substitutions required to transform the source string to the target string. A higher number of edit operations means less similarity between two strings. For example, the edit distance between “cat” and “dog” is 3. There are three edit operations needed to transform “cat” into “dog”. For Myanmar language, “Fate”-“ကံ”(kan) and “ကန်”(kan) (exact pronunciation with “ကံ” but different spelling and “kick”, “lake” in English), two edit operations are required. The Levenshtein distance is perfect for finding similarity of small strings, or for a small string and a big string, where the editing difference is expected to be a small number. The Levenshtein distance is defined recursively, as shown in Eq. (1).

$$dis_{a,b}(i,j) = \begin{cases} 0 & \text{if i=j=0} \\ i & \text{if j=0 and i>0} \\ j & \text{if i=0 and j>0} \\ min = \begin{cases} dis_{a,b}(i-1,j)+1 \\ dis_{a,b}(i,j-1)+1 \quad \text{otherwise} \\ dis_{a,b}(i-1,j-1)+1(a_i \neq a_j) \end{cases} \end{cases} \tag{1}$$

B. Damerau-Levenshtein Distance

The Damerau-Levenshtein distance is an algorithm that is similar to the Levenshtein distance; however, it additionally counts a transposition between adjacent characters as an edit operation [9]. For example, to transform string “CA” to string “ABC”, the Levenshtein distance counts three edits, whereas the Damerau-Levenshtein distance is 2. For Burmese, the Levenshtein distance between “ကလေး” (“baby”) and “ကောလေး”(wrong spelling of “baby”) is 3, whereas the Damerau-Levenshtein distance is 2. Variations of this algorithm assign different weights to the edit based on the type of operation, phonetic similarities between the sounds typically represented by relevant characters, and other considerations.

C. Hamming Distance

The Hamming distance between two strings of equal length measures the number of positions with mismatching characters [10]. The Hamming distance only applies to strings of the same length. It is mostly used for error correction in fields such as telecommunication, cryptography, and coding theory. For example, the Hamming distance between “apple” and “grape” is 4, and the distance between “ဘဝေ”(“father”) and “ဘဘော”(wrong spelling of “father”) is 1.

D. Jaro-Winkler Distance

The Jaro-Winkler distance is another string metric that measures an edit distance between two sequences [11]. The score ranges from 0 to 1, where 0 is “no similarity” and 1 is “exactly the same strings”. The Jaro-Winkler distance is used to find duplicates in strings, because the only operation that it considers is to transpose the letters in

a string. Eq. (2) describes the Jaro-Winkler distance dj of two given strings s1 and s2, where m is the number of matching characters, and t is half of the number of transpositions.

$$d_j = \begin{cases} 0 & \text{if m=0} \\ \frac{1}{3}(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}) & \text{otherwise} \end{cases} \quad (2)$$

E. Cosine Similarity

The cosine similarity between two vectors is a measure that calculates the cosine of the angle between them [12]. By calculating the cosine angle between the two vectors, we can decide if the vectors are pointing to the same direction or not. Two vectors with the same orientation have a cosine similarity of 1, which means that the two strings are equal. For two strings "ဇနီးမောင်နှံ" ("husband and wife") and "ကလေး" ("baby"), the cosine similarity is 0, but for "ဇနီးမောင်နှံ" ("husband and wife") and "စနီးမောင်နှံ" (wrong spelling of "husband and wife"), the similarity distance is 0.75, which is nearly 1. Eq. (3) shows the formula of cosine similarity.

$$similarity(A, B) = \frac{A.B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

F. Jaccard Similarity

The Jaccard similarity measures similarities between sets [13].It is defined as the size of the intersection divided by the size of the union of two sets. For example, for sets A= [1, 2, 3] and B= [1, 2, 4, 5], the Jaccard similarity is 0.4. The Jaccard similarity is calculated according to the following equation (4).

G. Soundex Algorithm

The Soundex algorithm is a phonetic algorithm [14].It is based on how close two words are depending on their pronunciation. For example, the code for "Flower" and the code for "Flour" is "F460" according to the Soundex encoding table, because they have the same pronunciation. Based on the idea of the Soundex algorithm, we propose phonetic and sound mappings for Burmese. All mappings aim to find words based on their phonetic similarity.

IV. Proposed Mappings

String similarity algorithms have some difficulties with Burmese because it is a tonal-based language and is composed of vowels, consonants, and medials. With Myanmar alphabets, many words have the same pronunciation but different meanings (e.g., "ကံ", "luck" in English and "ကန်", "lake" in English). Moreover, some words have similar pronunciations and different meanings (e.g., "ခုနစ်", "seven" in English and "ခုနှစ်", "year" in English). To consider phonetically similar words, we have proposed Phonetic Mapping and Sound Mapping for Myanmar words.

A. Phonetic Mapping

In our proposed methods, the first mapping is the phonetic mapping. Words with the same pronunciation are grouped together. For example, "ကလေး" and "ခလေး" have the same pronunciation. Therefore, "က" (Ka) and "ခ" (Kha) are clustered to "က" (Ka) group. Likewise, other consonants with same pronunciation, such as "ဂ" (Ga) and "ဃ"(Gha), "ပ" (Pa) and "ဖ"(Pha), "ဗ" (Ba) and "ဘ" (Bha) are put together as groups, respectively, and some diacritics, such as "ွ" (Wa Hswe) and "ှ" (Ha Hto), tone marks such as "့" (Aukmyit), "္" (Myanmar sign Virama) are considered to be removed. Mapped characters are using both Myanmar and English alphabets for simple reading and an easier practical implementation. The details of the phonetic mapping table are shown in Table I.

Char	Mapped Char	Char	Mapped Char
က ခ	က	ွ ှ	(delete)
ဂ ဃ	ဂ	ဣ ဤ ီ ဳ ဴ ည်	i
စ ဆ	စ	က် ပ် တ်	d
ဒ ဈ	ဒ	န် မ် ံ	n
ဋ တ	တ	ဲ ရ်	e
ဌ ထ	ထ	ဥ ဦ ု ူ	u
ဠ ဍ	ဠ	ဿ ျ	r
ဏ နာ	န	ဧ ဪ	a
ဒ ဓ	ဒ	့ း	(delete)
ပ ဖ	ပ	္	(delete)
ဗ ဘ	ဘ	ဩ ော် ဪ ော်	o
ယ ရ	ရ	၎င်း ဂ	ဂ
လ ဠ	လ	၊ ။	s
သ ဿ	သ	ံ င် ဉ	in
ျ ြ	y	?!.*-=#"<>[],+-	s

TABLE I: Phonetic Mapping

B. Sound Mapping

The second mapping is the sound mapping. This mapping is similar to the phonetic mapping, but the main difference is in processing Myanmar consonants. As the name of the sound mapping suggests, consonants that have the same movements of mouth, lips, and tongue, are grouped. For example, "က ခ ဂ ဃ င ဟ အ" (Ka Kha Ga Gha Nga Ha A) are clustered to "က" (Ka) group, "ည ဉ" (NyaGyi NyaLay) are clustered to "ည"(Nya) group, "ပ ဖ ဘ ဗ မ" (Pa Pha Ba Bha Ma) are clustered to "ပ" (Pa) group, "ယ ရ" (YaPetLet YaGauk) are clustered to "ရ" (Ya) group. The details of the sound mapping are shown in Table II.

C. Grapheme to Phoneme Mapping

The groups of characters according to their pronunciation based on unaspirated, aspirated, voiced and nasal tone are shown in Table III. [5]. Myanmar syllables containing unaspirated and aspirated consonants are pronounced as voiced consonants depending on the neighbor-

Char	Mapped Char	Char	Mapped Char
က ခ ဂ ဃ င ဟ အ	က	ွ ှ	(delete)
ည ဉ	ည	့	(delete)
စ ဆ ဇ ဈ	စ	ကဲ့ တဲ့	d
ဋ ဌ ဍ ဎ ဏ	တ	နဲ့ မဲ့	n
ပ ဖ ဗ ဘ မ	ပ	ဲ ရဲ	e
ယ ရ	ရ	ဥ ဦ ူ ဪ	u
လ ဠ	လ	ော ော်	r
သ ဿ	သ	ဧ ေ	a
ျ ြ	y	း ဳ	(delete)
။ ၊	s	ဩ ဪ ဩာ ဪာ	o
၎း ဂ	ဂ	ဣ ဤ ီ ဦ ဧည်	i
်င် ဥ်	in	?!.*-=#"<>[],+-	ဒ

TABLE II: Sound Mapping

ing context. The proposed group of Myanmar pronunciation features was designed to allow g2p conversion models to take these dependencies into account.

Unaspirated	Aspirated	Voiced	Nasal
က /k/	ခ /kh/	ဂ /g/ ဃ /g/	င /ng/
စ /s/	ဆ /hs/	ဇ /z/ �… /z/	ည/ဉ /nj/
ဋ /t/	ဌ /ht/	ဍ /d/ ဝ /d/	ဏ /n/
တ /t/	ထ /ht/	ဒ /d/ ဓ /d/	န /m/
ပ /p/	ဖ /hp/	ဗ /b/ ဘ /b/	မ /n/
ယ /j/	ရ /j/|/r/	လ /l/ ဝ /w/	သ /th/
	ဟ /h/|/r/	ဠ /l/ အ /a/	

TABLE III: Groups of Myanmar Consonants

V. Experiments

We compare 6 similarity measures on three mappings. They are Levenshtein, Hamming, Jaro-Winkler, Damerau-Levenshtein, Cosine, and Jaccard similarities. We conduct two experiments with two datasets that we have collected.

A. Datasets

We have collected two datasets: *Spelling Mistake Confusion Pairs* and *Word Similarity Dataset.*

1) Spelling Mistake Confusion Pairs

The dataset of spelling mistake confusion pairs was developed based on real-world spelling errors. Mainly, we collected general-domain text, especially from Myanmar news and social media websites, such as BBC (British Broadcasting Corporation) Myanmar, VOA (Voice of America) Myanmar, Facebook, and emails during March 2018 and July 2019. The dataset contains 2,381 pairs (i.e., 4762 words). Some examples of confusion pairs are as follows:

- ကိုကိုကြီး – ကိုကိုကြီး
- ကောင်းကောင်း – ကောင်းကောင်း
- ကောင်းကျပါတယ် – ကောင်းကြပါတယ်
- ခွင့်မလွတ်ပါနဲ့ – ခွင့်မလွတ်ပါနဲ့
- ငါ့စီ – ငါ့ဆီ
- စီးပွားေရး – စီးပွားရေး
- စွဲချက်တင်နိုင်ုင်သောကြောင့် –စွဲချက်တင်နိုင်သောကြောင့်
- တောင်ပန်အပ်ပါတယ် – တောင်းပန်အပ်ပါတယ်
- တိုင်ပြည်ချစ်စိတ် – တိုင်းပြည်ချစ်စိတ်
- ဒေါ်အောင်ဆန်းဴစုကြည် –ဒေါ်အောင်ဆန်းစုကြည်
- နက်နက်ရိှုင်းရိှုင်း – နက်နက်ရှိုင်းရှိုင်း
- ပြဿနာတက်မှာဆိုးပြီး – ပြဿနာတက်မှာစိုးပြီး
- ၂၀၁၀ – ၂၀၂၀
- ဝှူးရှူး – ဝူရှူး
- အဆောက်အဦ – အဆောက်အဦး

During the dataset collection, we found that some of the spelling mistakes are caused by encoding conversion between partial Unicode named "Zawgyi" and other Unicode fonts such as "Myanmar3" and "Padauk" (e.g., "ကိုကိုကြိီး – ကိုကိုကြီး", "တနလာနေ့ – တနလ်ာနေ့", "နိုင်ငံရေးဇ။ိ – နိုင်ငံရေးဇ၏ "). Moreover, the spelling mistakes based on pronunciation similarity (e.g., "ကျေးပွန်းစွား – ကျေးပွန်စွာ", "ငါ့စီ – ငါ့ဆီ", "ပြဿနာတက်မှာဆိုးပြီး – ပြဿနာတက်မှာစိုးပြီး") and shape similarity (i.e., glyph) of Myanmar characters are also found(e.g., "စီးပွားေဂး – စီးပွားရေး","ဝှူးရှူး–ဝူရှူး", "အဆောက်အဦ – အဆောက်အဦး").All the confusion pairs generally have one-to-one relationship between misspelled and correct words; thus, we assumed it is very useful for evaluating on our three mappings. However, this dataset has few homophones and rhyme words; therefore, it is not suitable for measuring pronunciation similarity.

2) Similar Pronunciation Dataset

We developed the similar pronunciation dataset to evaluate similarity scores provided by our three mappings. Based on the correct Myanmar word, we manually added one homophone and three more rhyme words, such as "Hat:Bat", "Fun:Sun", "Honey:Money". For example, the first column word "မြူးတူး"("festivity" in English) is the correct word, the second column "မြူးထူး"is the homophone word, and the other following columns "ဂျူးဖူး", "ကူးလူး" and "ပြူးတူး" are three rhyme words of the first column word (see Table IV). We collected 200 pairs for the similar pronunciation dataset, with 1,000 words in total. Examples of how three mappings encoded the words or strings can be seen as follows. All of these examples have different spellings but same meanings and make same sound.

- Example for Phonetic Mapping
  ဒဲပေါက်ထမင်း – `ɜb ʊard ɑ ɯk`
  ဓဲပေါက်ထမင်း – `ɜb ʊard ɑ ɯk`
- Example for Sound Mapping
  ပစ္စည်း – `ɒɛɛi`
  ပစ်စည်း – `ɒɛɛi`
- Example for Grapheme-to-Phoneme Mapping
  လေ ပြည် လေ ညင်း – `lei pji lei njin:`
  လေ ပြီ လေ ငြင်း – `lei pji lei njin:`

B. Ripple Down Rules-based (RDR)

Ripple-Down Rules (RDR) is an approach to building knowledge-based systems (KBS) incrementally, while the KS is in routine use [15] [16] present a new error-driven approach to automatically restructure transformation rules in the form of a Single Classification Ripple Down Rules (SCRDR) tree [15] [17]. A SCRDR can be notated as a triple < rule, X, N >, where X and N are the exception

Correct Word	Homophone	Rhyme1	Rhyme2	Rhyme3
မြူးတူး	မြူးထူး	ဂျူးဖူး	ဘူးလူး	ပြူးတူး
ပြဋ္ဌာန်း	ပြဋ္ဌာန်း	ရှုစမ်း	ကြာပန်း	ကျင်း
တချို့	တစ်ချို့	အချို့	သချို့	နှစ်ချို့
ကြွေးမြီ	ကျွေးမြီ	ခွေးမြီး	ကြွေးမှီ	ချွေးသီး
ဂဃနဏ	ဂဂနန	ခခယယ	မမထထ	ခခရရ
လက်ရွေးစင်	လက်ရွေးစဉ်	လက်ယွေးစင်	ရက်ရွေးစင်	လက်ရွေးဇင်

TABLE IV: Examples from the Similar Pronunciation Dataset

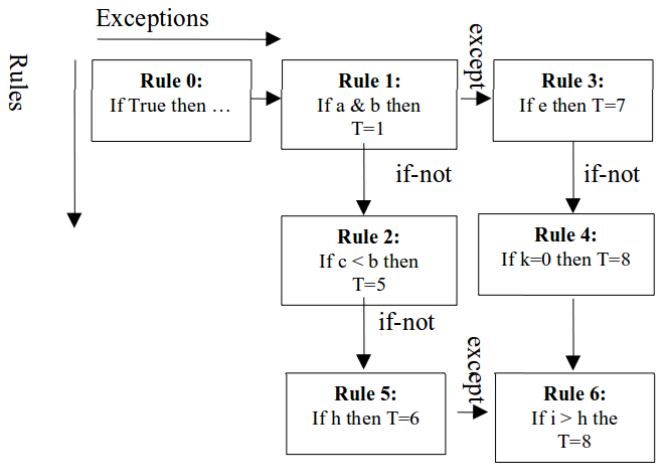


Fig. 1: A binary tree of Single Classification Ripple Down Rules

RDR and the succeeding RDR (i.e. if-not rules) respectively [18]. Cases in SCRDR are evaluated by passing a case to the root (Rule 0 in Figure 1). At any node in SCRDR tree (i.e. Rule 1 to Rule 6), if the condition of a node n met, the case is passed on to the exception child of n using except link if it exists. Otherwise, the case is passed on to the if-not child of n. In the SCRDR approach, a conclusion is always given by the last node in the process. To ensure that a conclusion is always given, the root node (also known as default node) is usually set up with the condition which is always satisfied.

For Grapheme-to-Phoneme conversion, we use Ripple Down Rules-based (RDR) to convert the strings from grapheme to phoneme. Preprocessing steps are prepared before parsing to the RDR model. Words are changed to syllable-level strings with Syllable Break method. We trained G2P with RDR model with syllable segmented words and thus alignment was done on syllable units. Some of the examples of G2P conversion can be seen as follows.

- ကွန် ပျူ တာ- kun pju ta (“Computer” in English)
- ကျန်း မာ ရေး- kyan: ma jei: (“Health” in English)
- လေ ပြည် လေ ညင်း- lei pji lei njin: (“Breeze” in English)

C. Evaluation

For the evaluation, we measured string similarity on each pair from both original datasets: “Spelling Mistake Confusion Pairs” and “Similar Pronunciation Dataset”. Next, we encoded or converted the original data with Phonetic Mapping, Sound Mapping and Grapheme-to-Phoneme Mapping. After that string similarity for two datasets is measured again. Finally, we counted the correct words or similar words based on the three thresholds “<=1”, “<=2”, and “<=3” for “Levenshtein, Damerau-Levenshtein, and Hamming distance measures” and “>=0.9”, “>=0.7”, and “0.5” for “Jaro-Winkler, Cosine, and Jaccard distance measures”.

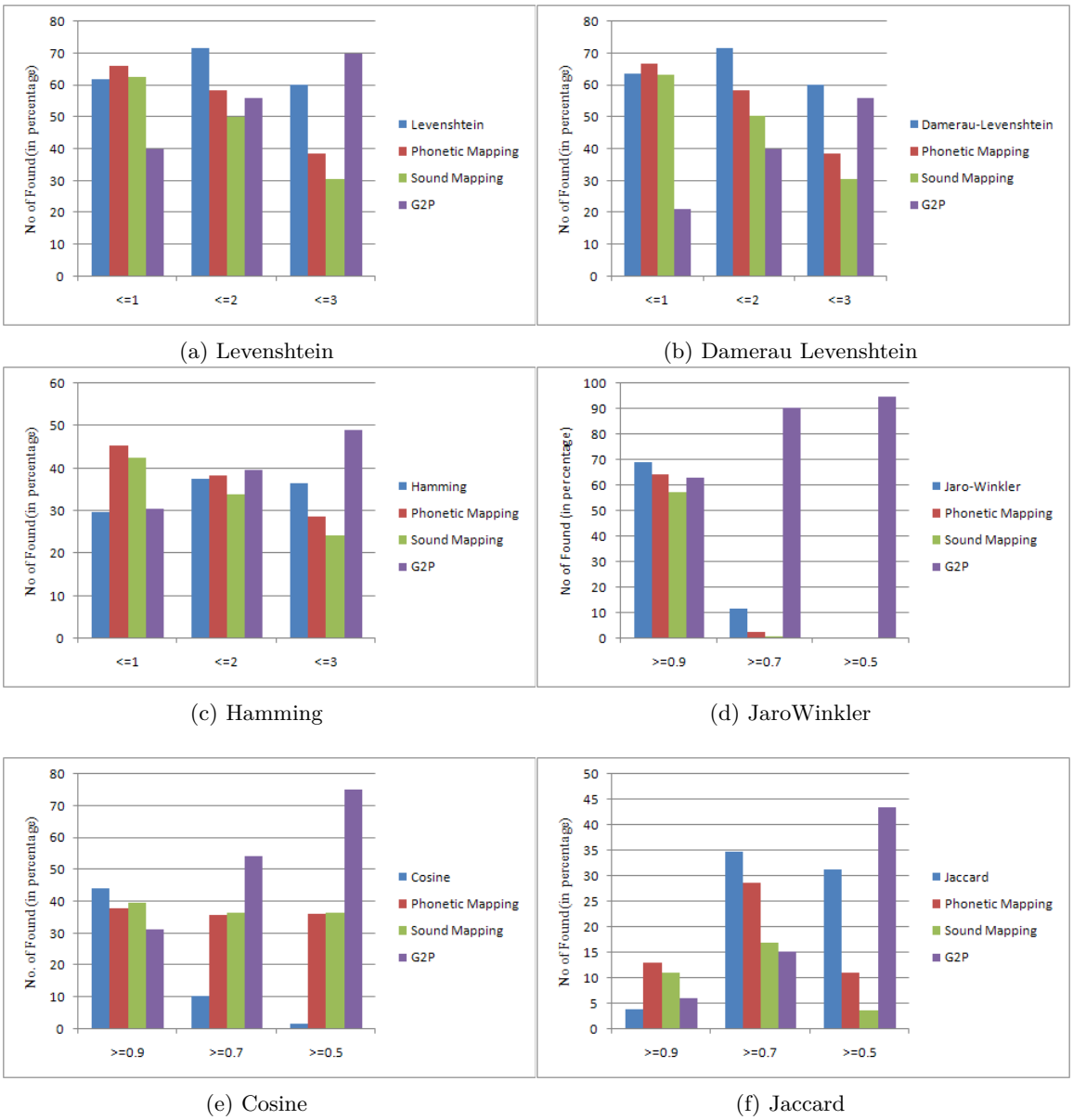
VI. Results and Discussion

The number of correct words found for six similarity measures on the “Spelling Mistake Confusion Pairs dataset” is shown in Figure 2. According to these experimental results, string similarity measurement base on G2P mapping gave a better word correction rate on all existing distance measures (Cosine, Jaccard, Jaro-Winkler, Levenshtein and Hamming) for threshold >=0.7 or >=0.5 but expect for Damerau-Levenshtein distance while the phonetic mapping and the sound mapping also achieved higher or comparable results, except for the Jaro-Winkler similarity.

In general, phonetic mapping and sound mapping are lower than raw Myanmar text input for thresholds “<=2” and “<=3” (“>=0.7”, “>=0.5” for Jaro-Winkler and Cosine similarity). However, G2P mapping, which is working between grapheme-to-phoneme converted strings, shows the best result for all string similarity metrics. According to these experimental results, our new two mappings (phonetic and sound mappings) are applicable for string similarity measurement on spelling mistake confusion words. Moreover, based on the current results for thresholds “<=2” and “<=3” (or “>=0.7” and “>=0.5”), we clearly found that the G2P mapping is able to retrieve approximately 70% of the correct words for Levenshtein, Jaro-Winkler and Cosine similarities.

The results of retrieving similar pronunciation words, such as homophones and rhyme words, with six similarity measures on the “Similar Pronunciation Dataset” is shown in Figure 3. As we expected, two of our proposed mappings, phonetic mapping and sound mapping, achieved the highest number of found errors for all thresholds of Levenshtein, Damerau-Levenshtein, Hamming, Jaro-Winkler and Jaccard similarities except for Cosine similarity. Additionally, the G2P mapping also obtained just about 50 % of the correct words for all measures with threshold “<=3” and “>=0.5” in general.

We did a detailed analysis on distance values, and we found that our proposed three mappings have a zero distance value (i.e., no distance value) for some similarly pronounced words. For example, the string similarity distances for the word “လက်ရွေးစင်” and similar pronunciation and rhyme words “လက်ရွေးစဉ်”, “လက်ယွေးစင်”, “ရက်ရွေးစင်” and “လက်ရွေးဇင်” for Levenshtein and our three mappings for the threshold “<=1” are shown in Table V. Moreover, three mappings retrieved similar words well, compared with inputting raw Myanmar text. For example, although Levenshtein distance (for the threshold “<=1”) retrieved



(a) Levenshtein (b) Damerau Levenshtein

(c) Hamming (d) JaroWinkler

(e) Cosine (f) Jaccard

Fig. 2: Results with the spelling-mistake confusion dataset

only one similar word of "လွင့်စဉ်" ("scatter" in English), our three mappings were able to retrieve three more similar words "လွှင့်စင်", "လွင့်စင်" and "လွန်စင်" (see Table VI). One more example of cosine and all three mappings' string similarity distances of the word "အကဲခတ်" ("to assess" in English) (for threshold ">=0.9") can be seen in Table VII. Here, "N\A" means "Not Applicable", and the expression is not contained in the threshold distance.

Word - Similar Word	Levenshtein	Pronunciation	Sound	Vowel
လက်ရွေးစင် လက်ရွေးစဉ်	1	0	1	0
လက်ရွေးစင် လက်ယွေးစင်	1	0	0	0
လက်ရွေးစင် ရက်ရွေးစင်	1	1	1	0
လက်ရွေးစင် လက်ရွေးဇင်	1	1	0	0

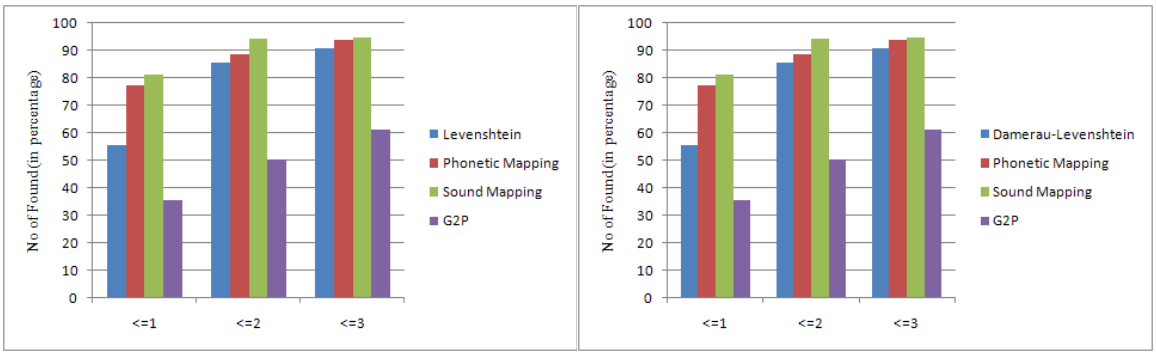
TABLE V: String similarity distances for the word "လက်ရွေးစင်" ("selection") in English

Word - Similar Word	Levenshtein	Pronunciation	Sound	Vowel
လွင့်စဉ် လွှင့်စင်	1	0	1	0
လွင့်စဉ် လွင့်စင်	N/A	0	1	1
လွင့်စဉ် လွင့်ဇင်	N/A	1	1	0
လွင့်စဉ် လွန်စင်	N/A	1	1	0

TABLE VI: String similarity distances for the word "လွင့်စဉ်" ("scatter" in English)

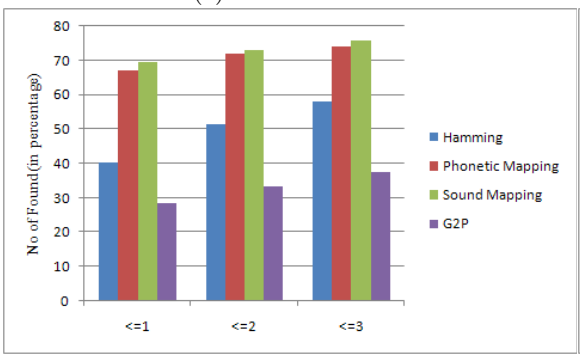
VII. Conclusion

In this paper, we have presented the first study of the string similarity measurement based on the pronunciation similarities for Burmese. We proposed two new mappings (phonetic mapping, sound mapping) and G2P mapping using RDR model for conversion grapheme to phoneme. We also proved a better retrieving of similarly pronounced

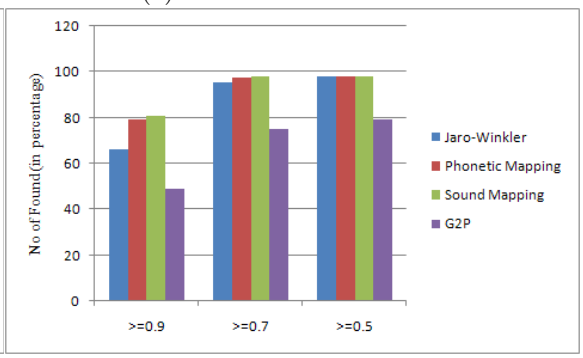


(a) Levenshtein

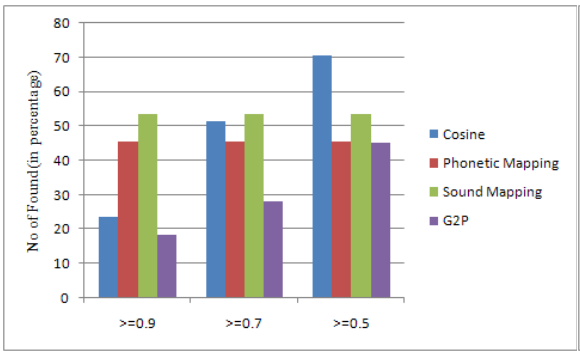
(b) Damerau Levenshtein



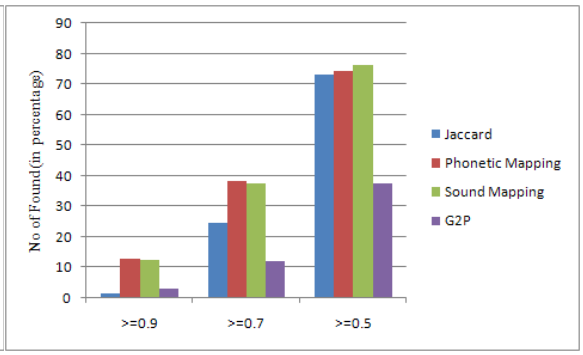
(c) Hamming



(d) JaroWinkler



(e) Cosine



(f) Jaccard

Fig. 3: Results with the similar pronunciation dataset

Word - Similar Word	Cosine	Pronunciation	Sound	Vowel
အကဲခတ် အကဲခပ်	N/A	1.0	1.0	1.0
အကဲခတ် အကဲဆတ်	N/A	N/A	N/A	1.0
အကဲခတ် အမြဲတက်	N/A	N/A	N/A	N/A
အကဲခတ် မဆဲတတ်	N/A	N/A	N/A	1.0

TABLE VII: String similarity distances for the word "အကဲခတ်" ("to assess" in English)

words, homophones, and rhyme words. Moreover, the G2P mapping is applicable for spelling correction by string similarity measurement of Burmese under the threshold "<=1". In the future work, we plan to expand the two datasets and conduct string similarity experiments to confirm our current mapping tables. Moreover, different G2P conversion models can be used to get better comparison between the words or strings.

References

[1] Ohnmar Htun, Shigeki Kodama, Yoshiki Mikami, "Measuring Phonetic Similarities in Myanmar IDNs", 2010.

[2] Ohnmar Htun, Shigeaki Kodama, Yoshiki Mikami, Cross-language Phonetic Similarity Measure on Terms Appeared in Asian Languages, International Journal of Intelligent Information Processing Volume 2, Number 2, June 2011

[3] Anil Kumar Singh, "A Computational Phonetic Model for Indian Language Scripts", Proceedings of Constraints on Spelling Changes: Fifth International Workshop on Writing Systems, 2006

[4] Burmese Language Wikipedia Page: https: //en. wikipedia.org /wiki/Burmese_language

[5] Ye Kyaw Thu, et al. Syllable Pronunciation Features for Myanmar Grapheme to Phoneme Conversion.

[6] Ye Kyaw Thu, et al. "The Application of Phrase Based Statistical Machine Translation Techniques to Myanmar Grapheme to Phoneme Conversion". Communications in Computer and Information Science, vol. 593, 2016, pp. 238–50, doi:10.1007/978-981-10-0515-2_17.
[7] Ye Kyaw, et al. "Comparison of Grapheme-to-Phoneme Conversion Methods on a Myanmar Pronunciation Dictionary". Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing, pages 11-22, Osaka, Japan, December 11-17 2016.
[8] Levenshtein, V. I., "Binary Codes Capable of Correcting Deletions, Insertions and Reversals", Soviet Physics Doklady, Vol. 10, p.707, 02/1966
[9] Damerau, Fred J., "A technique for computer detection and correction of spelling errors", Communications of the ACM, 7 (3): 171–176, March, 1964
[10] Hamming, R. W, "Error detecting and error correcting codes". The Bell System Technical Journal. 29 (2): 147–160, April 1950
[11] Matthew A. Jaro, Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, Journal of the American Statistical Association, 84(406):414-420, June 1989.
[12] Singhal, Amit, "Modern Information Retrieval: A Brief Overview", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43., 2001
[13] Jaccard, P., "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines", Bulletin de la Société Vaudoise des Sciences Naturelles 37, 241-272, 1901
[14] Odell, Margaret King , "The profit in records management Systems", New York, 20: 20, 1956
[15] Compton, P., & R. Jansen 1990. A philosophical basis for knowledge acquisition. Knowledge Acquisition, 2(3):241 – 258.
[16] Nguyen, Dat Quoc, Dai Quoc Nguyen, Dang Duc Pham, & Son Bao Pham 2016. A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. AI Communications, 29(3):409–422.
[17] Richards, Debbie 2009. Two decades of Ripple Down Rules research. Knowledge Eng. Review, 24(2):159–184.
[18] Scheffer, Tobias 1996. Algebraic Foundation and Improved Methods of Induction of Ripple Down Rules. Pages 23–25.
[19] Ye Kyaw Thu and Yoshiyori Urano, "Positional Mapping: Keyboard Mapping Based on Characters Writing Positions for Mobile Devices", Proceedings of the 9th International Conference on Multimodal Interfaces, ICMI 07, 110-117, 2007
[20] Thein Tun, "Acoustic phonetics and the phonology of the myanmar language", School of Human Communication Sciences, La Trobe University, Melbourne, Australia, 2007.
[21] Thein Tun, "The domain of tones in burmese", SST 1990 Proceedings, pp. 406–411, 1990.

Khaing Hsu Wai is currently pursuing Master of Engineering (Information Science and Technology) degree program at the University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar. Her current research interests are in the areas of Natural Language Processing, Machine Learning and Artificial Intelligence.

Ye Kyaw Thu is a Visiting Professor of Language & Semantic Technology Research Team (LST), Artificial Intelligence Research Unit (AINRU), National Electronic & Computer Technology Center (NECTEC), Thailand and Head of NLP Research Lab., University of Technology Yatanarpon Cyber City (UTYCC), Pyin Oo Lwin, Myanmar. He is also a founder of Language Understanding Lab., Myanmar and a Visiting Researcher of Language and Speech Science Research Lab., Waseda University, Japan. He is actively co-supervising/supervising undergrad, masters' and doctoral students of several universities including MTU, UCSM, UCSY, UTYCC and YTU.

Swe Zin Moe is currently an Assistant Lecturer at Myanmar Institute of Information Technology (MIIT) and also a Ph.D candidate at University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar. Her current doctoral thesis research focuses on machine translation between Myanmar sign language and Myanmar written text. She is interested in the general and related problems of natural language processing (NLP) such as machine translation, big data analysis and deep learning.

Hnin Aye Thant is Currently working as a Professor and Head of Department of Information Science at the University of Technology (Yatanarpon Cyber City), UTYCC, Pyinoolwin Township, Mandalay Division, Myanmar. I got Ph.D(IT) Degree from University of Computer Studies, Yangon, Myanmar in 2005. The current responsibilities are managing professional teachers, doing instructional designer of e-learning content development and teaching. I have 14 years teaching experiences in Information Technology specialized in Programming Languages (C, C++, Java & Assembly), Data Structure, Design and Analysis of Algorithms/Parallel Algorithms, Database Management System, Web Application Development, Operating System, Data Mining and Natural Language Processing. I am a member of research group in "Neural Network Machine Translation between Myanmar Sign Language to Myanmar Written Text" and Myanmar NLP Lab in UTYCC. I am also a Master Instructor and Coaching Expert of USAID COMET Mekong Learning Center. So, I have trained 190 Instructors from ten Technological Universities, twelve Computer Universities & UTYCC for Professional Development course to transform teacher-centered approach to learner-centered approach. This model is to reduce the skills gap between Universities and Industries and to fulfil the students' work-readiness skills.

Thepchai Supnithi received the B.S. degree in Mathematics from Chulalongkorn University in 1992. He received the M.S. and Ph.D. degrees in Engineering from the Osaka University in 1997 and 2001, respectively. Since 2001, he has been with the Human Language Technology Laboratory, NECTEC, Thailand.