University of Technology (Yadanarpon Cyber City)

Searching Mechanism Of Myanmar Words To Myanmar Sign Lauguage Videos

Ma Khaing Hsu Wai

2ME-IST-6

(Defense)

Supervised By

Dr. Hnin Aye Thant

Co-supervised By

Dr. Ye Kyaw Thu

Daw Swe Zin Moe

Date : 3.12.2019

Outline

- Abstract
- Objectives
- Introduction
- Theoretical Background
- Proposed Mapping
- Datasets and Evaluation
- Results and Discussion
- Implementation
- Conclusion

Abstract

- Natural language processing (NLP) is a brunch of artificial intelligence that deals with the interaction between computers and humans using the natural language.

- Information Retrieval (IR) is detecting information by locating documents with the terms specified in their queries.

- String similarity is the process of taking two or more strings and comparing them with each other to find out how similar they are.

Abstract (Cont'd)

- In this study, string similarity metrics have been calculated for Myanmar language (Burmese).

- The encoding table for Myanmar language has built based on the pronunciation similarity of characters and vowel combination positions with a consonant.

- The purpose of this proposed mapping is to use in retrieving or searching phonetically similar words from the Myanmar Sign Language Dictionary.

Objectives

- To study searching mechanism for Myanmar characters and videos
- To build video-based Myanmar Sign Language - Myanmar words dictionary with suitable word searching mechanism
- To search phonetically similar words according to the mapping and string similarity metrics
- To develop the knowledge of Myanmar Language for the children and people who are unable to hear
- To support better communication between normal people and the hearing disabilities

Introduction

- For Myanmar words to MSL search, the system searches the similar words compared to the user input based on the string similarity distance.
- String similarity is a measure to define the similarity between given two strings.
- In literature, a variety of approaches are proposed for string similarity.
- Most of them are character-based metrics and associated with English languages.

Introduction (Cont'd)

- In Myanmar Language, syllables or words are formed by combination of consonants and vowels.

- There are also many phonetically similar sounds of characters in Myanmar Language. (e.g. - uav; eSifh cav;)

- So, new approaches together with the existing string similarity metrics are needed to consider for Myanmar language.

Theoretical Background

- String similarity determines how similar the two strings are.
- Edit distance based metrics try to compute the number of operations needed to transforms one string to another.
- Token based metrics is a set of tokens and the purpose is to find the similar tokens in both sets.
- In sequence based group, the similarity is a factor of common sub-strings between the two strings.

Theoretical Background (Cont'd)

- The more the number of operations, the less is the similarity between the two strings and the more the number of similar tokens or sequences found, the higher is the similarity score.

- Six similarity measures are used to evaluate in this research:
  - Levenshtein Distance
  - Damerau-Levenshtein Distance
  - Hamming Distance
  - Jaccard Similarity
  - Cosine Similarity
  - Jaro Winkler Similarity

Levenshtein Distance

- Given two character strings s1 and s2, the edit distance between them is the minimum number of edit operations required to transform s1 into s2.
- The edit operations are insertion, deletion and substitution.
  - 
  - 

Levenshtein Distance = 2

Damerau Levenshtein Distance

- The Damerau-Levenshtein distance differs from the Levenshtein distance by including transpositions among its allowable operations in addition to three classical single character edit operations.
  - 
  - 

Damerau Levenshtein Distance = 1

Hamming Distance

- The Hamming distance between two strings of equal length measures the number of positions with mismatching characters.
  - ಅಆಇ
  - ಅಆಈ

Hamming Distance = 1

Jaccard Similarity

- Jaccard similarity measures similarities between sets.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- A=[ಣ,ತ,ನ, ಪ] and B =[ಣ, ತ, ನ, ಪ, c, ಱ]
- J(A,B) = |A ∩ B| / |A| + |B| - |A ∩ B|
  = 4 / 4 + 6 - 4 = 4 / 6 = 0.67

Jaccard Similarity = 0.67

Cosine Similarity

- Cosine similarity between two vectors is a measure that calculates the cosine of the angle between them.
- The two vectors with the same orientation have a cosine similarity of 1.

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

- where the denominator is the product of the vectors and the numerator shows the dot product which is the inner product of the vectors.

```
Cosine Similarity
-----------------
ဇနီးမောင်နှံ and ကလေး: : 0.0
```

```
Cosine Similarity
-----------------
စနီးမောင်နှံ and ဇနီးမောင်နှံ: 0.75
```

```
Cosine Similarity
-----------------
ဇနီးမောင်နှံ and ဇနီးမောင်နှံ: 1.0
```

Jaro Winkler Similarity

- Jaro Winkler distance is a string metric measuring an edit distance between two sequences.
- The score ranges from 0 to 1 where 0 is no similarity and 1 is exact the same strings.

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

- where $|s_i|$ is the length of the string s_i
- m is the number of matching characters
- t is half the number of transpositions

```
Jaro Winkler
--------------------
ခေါင်းစဉ် and ခေါင်းစဉ် : 1.0
```

```
Jaro Winkler
--------------------
ခေါင်းစဉ် and ခေါင်းစဉ် : 0.9555555555555556
```

Soundex

- Soundex algorithm is a phonetic algorithm.

- It is based on how close two words are depending on pronunciation.

  - Flower = F406
  - Flour = F406

Letters	Code
A,E,H,I,O,U,W,Y	0
B,F,P,V	1
C,G,J,K,Q,S,X,Z	2
D,T	3
L	4
M,N	5
R	6

Soundex (Cont'd)

- Like Soundex algorithm, the proposed approach aims to search words based on phonetic similarity.
- Based on the idea of Soundex algorithm, there are three proposed mappings for Myanmar Lauguage.
- All of the mappings are aimed to used to compare the two data strings that may be spelled differently but sound the same.

Proposed Mapping for Phonetically Myanmar Similar Words

- Phonetic Mapping
- Sound Mapping
- Vowel Position Mapping

Phonetic Mapping

Letters	Code	Letters	Code
က ခ	က	◌ွ ◌့	(delete)
ဂ ဃ	ဂ	ကြက်၏ဝိဓိဉ်	i
စ ဆ	စ	က်ပ်တ်	d
ဇ ဈ	ဇ	နှ်မ်ံ	n
ဋ တ	တ	ဲရ်	e
ဌ ထ	ထ	ဥဦ ◌ု ◌ူ	u
ဍ ဎ	ဍ	တ ါ	r
ဏ နှ	နှ	ဧ ေ	a
ဒ ဓ	ဒ	◌ိ ◌ီး	(delete)
ပ ဖ	ပ	◌်	(delete)
ဗ ဘ	ဘ	ဩ ြော် ဩ ြော်	o
ယ ရ	ရ	၎င်း ၎	၎
လ ဠ	လ	၊ ။	s
သ ဿ	သ	်ဳင်ဂဥ်	in
ျ ြ	y	?!.*-=\#"<>{}[],+-	s

Table : Phonetic Mapping

- Same pronunciation words are grouped together.
- For example, ကလေး and ခလေး have same pronunciation.
- Some diacritics such as ◌ွ (WaHswe) and ◌့ (Ha Hto), tone marks such as ◌ိ (Aukmyit), ◌် (Myanmar sign Virama) are considered to be removed.
- Source String = ပစ်စီး
- Target String = ပစ္စည်း
- Encoded Source String using Phonetic Mapping = ပစစi
- Encoded Target String using Phonetic Mapping = ပစစi

Example for Phonetic Mapping

Source String Target String	ပစ်စီး ပစ္စည်း	ပစၥi ပစၥi
Levenshtein Distance	3	0
Damerau-Levenshtein Distance	3	0
Hamming Distance	4	0
Jaro Winkler Distance	0.7968	1
Cosine Similarity	0.4999	1
Jaccard Similarity	0.4444	0.6

Table : Calculation of string similarity between two strings with Phonetic Mapping

Sound Mapping

Letters	Code	Letters	Code
ကခဂဃငဟအ	က	◌ွ ◌့	(delete)
ညဉ	ည	ကြဉ်၏ ◌ီ ◌ိ ည်	i
စဆဇဈ	စ	က်ပ်တ်	d
ဋဌဍဎဏထဒဓန	တ	န်မ် ◌ံ	n
ပဖဗဘမ	ပ	◌ဲ ရ်	e
ယရ	ရ	ဥဦ ◌ု ◌ူ	u
လဠ	လ	◌ာ ◌ါ	r
သဿ	သ	ဧ ◌ေ	a
ျ ြ	y	◌ှ ◌း	(delete)
၊ ။	s	◌္	(delete)
၎င်း ၎	၎	ဩ ◌ော် ဩ ◌ော်	o
◌် ◌င ◌ဉ်	in	?!.*-=\#"<>{}[],+-	s

Table : Sound Mapping

- Sound Mapping is similar to Phonetic Mapping.
- The main different part of Phonetic mapping is at the Myanmar consonant.
- As the name of sound mapping, consonants, which have the same movements from mouth, lips and tongue, are grouped.
- For example, က ခ ဂ ဃ င ဟ အ (Ka Kha Ga Gha Nga Ha A) are clustered of က (Ka) group. ပ ဖ ဘ ဗ မ (Pa Pha Ba Bha Ma) are clustered of ပ (Pa) group etc.

Example for Sound Mapping

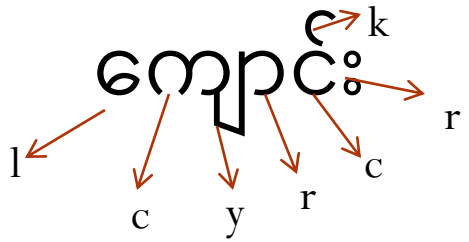
Source String Target String	ဖုရား ဘုရား	ပရာ ပရာ
Levenshtein Distance	1	0
Damerau-Levenshtein Distance	1	0
Hamming Distance	1	0
Jaro Winkler Distance	0.8667	1
Cosine Similarity	0.4999	1
Jaccard Similarity	0.6667	1

Table : Calculation of string similarity between two strings with Sound Mapping

Vowel Position Mapping

Letters	Code	Letters	Code
a-z A-Z	F	က-အ	c
ျ ြ	y	့	p
ေ	l	ာ ါ ိ ီ ု ူ း	r
ိ ီ ံ ံ	u	ု ူ ု ူ	d
်	k	။	s
ဣ ဤ ဥ ဦ ဧ ဩ ဪ သ ၌ ၍ ၏	i	?!.*-=\#"<>{}[],+-	$
၀-၉	n	0-9	D

Table : Vowel Position Mapping



- The third mapping based on the syllable formation of Myanmar language.
- The vowels written on the left side of the consonant are under the left (l) group, the right side vowels are under the (r) group, the upper vowels are under the (u) group, the lower vowels are under the (d) group respectively.

Example of Vowel Position Mapping

Source String Target String	ကျောင်း ချောင်း	cylrckr cylrckr
Levenshtein Distance	1	0
Damerau-Levenshtein Distance	1	0
Hamming Distance	1	0
Jaro Winkler Distance	0.9047	1
Cosine Similarity	0.4999	1
Jaccard Similarity	0.75	1

Table : Calculation of string similarity between two strings with Vowel Position Mapping

Data Sets and Evaluation

- Spelling Mistake Confusion Pair
  - Develop based on the real world spelling errors
  - Collect from Myanmar web news and social media website (BBC, VOA, Facebook)
  - 2,381 pairs in total (i.e. 4,762 words)

- Word Similarity Data Set
  - Develop similar pronunciation data set to measure how well the similarity scores provided by three mappings
  - Manually added the correct words together with homophone and rhyme words
  - 200 pairs in total (i.e. 1,000 words)

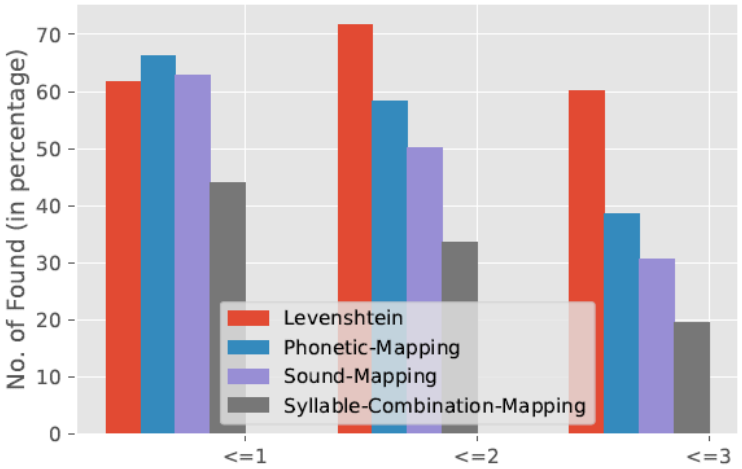
Data Sets and Evaluation (Cont'd)

- For the evaluation, string similarity on each pair of data set is measured.

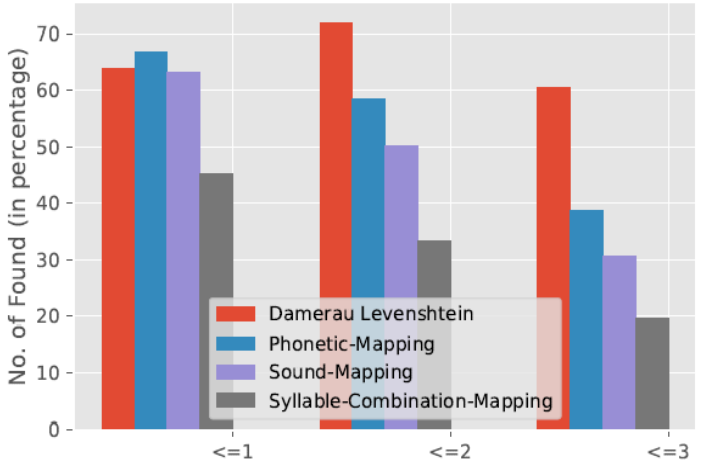
- After that, the original data is encoded or converted with three proposed mappings and string similarity is measured again.

- Finally, count 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 "Jaro Winkler", "Cosine" and "Jaccard" distance measures.

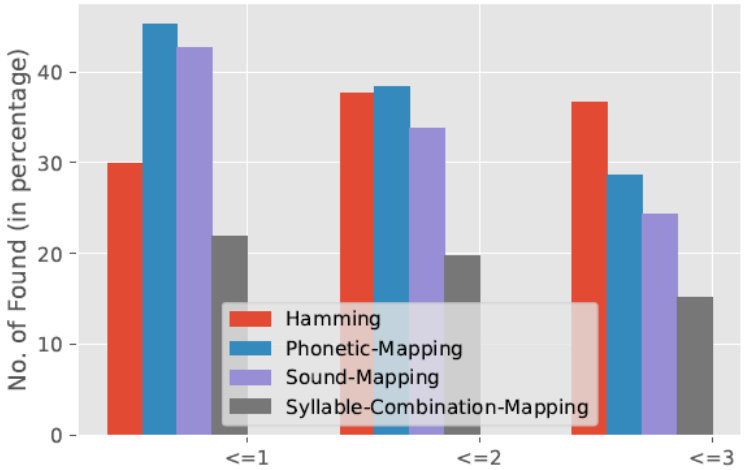
Result and Discussion



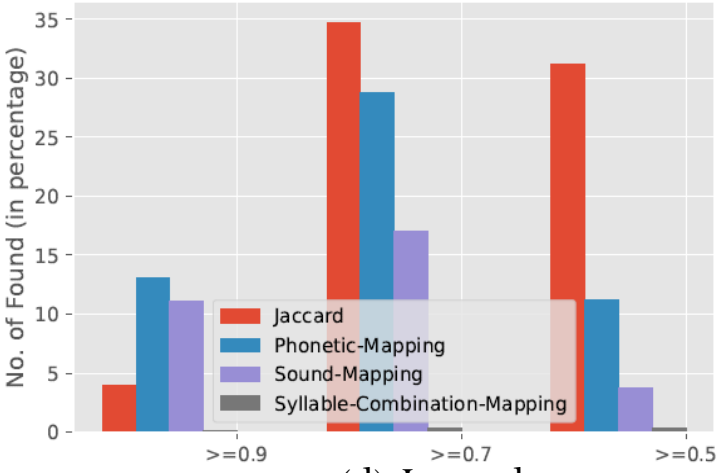
(a) Levenshtein Distance



(b) Damerau-Levenshtein Distance

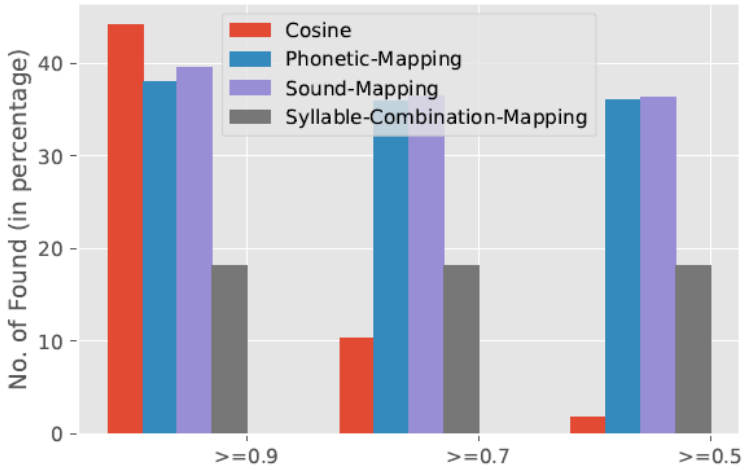


(c) Hamming Distance

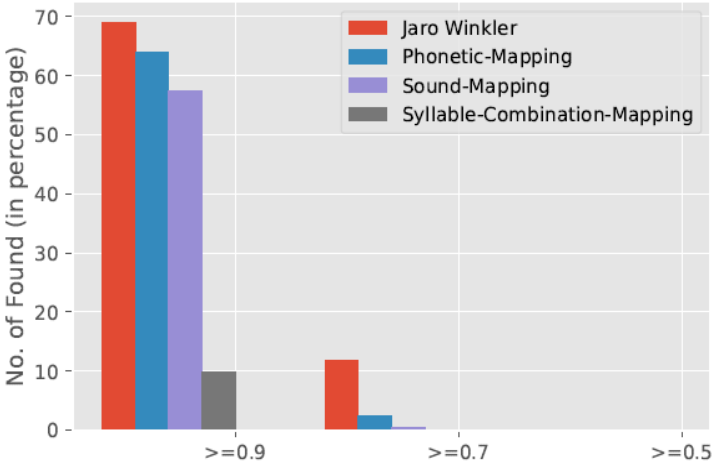


(d) Jaccard

Result and Discussion (Cont'd)



(e) Cosine Similarity



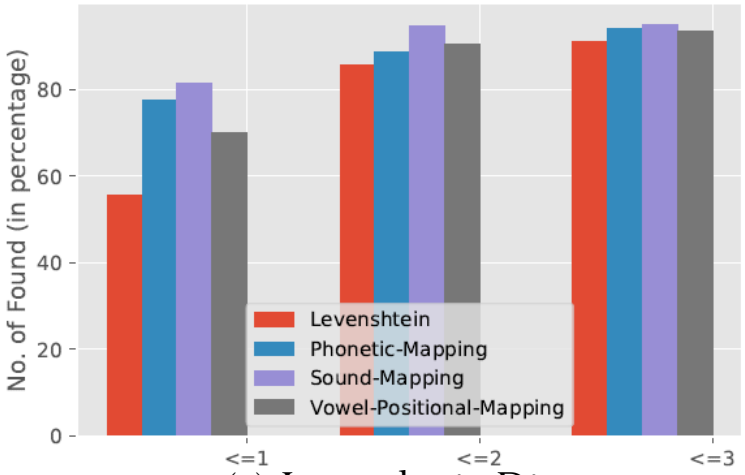
(f) Jaro Winkler

Figure : Results with the spelling-mistake confusion dataset

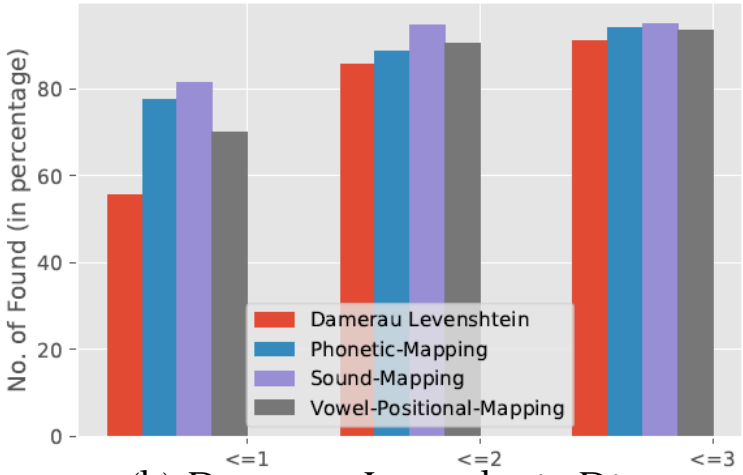
- According to the results, Phonetic and Sound mappings are applicable for string similarity measurement on spelling mistake confusion words.

- It is assumed to be very useful for evaluating on three mappings but this data has few homophone and rhyme words, so it is not suitable for measuring pronunciation similarity.

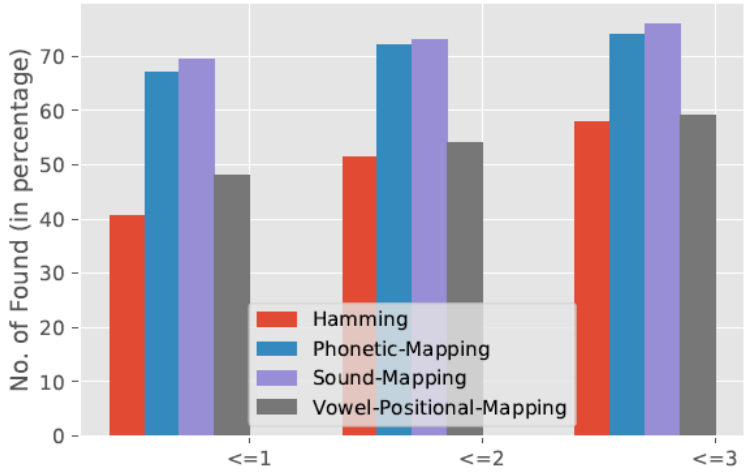
Result and Discussion (Cont'd)



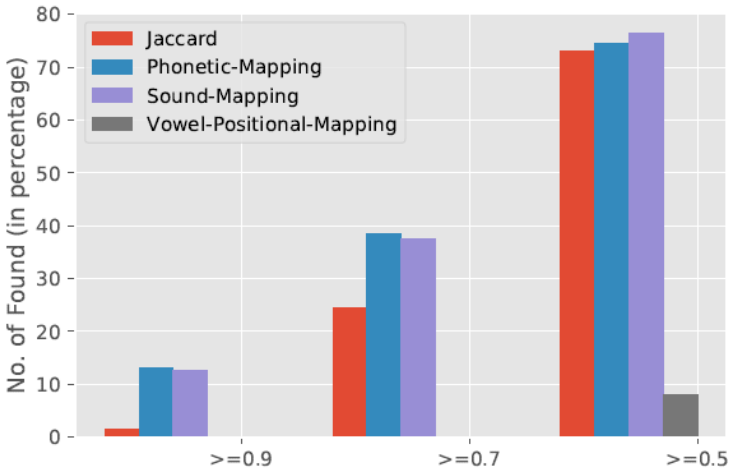
(a) Levenshtein Distance



(b) Damerau-Levenshtein Distance

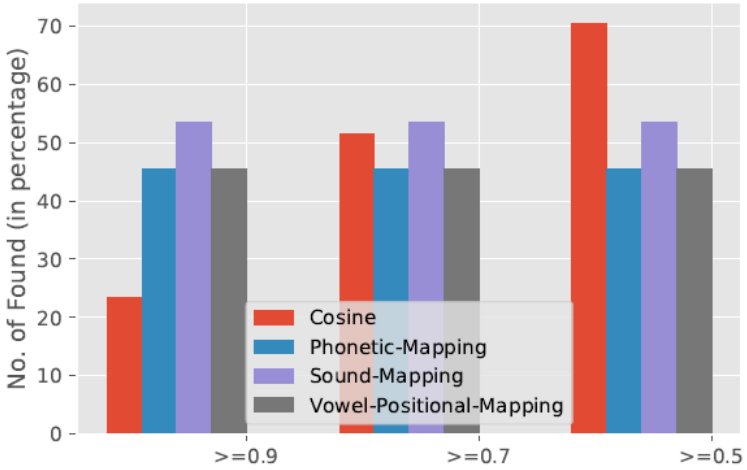


(c) Hamming Distance

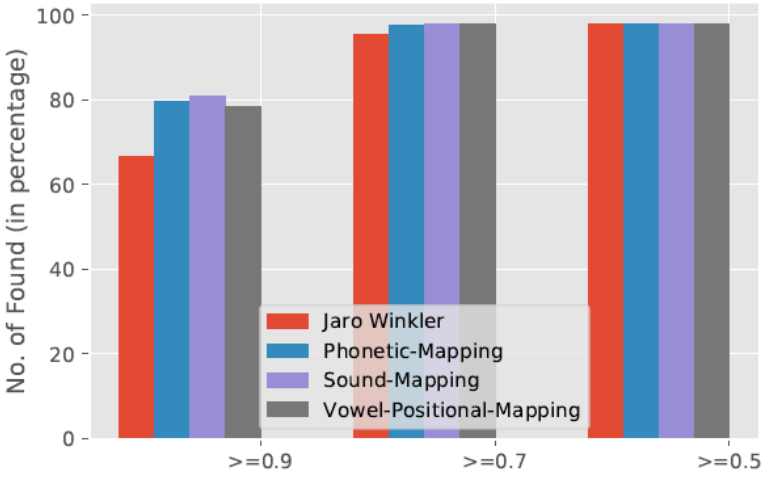


(d) Jaccard

Result and Discussion (Cont'd)



(e) Cosine



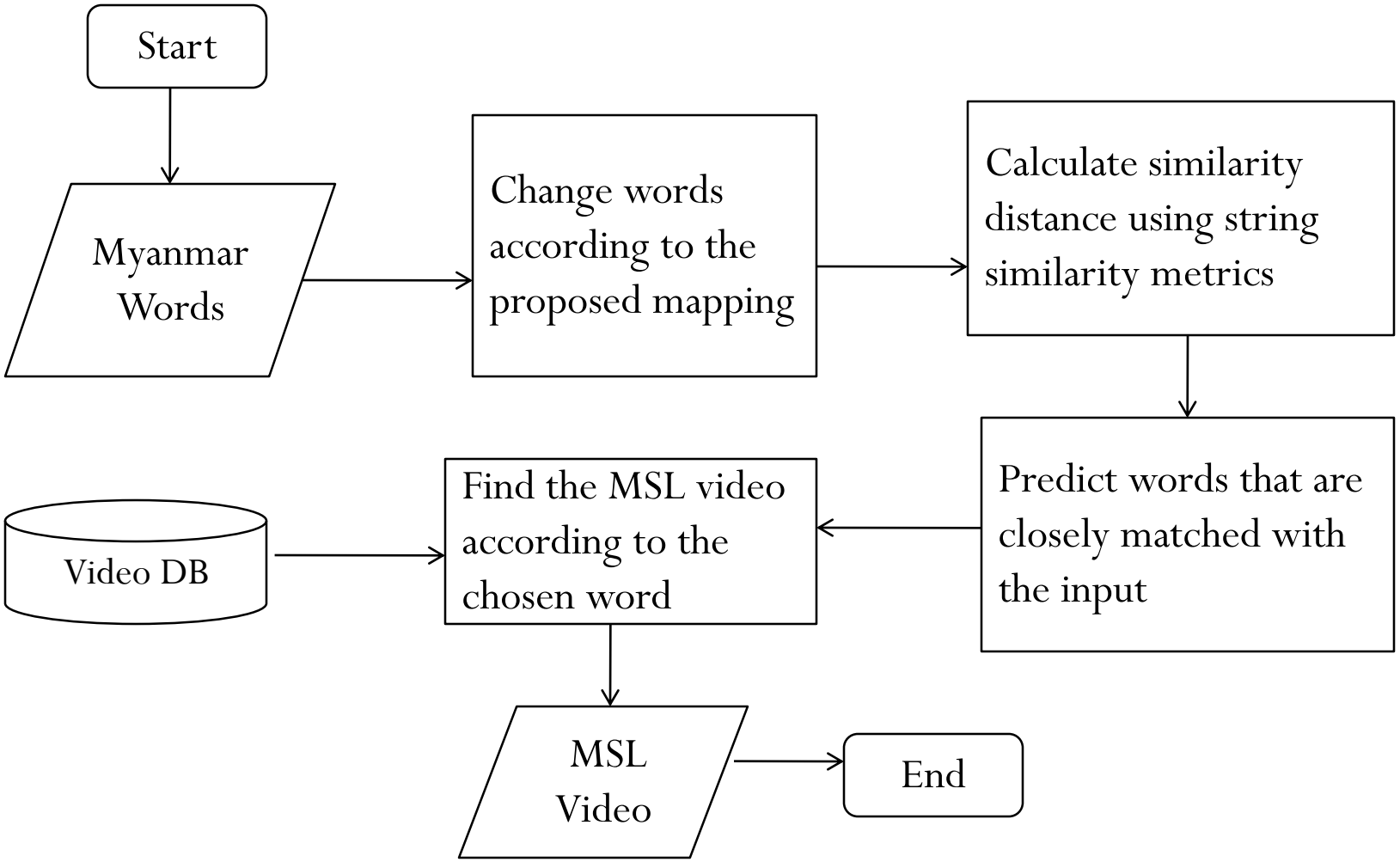
(f) Jaro Winkler

Figure : Results of the Similar Pronunciation Dataset

- As expected, our proposed two mappings: Phonetic Mapping and Sound Mapping achieved highest number of founds for all thresholds of Levenshtein, Damerau Levenshtein, Hamming, Jaro Winkler, Cosine and Jaccard distance measures.

- Additionally, the Vowel Positional Mapping also gives the highest results for existing five distance measures except for the Jaccard distance measure.

System Design



Myanmar To Myanmar Sign Language System Overview

Words and Video Data Preparation

- Collect video data from Myanmar words based on the Myanmar Sign Language Dictionary distributed from School for the Deaf (Mandalay) and Myanmar Dictionary from Myanmar Language Commission

- Edit raw video data with Cyber link Power Director, remove noise and sound with regular expressions command and match with their respective words as a database

System Implementation

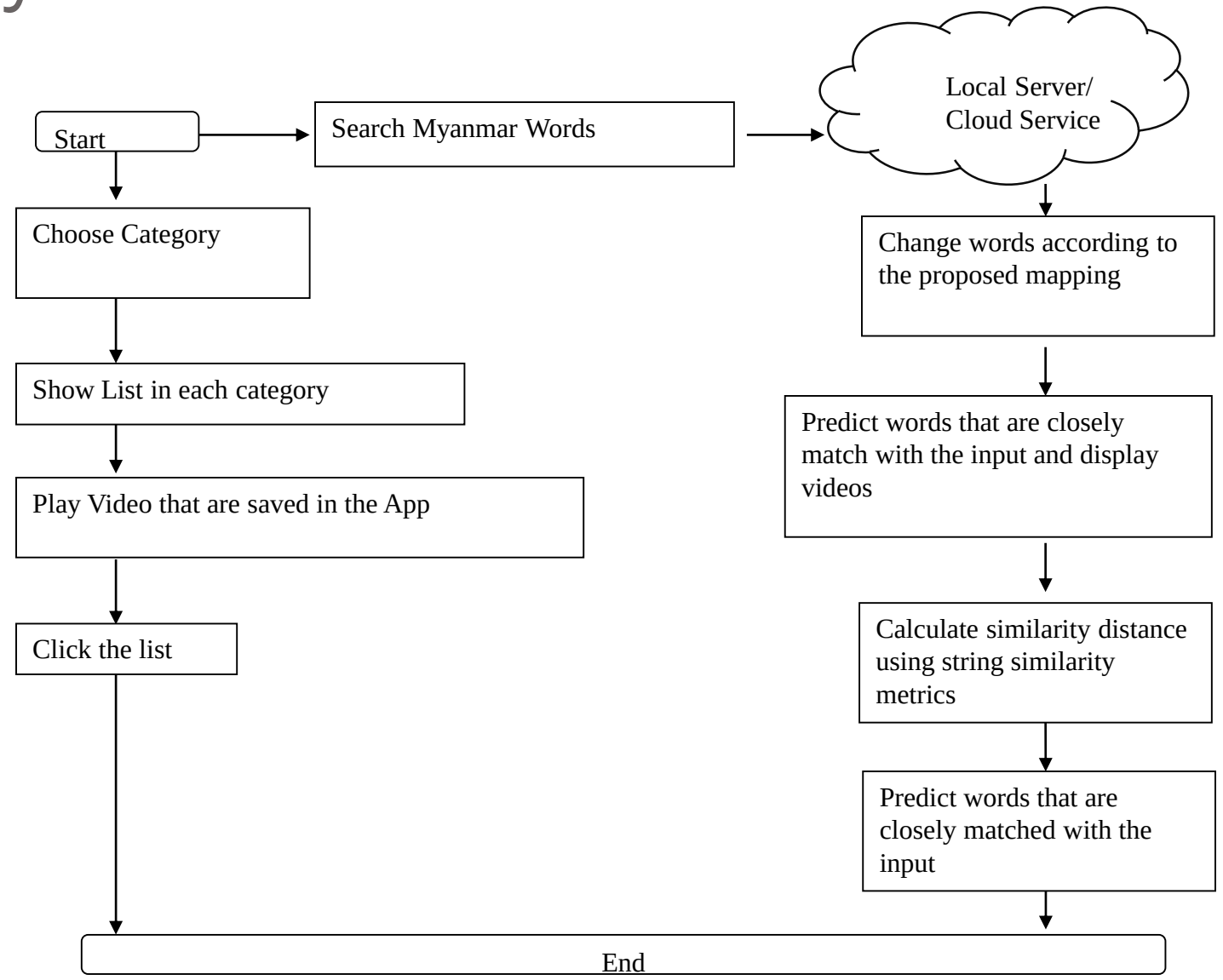
- Phonetic Mapping and Levenshtein distance is used to measure the string similarity distance between the words.

- The overall accuracy is calculated based on the 141 input words.

- For threshold value <=1, the overall accuracy is 8% which is very low. However, for threshold value <=2 and <=3, the overall accuracy is 33% and 28% respectively.

Average overall f-measure with threshold value <=1	0.0802
Average overall f-measure with threshold value <=2	0.3296
Average overall f-measure with threshold value <=3	0.2805

System Flow



MSL Application System Flow

Conclusion

- In this proposed system, simple text searching method and video similarity approach for classification of MSL videos will be used to make a useful Myanmar to MSL searching.

- The major goal is to have a better communication between the deaf and the hearing people.

- With this system, it is easy to search the words between Myanmar Language and MSL.

Limitations and Further Extension

- As this system is about searching mechanism between Myanmar words and Myanmar Sign Language videos, the users can search the words and their related sign language videos.

- However, the words in the dataset is limited to about 3000 words according to the time limitations and most of them are used in the deaf schools.

- The sign language videos are collected with the help of Myanmar Deaf society.

Limitations and Further Extension (Cont'd)

- The words can be added and the more the sign language videos data, the better is the system searching mechanism.

- The searching and retrieving will be extended and implemented by using other string similarity metrics.

List of Publications

- Khaing Hsu Wai, Ye Kyaw Thu, Hnin Aye Thant, "Learning String Similarity Metrics for Myanmar Language", at 9th Workshop on Natural Language Processing, ICCA2019, Yangon, Myanmar (28th Feb 2019)

- Khaing Hsu Wai, “Searching Mechanism for Myanmar Sign Language Video”, Machine Learning Research School (MLRS2019), 7 August 2019, Bangkok, Thailand (Poster Presentation)

- Khaing Hsu Wai, Ye Kyaw Thu, Hnin Aye Thant, Swe Zin Moe, Thepchai Supnithi, "String Similarity Measures for Myanmar Language (Burmese)", The First Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019), 11-13 September 2019, Trento, Italy

- Khaing Hsu Wai, Ye Kyaw Thu, Hnin Aye Thant, Swe Zin Moe, Thepchai Supnithi, “Myanmar (Burmese) String Similarity Measures based on Phoneme Similarity”, iSAI-NLP2019, 30 October 2019, Chiang Mai, Thailand

Thank You