Development of Cross-Lingual Phonetic Similarity Metrics
言語横断型音声類似度指標の開発

by

OHNMAR HTUN

Supervisor: Professor Yoshiki Mikami

Dissertation

Presented to the Faculty of the Graduate School of Engineering

Nagaoka University of Technology

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Engineering

(Information Science and Control Engineering)

Nagaoka University of Technology

September 2013, Japan

Dedication

I wish to dedicate this dissertation to my supervisor, Prof. Yoshiki Mikami for his great effort to me from the beginning till the end of my doctoral studies, to my mother and brothers, sister for their love and affection.

Acknowledgements

With a deep sense of gratitude, I wish to express my sincere thanks to my supervisor, Professor Yoshiki Mikami for his continuous guidance and great support throughout my doctoral program. With him, I have learned about the process of research, how to study scientific logic, how to write papers, and how to make presentations. Besides being an outstanding advisor he taught me how to be a critical thinker and prominent researcher. I am thankful to co-supervisor Andrew Finch (NICT) and co-author Shigeaki Kodama for their valuable guidance and help to my research. I am also indebted to Dr. Eiichiro Sumita (NICT), Prof. Jay Rajasekera (IUJ), Prof. Yoshinori Sagisaga (Waseda University), Assoc.Prof. Ashu Marasinghe (NUT), and Assoc.Prof. Kanou Mitsuru (NUT) for all the help and suggestions.

I am extremely grateful to all the colleagues and friends at NUT who helped and encouragement me during my struggle time. I also wish to express my gratitude to my host parents: Katsumata Shunsuke and Katsumata Setsuko, to my friends: Maung Maung Thant, Swe Swe Thant, and Ye Kyaw Thu (NICT) for their timely support and encouragements.

I am very obliged to Japan Student Services Organization (JASSO) for awarding me with the scholarship to study at NUT. Similarly I extend my appreciation to the division of international affairs (Kokusai-ka) for their administrative support.

Abstract

This dissertation presents a cross-lingual phonetic similarity metric (CLPSM) that can effectively measure phonetic similarity of words which belongs to different languages. CLPSM is a methodology that integrates the Soundex algorithm, Levenshtein Edit Distance, Stochastic Edit Distance, and Bayesian Alignment Model. The dissertation consists of seven chapters.

Chapter 1 describes the background and motivation of this research. Afterwards, related works are reviewed. In the multilingual webspace, it often happens that more rich information is available in other languages than in the original language. In that case, cross-language information retrieval (CLIR) applications and machine translation (MT) systems are helpful to provide access to that information for users. And CLPSM can support CLIR by providing phonetic borrowing word pairs and can support MT systems by providing transliteration mining tools.

Chapter 2 presents the concepts of linguistic borrowing, the background theories, and the definition of terms employed in this research (e.g., phonetic similarity, linguistic borrowing, phonetic transcription, Soundex phonetic coding system, and edit distance similarity measures).

Chapter 3 describes the development of cross-language sound grouping (CLSG) and CLPSM in two approaches. Firstly, development of CLSG for Asian languages is reported in detail. Secondly, development of CLPSM based on classical edit distance (Levenshtein Edit distance) in a manual designing approach is reported. Thirdly, development of CLPSM based on different stochastic models, namely, stochastic edit distance (EM), stochastic edit distance with noise (EMn), and Bayesian alignment with noise (BAYESn) in a learning approach are reported.

Chapter 4 presents the experimental results and evaluation in a manual designing approach by using Levenshtein Distance (LD) and Normalization of Levenshtein Distance (NS). The experiment uses the names of 92 chemical elements words in eight Asian language pairs: English-Japanese, English-Korean, English-Malay, English-Myanmar, English-Thai, English-Indonesian, English-Vietnamese, and English-Chinese. The results of two CLSG versions are compared in two metrics (LD and NS).

Chapter 5 presents the experiment results and evaluation in a learning approach by using Stochastic Edit Distance (EM), Stochastic Edit Distance with noise model (EMn) and Bayesian Alignment with noise model (BAYESn). The data for this experiment was prepared by procedures consists of segmentation and romanization of Myanmar language data, and building Myanmar-English bilingual training corpus consisting of 3,100 single word pairs and 14,891 multiple word pairs. The results of the three different models are also compared.

Chapter 6 discusses the analysis of methodologies and experimental results to evaluate the performance of CLSG and compares between manual designing and learning approaches. The pros and cons of methodologies in two approaches are presented.

Finally, chapter 7 summarizes the contributions of this work first, then gives brief discussions on limitations of proposed techniques, and future tasks are presented.

Table of Contents

List of Tables

List of Figures

Chapter: 1

Introduction

1.1 BACKGROUND AND MOTIVATION

Languages on WWW

There are around six thousand living languages used by the people of the world. Today, the information provided on the Internet is also available in various languages. With reference to Internet world states (2011), there were 2,099,926,965 Internet users in the world, the group of top ten speaking languages accounted 82.2% and the rest of the speaking languages was only 17.8%. Among these languages, English and Chinese speakers account for more than 50% of all Internet language users in the world (See in Figure 2.1). Nowadays, Google provides translation service to 66 languages and an advanced search function (i.e., pages in the language selected) service to 46 languages.

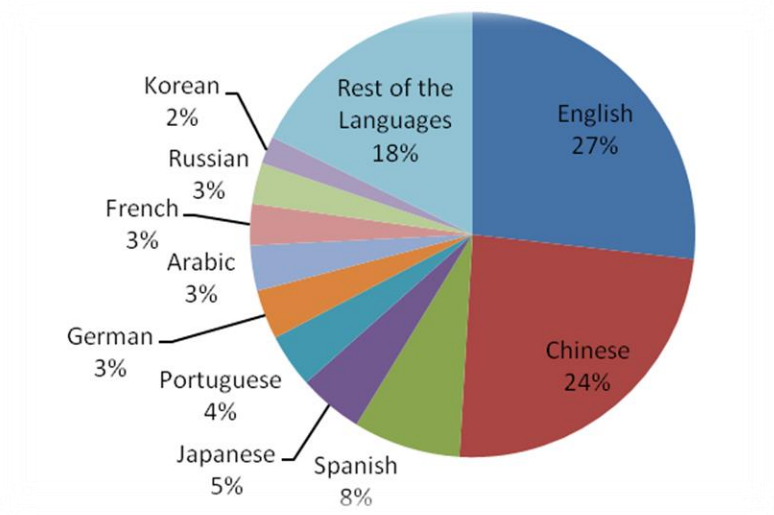


Figure 1.1: Top Ten Languages Speaking Users in the Internet 2011

ref: Internet World States-www.internetworldstates.com/stats7.htm, May 31 2011.

Cross Language Information Retrieval

Due to the exponential increase in non-English users on the Internet in recent years, there has been a demand from end-users for a Cross-Language Information Retrieval (CLIR) system which is more efficient and applicable to modern technology. This is because, if users have limited foreign language vocabulary, it can be difficult for them to use effective query words to retrieve the relevant documents. In most of the cases where the information is only available in a foreign language, CLIR applications and Machine Translation (MT) systems are helpful to provide access for local language users. In principle, CLIR refers to the task of retrieving documents in one language (e.g., English) with a query in another language (e.g., Japanese), but a major challenge exists in CLIR that is required to cross the language barrier in some way, typically involving translating either the query or the document from one language to the other [Zhai, 2009].

Machine Translation

Machine Translation (MT) is the process of translating from source language text into the target language. The use of a good MT system can help to translate the queries in CLIR system. In fact, the availability and performance of CLIR and MT depend on the availability of the lexicon/parallel corpora. Particularly, most of the lexicons do not provide a good coverage of proper nouns (e.g., people name, toponyms, science, and engineering terms, etc.) which are represented as phonetic borrowing terms/ transliteration/Out-Of-Vocabulary (OOV) terms in the languages.

Transliteration Mining

There are various tasks in the process of an MT system; transliteration mining is a significant task of MT to find the transliterated word pairs in parallel or comparable corpora in the word alignment phase of building a MT system. In particular, transliteration mining can be used both to handle OOV query words in CLIR system and to improve alignment during training time and help enrich phrase tables with name entities that may not appear in parallel training data in MT system [Kahki, 2012].

Problem of Out-of-Vocabulary terms

Proper nouns and terminologies form independently in any language and these new names and terms appear among the various spoken languages every day. When the queries may concern current affairs, they contain either new words that are out of scope in the translation lexicon, or recently appeared proper nouns such as personal names, place names, brand names, terminology that is not included in the translation lexicon [Ying, 2006]. There is no translation lexicon which can cover all recently appeared names and terms and thus the problem of OOV terms is a persistent problem for CLIR and MT systems [Raghavendra, 2009]. For example, users from Myanmar want to search for information related to the current Myanmar President's news that appears in various languages on the web, but most users have limited multilingual knowledge and they might use either Google language translation or a manual translation service as per their target search languages.

A proper noun example

For example, Table 2.1 presents the number of hits for a query word using Myanmar President's name in multiple languages such as "Thein Sein" in English, "テイン・セイン" in Japanese, "吳登盛" in Chinese, " เทียน เส่ง" in Thai, "Тейн Сейн" in Russian, " سـ ين ثـ ين الرئ يس " in Arabic, "테인 세인" in Korean, and "သိန်းစိန်" in native Myanmar language. It was found that a common query in English returned a total number of 1,460,000 hits for relevant pages. And also queries in some other languages returned a higher number of hits too (see in Table. 2.1).

Language	Query Word	No. of Hits
English	Thein Sein	1,460,000
Japanese	テイン・セイン	59,400
Chinese	吴登盛	75,500
Thai	เทียน เส่ง	32,300
Russian	Тейн Сейн	20,200
Arabic	سين ثين الرئيس	375,000
Korean	테인 세인	44,500
Myanmar	သိန်းစိန်	49,700

Table 1.1: Myanmar President's name in multilingual search (attempted on 2013/08/15)

Examples of a well-known proper noun and a technical term

Language	Query Word	No. of Hits
English	Bangkok	199,000,000
Japanese	バンコク	14,300,000
Chinese	曼谷	24,400,000
Thai	บางกอก	10,200,000
Russian	Бангкок	210,000,000
Arabic	بانكوك	3,150,000
Korean	방콕	213,000,000
Myanmar	ဘန်ကောက်	114,040

Table 1.2: City name "Bangkok" in multilingual search (attempted on 2013/08/15)

For examples, consider searching for the well-known proper noun "Bangkok" and the term "Solar system", Google returned a massive number of hits not only in common English, but also in each translated query in different languages such as Japanese, Chinese, Thai, Russian, Arabic, and Myanmar (i.e., see in Table 2.2 & 2.3). If the cross-language search was enabled, users could access the information regardless of language boundaries.

Language	Query Word	No. of Hits
English	Solar Systems	3,250,000
Japanese	ソーラーシステム	1,500,000
Chinese	太陽能系統	175,000
Thai	ระบบสุริยจักรวาล	52,600
Russian	солнечная система	920,000
Arabic	ال شمسي ال نظام	335,000
Korean	태양 광 발전 시스템	3,950,000
Myanmar	ဆိုလာ စနစ်	22
Hebrew	שמש מערכת	4,710

Table 1.3: The term "Solar Systems" in multilingual search (attempted on 2013/08/15)

Lack of technology

Although a lot of literature in CLIR includes many techniques for retrieving OOV words/phonetic borrowing words from more or less languages, such techniques are still rare for all languages. In spite of the fact that currently Google provides translation service to 66 languages, cross-language search methodology is still lacking so far. In addition, most of transliteration models in MT system applied phonetic and orthographic transformation rule, but coverage is only for specific languages (i.e., source and target bilingual).

Motivation

In order to overcome the problem of OOV terms in CLIR and MT systems, it is necessary to develop a methodology that can be used to retrieve phonetic borrowing words/ transliterated words/OOV terms accurately. Moreover, the advantages of a CLIR system are not only limited to individual users of the Internet, but many business, government, social, education, and multi-international organizations can also benefit from the ability to perform searches across different languages [Ying, 2006]. In fact, the performance of CLIR and MT systems depends on good support from transliteration model to mine accurately the transliteration word pairs. Therefore, we believe our methodology in learning approach of transliteration mining has many potential applications such as mining training data for transliterations and improving lexical coverage for MT and CLIR via translation resource expansion.

1.2 RESEARCH GOAL AND OBJECTIVES

Goal

The primary goal of this research is to develop a cross-language phonetic similarity metrics (CLPSM) to improve the performance of CLIR and various MT tasks.

Objectives

The objective of the study is to develop a cross-language phonetic similarity measurement methodology to address the issue of phonetically similar terms (phonetic borrowing word/transliterated words/out-of-vocabulary terms) in the applications of CLIR, MT, and various linguistics researches. The highlights are as follow:

- Based on Soundex, a Cross-Language Sound Grouping (CLSG) table for Asian languages is developed to help the distinction between loanwords/phonetic borrowing words and semantic adopting words.

- The method of the learning approach is built on stochastic edit distance within two frameworks: one-to-one Expectation Maximization (EM) [Ristad & Rianilos, 1998] and many-to-many Bayesian alignment [Fukunishi et al., 2012]. One-to-one EM alignment is useful when aligning two sequences of romanized characters, whereas many-to-many Bayesian alignment (our approach extends [Fukunishi et al., 2012]) is advantageous to use non-romanization scripts without restriction of the source and target sequence lengths.
- Adding a noise model (i.e., non-transliteration sub-model) to the EM and Bayesian models substantially improved their mining performance.
- Effectiveness of methodology in the learning approach that can be applicable in CLIR, MT, and various linguistic researches without employing any rich linguistic data resource.

Attempted experiments

The following methods were developed as part of experimental work to pursue the above mentioned objectives:

- An initial attempt applied Levenshtein edit distance in manual designing approach to test our CLSG;
- Stochastic edit distance based learning models attempted to mine the transliterations/phonetic borrowing words, and evaluate the performance of the methodology.

Expected Benefits

The work suggests the usefulness of CLSG (i.e., an extension of Soundex) for Asian languages that can be employed in CLIR, MT, and various linguistic applications such as mining training data for transliterations/OOV terms, improving lexical coverage for MT, and CLIR via translation resource expansion. Experimental results show that successful judgment of the phonological similarity between the original/semantic adopting words and phonetic borrowing words significantly improves performance of learning tasks. This dissertation points towards it being most useful when working with low resource languages. There is knowledge of CLSG based on the phonemes of Asian languages, and the method allows room for modifications and can be easily adopted as per required additional language. With a practicable approach of phonological knowledge, this methodology confirming to a standard framework can be part of any CLIR and MT applications.

1.3 RELATED WORKS

The notion of phonetic distance or, conversely, phonetic similarity is operated in many theoretical and applied areas of computational linguistic researches [Nerbonne et al., 2006]. As per literature, research studies into similarity measurements can be categorized into two main classes, namely, multilingual studies and, bilingual or monolingual studies. In order to measure phonetic similarity between languages, there are several extensions of the fundamental edit distance algorithms in two principle approaches: manual designing approach and learning approach [Kondard et al., 2006]. In this research, both approaches are attempted to measure phonetic similarity across Asian languages. Here, some reviews of significant research in the two principle approaches are discussed.

1.3.1 Reviewing Manual Designing Approach

Thai-English cross-language transliterated-word retrieval

Researchers looking at measuring phonetic similarity have presented an algorithm for Thai-English cross-language transliterated-word retrieval on the World Wide Web and CD-Rom applications based on the Soundex algorithm [Suwanvisat et al., 1998]. The Soundex phonetic coding table is modified to incorporate Thai characters by adding three more groups for using vowels as consonants in Thai language (e.g., ว/w/) and extending the code to unlimited length. It supports only Thai-English (bilingual) word retrieval and their results scored 80% accuracy on recall and precision measurements. However, they could only efficiently evaluate words with more than 4 characters.

Myanmar personal name matching system

A similar study had been done in Myanmar language to match the personal names that can be applicable in various information systems such as national citizen database, text and web mining, information retrieval, online library system, e-commerce, and record linkage system [Yuzana, 2008]. It has offered a sound-group mapping algorithm based on seven types of consonants related to the place of articulation in Myanmar syllables and measured the phonetic similarity using a pattern matching algorithm. Although their collation accuracy achieved 95.88%, it is only mono-language matching (i.e., within Myanmar characters).

Phonetic models for generating spelling variants

Bhagat et al. (2007) have proposed two phonetic models to generate the spelling variants of personal name. Firstly, using the CMU pronunciation dictionary[1], the EM algorithm is employed to learn the mappings between letters and the corresponding

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

phonemes in both directions. Then combing a noisy channel model based translator generated the best phoneme sequences, whereas a revised translator generated the variants of phoneme sequences under a supervised learning approach. Secondly, the Soundex method is used to generate numerous candidate variant spellings of a name. They used a list of names containing about 89,000 last names and 5,500 first names from the US census as their test bed. Their baseline system is used to measure the variants by Levenshtein edit distance and compared the results. Even though the results showed improvement with a respectable precision rate of 68%, their pronunciation learning and Soundex algorithms could not work accurately with the names derived from different languages.

Phonetic string matching using several algorithms

A similar attempt has been done by Zobel et al. (1996) who developed several new algorithms (i.e., Ipadist, Editex, Tapered editex, Tapered edit and Best agrep) and used some existing algorithms (i.e., Soundex, Phonix, and Phonix+) to compare for phonetic matching, and applied measurement techniques (i.e., Edit distance, Q-gram) for information retrieval to judge them. Methods in Editex and Ipadist combine the properties of edit distance with Soundex and Phonix, performance of these methods are significantly improved, whereas other methods gave poor results. Even though their successful algorithms accurately calculated the measurements, it worked only on monolingual data in English.

Improving Soundex retrieval

Similar research of integrating several phonetic algorithms has been attempted by Holmes et al. (2002). Their approach fused Fuzzy Soundex, Celko, Russell, Pfeifer techniques used to assign phonetic codes of multiple length and Dice Co-Efficient, Ngrams techniques to score the similarity. Experiments tested using a corpus containing 14,972 surnames, their approach of integrating multiple phonetic algorithms apparently improved recall and precision, whereas only Russell Soundex retrieved 658 of the 1,187 relevant names in search. However, their approach is limited to mono- or bilingual studies.

Cross Linguistic name matching in English and Arabic

A quite different study was conducted by Freeman et al. (2006) who experimented on cross-language name matching between English and Arabic. They used the Basis Artrans transliteration tool to transform Arabic letters into English and created an equivalent sound class (Character Equivalence Class) and developed two new algorithms called baseline and enhancements that are based on Second-String and Levenshtein distance. Enhancements integrate on the character equivalence classes and normalization of character strings (i.e., both in Arabic and in English). Results confirmed that the enhancement method is more effective than the baseline method.

The proposed approach in manual designing

Therefore, most of the research conducted on non-European languages has been limited to monolingual and bilingual studies; this is mainly because the Soundex algorithm is monolingual and Levenshtein distance does not directly use the knowledge base for measurements. The initial attempt in manual designing approach uses knowledge of CLSG (extension of Soundex) for Asian languages (i.e., multilingual) and integrates Levenshtein edit distance to measure the phonetic similarity across the languages.

1.3.2 Reviewing Learning Approach

Traditional mining system

In the statistical learning approach, traditionally transliteration mining systems have been applied to reasonably large-scale data resources in various language scripts, which have been studied in several prior works.

Improving CLIR by transliteration mining and generation

In order to improve the performance of cross-language information retrieval (CLIR), in [K Saravanan et al., 2011] the effect of integrating transliteration mining and transliteration generation techniques into CLIR was studied. They found that transliteration mining techniques were able to give better results than applying transliteration generation techniques. An experiment was done in the context of Hindi-English and Tamil-English on the standard FIRE 2010 dataset[2]. The transliteration similarity model was built using a W-HMM word alignment model [He, X., 2007] to determine whether document term was a transliteration of the query term. The expectation maximization (EM) algorithm was used to estimate the model parameters and transliteration similarity score of each source and target pair (ws,wt) was defined to be log(wt|ws). They combined both techniques, but this approach did not produce significantly better results than using transliteration mining alone.

Transliteration mining with phonetic conflation

In [Kareem, 2010], a generative transliteration model was trained using limited resources by using two methods: phonetic conflation and iterative training of the transliteration model. Phonetic conflation used a Soundex like conflation scheme for English. The experiment tested transliteration from ACL 2010 NEWS workshop shared

[2] Forum for Information Retrieval Evaluation (FIRE) http://www.isical.ac.in/~clia/data.html

transliteration mining task data containing five source languages (Arabic, Chinese, Hindi, Russian, and Tamil) to the target, English. The transliteration model with phonetic conflation gave much improved recall and F-measure in general, but the recall for transliteration mining between English and Chinese was very low. The model without phonetic conflation gave improved recall but often at the expense of precision.

Transliteration mining using graph reinforcement

Arabic-English transliteration mining using large training and test datasets was performed by applying a graph reinforcement method in [Kahki et al., 2011] [Kahki et al., 2012]. The baseline transliteration mining was trained by using a Bayesian generative model and the alignment of the character pairs was done by using an HMM based aligner [He, X., 2007]. Each source/target character sequence used in the alignment had a maximum length of 3 characters along with their associated mappings into the target language. Although a large amount of training data yielded more correct initial mappings, it tended to increase the errors. A method of graph reinforcement that led to sizable improvement in precision was introduced.

Stochastic edit distance

A classical stochastic model that learned a string edit distance function from a corpus of examples was proposed in [Ristad & Yianilos, 1998]. Edit weights were learned for four primitive edit operations: identity, insertion, deletion, and substitution (the edit operators that are used in the standard Levenshtein distance). The generative model learned multiple edit paths by using an expectation-maximization (EM) algorithm in an unsupervised manner. The expectation step accumulated expected counts for each edit operation on the training corpus. The maximization step set the model parameter values using these expectations. The total probability of all edit operations beings normalized in the maximization step. This approach is applicable to various applications involving string similarity and it was shown empirically to reduce the error with respect to using untrained Levenshtein distance with unit edit costs. We use their model as baseline system that will refer to as EM (Expectation Maximization). Their model is fused by the knowledge of CLSG in our proposed method.

Stochastic edit distance with noise model

A model of semi-supervised transliteration mining was proposed in [Hassan et al., 2012] which incorporates an explicit model for the generation of non-transliteration pairs (which will be referred to as the noise model). The model classifies unseen pairs by comparing the probabilities assigned by the transliteration sub-model, and the noise sub-model. In EM training, a parameter λ, the prior probability of generating a non-transliteration pair is learned along with model parameters representing the probabilities of each edit operation. Experiments were conducted on four language pairs: English-Arabic, English-Hindi, English-Tamil and English-Russian. Their results show that semi-supervised mining performed much better than an unsupervised approach. The current system is limited to learning unigram character alignment. In this experimental section, their approach is used as a baseline system that will refer to as EMn (since this model is trained using the EM algorithm, and includes a model for the noise). The proposed method extends their model to allow the integration of human knowledge of phonetic features (CLSG) and describe this in detail in the next chapter 3.

Bayesian alignment

As an alternative to the EM alignment approach of [Hassan et al., 2012], a non-parametric Bayesian alignment approach was proposed by the authors in [Fukunishi et al., 2012], [Finch et al., 2010]. This Bayesian approach has a tendency not to build models that over-fit the data and is therefore suitable for learning a many-to-many bilingual alignment model. In [Fukunishi et al., 2012] the model was used to align the data to be mined, and features from this alignment were used to classify the data. The classifier was trained on a set of seed sentences that were known to be correct (supplied as part of the NEWS workshop task), and a set of negative examples that were selected from the data. Their results yielded levels of precision and recall that were comparable with the best systems in the NEWS2010 shared task, for all of the language pairs tested. One weakness of their approach is that they make no attempt to screen out noisy data when training their alignment model, and erroneous parameters learned from the noise may degrade the performance of the model for some types of data (such as the dictionary data used in these experiments). In other words, their system may learn to model the noise, and consequently learn to mine pairs that are similar in character to noisy examples that were trained on.

My approach in learning

This study extends their approach to include an explicit noise model in order to mine word pairs in a low-resource environment, where knowledge of cross-language sound grouping can compensate for a lack of data resources. The alignment technique described in Section 3.3 is based on a Bayesian non-parametric model in which the prior and posterior distributions are not parametric distributions, but stochastic processes. The model is termed non-parametric not because it has no parameters, but because the number of parameters is not fixed and evolves during the training of the model. The detail of methodology is described in the chapter 3.

1.4 STRUCTURE OF THE THESIS

This thesis is organized into seven chapters.

Chapter 1:Introduction

Firstly, the background and motivation of this research are introduced, as well as the main goal and objectives. Afterwards, related works are reviewed in two different approaches: manual designing and learning approaches. Then, the structure of the thesis is described.

Chapter 2:Definition of Basic Terms

A review of the concepts and theories employed in this research is presented. The basic concepts and terms (i.e., phonetic similarity, linguistic borrowing, phonetic transcription, Soundex phonetic coding system, and edit distance similarity measures) that are necessary to understand the methodology are explained.

Chapter 3:Methodology

The main contribution of this dissertation begins in this chapter. The development of Cross-Language Sound Grouping (CLSG) and edit distance based similarity measure techniques in two principle approaches are demonstrated. Initially, development of CLSG for Asian languages is reported in detail. Afterword, development of CLPSM based on classical edit distance (Levenshtein Edit distance) in a manual designing approach is reported. Finally, development of CLPSM based on three stochastic models, namely, stochastic edit distance (EM), stochastic edit distance with noise (EMn), and Bayesian alignment with noise (BAYESn) in a learning approach are reported.

Chapter 4:Experiment of Classical Edit Distance Measures

The experimental results and evaluation in a manual designing approach are presented in detail. The experiment uses the names of 92 chemical elements words in eight Asian language pairs: English-Japanese, English-Korean, English-Malay, English-Myanmar, English-Thai, English-Indonesian, English-Vietnamese, and English-Chinese.

The results of two CLSG versions are compared in two metrics (LD and NS). We published these experimental tasks and findings in International Journal of Intelligent Information Processing [Htun et al., 2011].

Chapter 5:Experiment of Stochastic Edit Distance Measure

The experiment results and evaluation in a learning approach by using Stochastic Edit Distance (EM), Stochastic Edit Distance with noise model (EMn) and Bayesian Alignment with noise model (BAYESn) are presented. The experimental procedure consists of segmentation and romanization of Myanmar language data, and building Myanmar-English bilingual training corpus consisting of 3,100 single word pairs and 14,891 multiple word pairs. The results of the three different models are also compared. We published these experiments in International Journal of Computer Applications [Htun et al., 2012].

Chapter 6:Discussion

In this chapter, we discuss about the analysis of our methodologies such as experimental results, the nature of tested languages, and the feature of datasets. Finally, the pros and cons of methodologies in two approaches are determined.

Chapter 7:Conclusion and Future Research

The research findings and contributions of this work are summarized. It is also concludes some brief discussions on limitations of proposed techniques, and presents the future direction of work.

Chapter : 2

Definition of Basic Terms

2.1 PHONETIC SIMILARITY

What is phonetic similarity

Phonetically similar segments are two or more sounds which share phonetic features and are frequently found as variants of a single phonological unit not only in a language, but also across the languages. Most phonetically similar segments are adjacent to each other in a phonetic chart, and differ only slightly in one or two articulatory features. Examples of phonetically similar segments and the difference between the given segments are shown in Table 2.1 [Eugene et al. 2003].

The phonetically similar segments	Differences in articulatory features
[p] and [b]	voicing
[p] and [f]	place of articulation and manner of articulation
[l] and [r]	manner of articulation
[p] and [t]	place of articulation

Table 2.1: Examples of phonetically similar segments and the difference between the given segments

2.2 LINGUISTIC BORROWING

Linguistic borrowing

Usually languages borrow words freely from one another, linguistic borrowing happens when some new object or institution is developed for which the borrowing language (recipient language) has no word of its own for example [Encyclopedia Britannica]. However, many words are shared in different languages a result of linguistic borrowing in various language contexts, but it depends on subject domain, economic situation, and political situation of the country as well as historical ties. Many Asian languages employ the words in phonetic borrowing from European languages especially for science and engineering terminologies.

Sounds of Asian languages

There are many loanwords in Japanese that are from Chinese and European languages. Especially from English, some vocabulary items and sounds were not in the phonological inventory of Japanese before the borrowing occurred. As an example, the Japanese phonological system did not have the /f/ sound before the Meiji era, but it was introduced into Japanese by borrowing the /f/sound from English [Kodama, 2010]. Katakana (i.e., a domestic phonetic script) is mostly used to transliterate loanwords (except those borrowed from Chinese). As another example, Myanmar consonantal gaps in comparison to English are the lack of /f, v, ɹ, ʒ/. Therefore, English [p^h] is consistently adapted with the Myanmar ပ/p/ rather than the Myanmar ဖ/p^h/, which is used instead to represent the English [f] [Charles, 2008].

History of borrowing word in Asian languages

In addition, Sanskrit and Pali loanwords have influenced Asian languages such as Myanmar, Thai, Malay, and Indonesian, especially related to religious and scholarly terms. Because of the influence of France while Vietnam was a French colony, the Vietnamese adopted many French words [Barker, 1969]. There are about 20,000 loanwords in Korean, of which almost 90% are from English [Ho-Min Sohn, 1999]. There are not only a considerable number of loanwords from Japanese, but also many Western loanwords were re-borrowed from Western loanwords in Japanese. Most of the loanwords are from specific terminologies (e.g. science and engineering). In addition, Kana in Japan and Rito moji (吏読文字) in Korea are based from Chinese characters writing system [Mikami, 1999] and these languages are created by combining a semantically relevant word-root of Chinese. Malay and Indonesian are known for their use of loanwords. They have adopted a number of words from different languages throughout their history such as those coming from Arabic, Chinese, English, Dutch, Hindi, Japanese, Portuguese, Sanskrit, Tamil, etc [Hiramoto, 2007].

Borrowing Types

When a word in a language is similar to a word in another language that can be considered as two types of word borrowing [Matsuda et al., 2008], [Kodama, 2010]: Semantic borrowing (Calque[3]) and phonetic borrowing (Importation). In this section, the sequence in the brackets [ ] means romanization, sequence in slashes / / means IPA phoneme, and the letter in < > indicates the translation of the word into English.

2.2.1 Semantic Borrowing

What is semantic borrowing

An expression of foreign/source language is translated directly into the recipient language, element by element (i.e., creating a new word by combining a semantically relevant word-root or word in the recipient language).

Semantic borrowing example in Japanese

English: <Biology >→ Japanese: 生物学/[seibutsugaku] [MTD, 2009]

In this example, the term <Biology> is translated directly into a compound of words 生物学/[seibutsugaku] in Japanese; 生/[sei] means <life, living>, 物/[butsu] means <matter, object, thing>, and 学/[gaku] means <learning, science, study> [Denshi

[3] A calque is a complex form that was created on the model of a complex form in a donor language and whose can constituents correspond semantically to the donor language constituents [Martin et al., 2009]

Jisho]. Each kanji character is related to the definition of the Chinese character and were combined by coining more than a hundred years ago.

Semantic borrowing example in Korean

English: <Carbon> → Korean: 탄소/[tanso] [MTD, 2009]

<Carbon> is semantically translated into Korean as 탄소/[tanso], but it refers to the word-root of Chinese 碳/[Tàn] which is the same as in Japanese.

Semantic borrowing example in Myanmar

English: <Frequency> → Myanmar: ကြိမ်နှုန်း/[krim nhun:] [Myanmar-English, 2008]

The term <frequency> is purely created as the new word ကြိမ်နှုန်း/[krim nhun:] in Myanmar.

Semantic borrowing example in Vietnamese

English: <Gold> → Vietnamese: vàng/[vang] [MTD, 2009]

<Gold> is translated originally in Vietnamese word as vàng/[vang]. Because the color of gold is yellow, which is the name of the color called "vàng" in Vietnamese. So that [vang] means golden or yellow.

Semantic borrowing example in Malay

English: <Mercury> → Malay: raksa/[raksa] [MTD, 2009]

The term <mercury> is semantically translated into Malay as "raksa".

Semantic borrowing example in Indonesian

English: <Image> → Indonesian: gambar/[gambar] [MTD, 2009]

<Image> is translated into Indonesian as "gamber".

English: <System> → Thai: ระบบ/[Rabb] [MTD, 2009]

The term <system> is translated in Thai original word as " ระบบ/[Rabb]".

Semantic borrowing example in Chinese

English: <Computer>→ Chinese: 电脑/diàn-nǎo/ [MTD, 2009]

The Chinese word for <computer> is 电脑 /diàn-nǎo/ that literally means "electronic brain".

2.2.2 Phonetic Borrowing

What is phonetic borrowing

An expression is introduced into the recipient language by representing a pronunciation that is close to that of the source language because some vowel and consonant sounds in English or other source language that do not exist in the recipient language are represented by the nearest recipient language equivalents. Usually loanwords are affected by phonological adjustments within the borrowing languages that can be classified into four types of phonetic adjustment [Kodama, 2010] as follows:

i. Phoneme insertion

Insert extra segments into the loanword.

English: <Chip> → Japanese: チップ/[chippu]

English: <Pipe> → Korean: 파이프/[paipeu]

English: <Domain> → Thai: โดเมน/[doomeen]

ii. Phoneme deletion

Drop the final consonant of the borrowed word.

English: <Series> → Malay: siri/ [siri]

English: <Manganese> → Vietnamese: mangan/ ˈmæŋgən /

iii. Incorporation of phoneme of borrowed language

A phoneme in the borrowed language is accepted and incorporated into the phonological inventory of the borrowing language (i.e., the same phoneme sequence).

English: <Keyboard> → Myanmar: ကီးဘုတ်/[ki: bhut]

Japanese: 柔道/[judo] → English: <Judo>

iv. Replacement of phoneme in the source language

Sound change in phonetic borrowing

A phoneme in the source language is replaced by the phonologically similar phoneme in the destination language. Usually speakers of one language have difficulty to reproduce the sounds of another language that do not exist in their native language. Most foreign words are changed phonetically when transcribed into recipient Asian languages. In this case, the nearest sounds in each recipient language represent the equivalent sounds of foreign words.

Sound change examples from English to Japanese

Some vowel and consonant sounds in English do not exist in Japanese. The following examples are phonologically changed between English and Japanese word pairs [Gillian Kay, 1995]. Each sound in English is replaced by the closet sound in Japanese.

Sound change	**English**	**Japanese**
[f]/ f / → [h]/ h /	<earphone>	イヤホン/[iyahon]
[d]/ d /→ [j]/ j /	<rhodium>	ロジウム/[rojiumu]
[l]/ l / → [r]/ r /	<gallium>	ガリウム/[gariumu]
[s]/ s / → [sh]/ ʂ /	<silicon>	シリコン/[shirikon]
[th]/ θ / → [s]/ s /	<thread>	スレッド/[sureddo]

[t]/ t / → [ch]/ tʂ^h /	<platinum>	プラチナ/[purachina]
[v]/ v / → [b]/ b /	<vanadium>	バナジウム/[banajiumu]
[z]/ dz /→ [j]/ ɟ /	<zirconium>	ジルコニウム/[jirukoniumu]

Sound change example from Chinese to Japanese

Another example of Japanese Kanji orthography, /ŋ, v, t^h/ in Chinese are replaced by /g, b, t/ in Japanese [Kodama, 2010]. For example, 五/ŋ˩/ in Chinese (Wu) means "five" in English is changed into Japanese as 五/go/.

Sound change	**Chinese (Wu)**	**Japanese**	
/ ŋ / →/ g /	五/ ŋ˩ / <five>	五/ go /[go]	[Wiki-1, 2013]

Sound change example from English to Thai

In loanword adoption, the voiced labiodental fricative /v/ in English is changed by the Thai speakers into /w/ which is a closer sound [Kenstowicz, 2006].

Sound change	**English**	**Thai**
/ v / → / w /	<vanadium>	วาเนเดียม/ [wānedeīym]

Sound change example from English to Myanmar

Similarly, Myanmar consonantal gaps in comparison to English are the lack of /f, v, ɹ, ʒ/ sounds. However, the phoneme /p^h/ in English is consistently adapted with the Myanmar consonant ပ/p/ rather than ဖ/p^h/ which is used instead to represent /f/ in English; the phoneme /p/ is changed in Myanmar to ပ/p/ [Charles, 2008].

Sound change	**English**	**Myanmar**
/ f /→ /p^h /	<fluorine>	ဖလိုရင်း/[p^{h}luirang]

Sound change example from French Vietnamese

When French words are adopted into Vietnamese [Barker, 1969], they are usually changed to fit the phonemic system of Vietnamese. Because of the lack of an initial labial-plosive /p/ in Vietnamese, it is substituted by /b/ in initial position.

Sound change	**French**	**Vietnamese**
/ p / → / b /	pupe/ [pupe]/<doll>	búp-be /[bup-be]

Sound change example from English to Malay

The alphabetical order of Indonesian and Malay is identical with that of Dutch and English, though orthography of those languages has some variances. There are (29) consonants and (6) main vowels in Malay, but only six consonants (i.e., /m/, /n/, /f/, /l/, /s/, and /y/ are pronounced as same as in English [Seman, 2008].

Sound change	**English**	**Malay**
/ v / → / p /	< vase >	pasu /[pasu]

Sound change example from English to Indonesian

In early times, /j/ sound was used in Indonesian (i.e., borrowed from Dutch). After 1972, /j/ was changed into /y/ sound [Hiramoto, 2007].

Sound change	**Dutch**	**Indonesian**
/j/ → /y/	jodium/[jodium]<iodine>	yodium/[yodium]

Sound change example from English to Korean

Phonetic representations of loanword in Korean are interpreted and structured according to the contrastive categories of the native language. Initial consonants /p, t, k/ are changed to /pp, tt, kk/, whereas /b, d, g/ are altered to /p, t, k/ [Iverson, 2006]. Also the L-sound from the source language is changed to an R-sound like Japanese.

Sound change	English	Korean
/ l /→ / r /	<lobby>	로비/[robi]

Sound change example from English to Chinese

Phonetic borrowing in Chinese is extremely rare, a least as far as the standard and various written forms are concerned. The syllables `ga`, `ka`, and `ha` are frequently used in phonetic loans (i.e., especially for inner-Asian place names) [Martin, 2009].

Sound change	English	Chinese
/c /→ /g/	<nicotine>	尼古丁/[niguding]

2.3 Phonetic Transcription

Definition of phonetic transcription

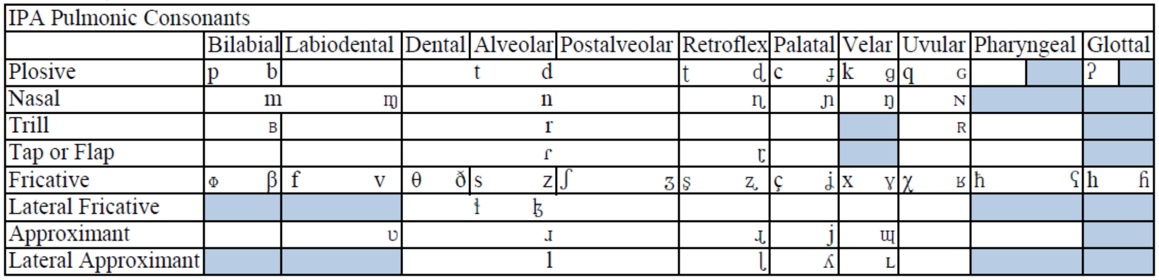
Phonetic transcription is the use of phonetic symbols to represent speech sounds [Wells, 2006]. Alternatively the characters of one language are converted to the characters of another language in accordance with the pronunciation of the target language. Depending on the target language (e.g. Japanese or Thai) for one and the same source language can be used. The spelling of a word in a language is often not the same in pronunciation; usually a language dictionary includes phonetic transcriptions (i.e., phonetic notation). The International Phonetic Alphabet (IPA) is widely used for phonetic transcription of speech sounds for English and IPA letters are incorporated into the alphabets of various languages. For example, 'mother' in English and 'အောင်ဆန်းစုကြည်' [Aung San Suu Kyi] in Myanmar are represented in IPA notation as /mʌðə/ and /ʔaʊ̃sʰã̃sṵkɹɲ/[4]. The following section describes a brief review of IPA.

[4] Accessed 2013/03/26, http://dl.dropbox.com/u/8589366/Files/IPA%20converter/converter_ipabur.html

2.3.1 International Phonetic Alphabet

IPA symbols

The International Phonetic Alphabet is based on the Latin alphabet, using a few non-Latin forms for some extra-ordinary sound values, because IPA symbols include one or more elements of two basic types, letters, and diacritics. For examples, sounds like /tʃ/ in church and /θ/ in thin, author. There are 107 letters, 52 diacritics, and four prosodic marks; any given language normally involves exploiting only a small subset. There is no letter holding context-dependent sound values, for example, /c/ does in English and several other European languages. The IPA letters are organized into three categories: pulmonic[5] consonants, non-pulmonic consonants, and vowel. Most IPA letters are pulmonic consonants, for example, all consonants in the English language fall into these categories. Pulmonic consonant letters are arranged and grouped in horizontal columns with the place of articulation such as labial, coronal, dorsal, radical, and glottal. Moreover, in vertical columns with the manner of articulation are followed such as nasal, stop, fricative, approximant, flap or tap, trill, lateral fricative, lateral approximant, and lateral flap. Other groups are organized as diacritics, other symbols, supra segmental, and ones & word accents [IPA]. Figure 1.1 illustrates IPA symbols relating to consonants [Ladefoged, 2001].



IPA Pulmonic Consonants

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral Fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral Approximant				l		ɭ	ʎ	ʟ			

Figure 2.1: IPA symbols relating to consonants

[5] A pulmonic consonant is a consonant produced by air pressure from the lungs, as opposed to ejective, implosive and click consonants.

2.3.2 Romanization

Definition of romanization

Romanization is the representation of either a written word or spoken speech with the Roman (Latin) script, or an original word or a language in a non-Roman (non-Latin) writing system is changed by representing in the Roman (Latin) script [Romanization]. In order to romanize the script from non-Latin languages, methods includes transliteration and transcription for representing the written text and the spoken word respectively. Transcription in spoken word can be subdivided into phonemic transcription (i.e., records the phonemes or units of semantic meaning in speech) and phonetic transcription (i.e., records precisely the speech sounds). Each romanization of language has its own set of rules for pronunciation of the romanized words. Beside there are several different romanization schemes in every language. For example, however Japanese has mainly three romanization systems: Hepburn romanization, Kunrei-shiki Rōmaji (ISO 3602), and Nihon-shiki Rōmaji (ISO 3602 Strict), variants of the Hepburn system are the most widely used [Romanization of Japanese]. Another example, Myanmar uses various romanization systems such as the Myanmar Language Commission Transcription System (MLCTS) [Naing, 2006], ALA-LC Romanization [CONGRESS, 1997], BGN/PCGN Romanization [US, 1994], and Okell [Okell, 1971]. Many other examples in non-Latin script languages can be found in different romanization schemes: The Royal Thai General System of Transcription [Royal Thai, 2002] for Thai, Romanization of Korean and Revised Romanization Korean [The Korean language, 2000] for Korean, and Pinyin Romanization of Mandarin Chinese [Francis, 1984] for Chinese.

2.4 SOUNDEX PHONETIC CODING SYSTEM

A brief history of Soundex

The originative approach to phonetic similarity measuring is the Soundex indexing system, the best known algorithm that was developed by Robert Russell and Margaret Odell (1918 and 1922). In principle, Soundex is a phonetic algorithm for indexing names by their sound when pronounced in English. Names with the same pronunciation can be encoded to the same string and matching occurs despite minor differences in spelling. The original Soundex has been revised and alternatives of Soundex have been developed in English and many other languages for their particular purposes. Thus Soundex is usually adapted for use with other language by taking into account the characteristics of the particular language, for examples, Double Metaphone, Spanish Metaphone, Brazilian Portuguese Metaphone, the Daitch-Mokotoff Soundex for German, Fuzzy Soundex, Phonix, etc. A sort of Soundex-based phonetic matching technique that has found widespread application in the linguistics domain over a long period of time in the indexing system of database (e.g. U.S. National Archives), and is still finding application in recent years for example in Oracle , MySQL, Microsoft SQL Server etc.

Soundex rules

In the Soundex algorithm, phonetic representations of the letters of the alphabet were divided into distinct categories and assigned a numeric value in each category. The exact principle is as follows [Soundex][Soundex-Ex]:

a. Retain the first letter of the string

b. Remove all occurrences of the following letters, unless it is the first letter: a, e, h, i, o, u, w, y

c. Assign numbers to the remaining letters (after the first) as follows:

Code	Letters	Description of Speech Sounds
1	B, F, P, V	Labials
2	C, G, J, K, Q, S, X, Z	Gutturals & Sibilants
3	D, T	Dental
4	L	Long Liquid
5	M, N	Nasal
6	R	Short Liquid
Delete	A, E, H, I, O, U, W, Y	Vowels plus H, W, & Y

Table 2.2: The Original Soundex Coding System

d. If two or more letters with the same number were adjacent in the original name (before step a), or adjacent except for any intervening h and w, then omit all but the first

e. Return the first four bytes appended with zeros

2.5 EDIT DISTANCE SIMILARITY MEASURES

Concept of similarity and distance relationship

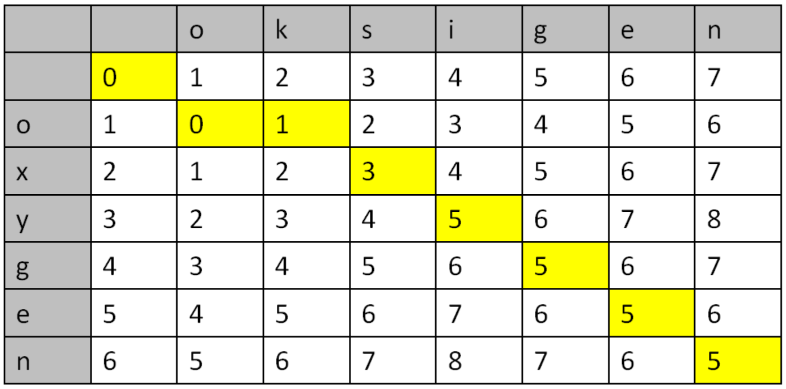
The concept of similarity is fundamentally important in almost every scientific field [Ennis et al. 2007]. One assumption of the psychological theory of similarity is that closer objects are more similar than objects that are far apart. Typically, the similarity between two objects is related to their similarities (commonalities) and differences. From a theoretical point of view, similarity and distance ought to be each other inverses [Herringa et al., 2006]. In principle, similarity measurement can be derived from the distance measure using a concept (1 - distance). Similarity measurements are used to determine the verity of mathematical foundations of common techniques such as

Manhattan distance, Euclidean distance, Jaccard distance, Dice's coefficient, Cosine similarity, Hamming distance, Levenshtein distance, and Soundex distance.

2.5.1 Levenshtein Distance

What is the Levenshtein distance

The Levenshtein distance [Levenshtein, 1966] is also referred to as edit-distance and is primarily an algorithm used to investigate a channel model considering the problem of constructing optimal codes capable of correcting deletions, insertions, and substitution (i.e., also known as edit operations). It is a string comparison metric that counts the least number of edit operations that are necessary to modify one string to obtain another string. The cost is normally set to one unit (cost/weight) for each edit operation. For example, the Levenshtein distance between English and Malay phonetically similar word pairs "oxygen" and "oksigen" is distance 3. However, if the cost of substitution is set to 2, the distance between these words becomes 5 (See in Figure 2.2).



		o	k	s	i	g	e	n
	0	1	2	3	4	5	6	7
o	1	0	1	2	3	4	5	6
x	2	1	2	3	4	5	6	7
y	3	2	3	4	5	6	7	8
g	4	3	4	5	6	5	6	7
e	5	4	5	6	7	6	5	6
n	6	5	6	7	8	7	6	5

Figure 2.2: Calculation of edit operations in Levenshtein Distance .

Example applications

Edit distance has been applied successfully in a wide range of applications, for instance, spell checkers, correction systems for optical character recognition, etc.

2.5.2 Stochastic Edit Distance

Stochastic edit distance model

The standard Levenshtein distance for string sequences was given a stochastic interpretation by [Ristad & Yianilos, 1998] (i.e., stochastic model allows one to learn the string edit distance function from a corpus of examples), where a stochastic transduction was used to define two string edit distances. The first distance called Viterbi edit distance is defined by the most likely transduction between the two strings (i.e., the negative logarithm of the probability of the most likely edit sequence for the string pair). The second distance (i.e., also known as stochastic edit distance) is defined by aggregating all transductions between the two strings (i.e., the negative logarithm of the probability of the string pairs according to the transducer). In estimation, the parameters of the memoryless stochastic transducer (which will be reinterpreted as edit costs) are learned with the Expectation Maximization (EM) algorithm using a forward-backward dynamic programming that consists of two steps. In the expectation step, each edit operation of generated string pairs is accumulated. In the maximization step, each parameter value is set on its relative expectations on training data. The Expectation Maximization algorithm finds a locally optimal set of parameters in terms of the likelihood of the model given the data. It is applicable to any string classification problem that may be solved using a similarity function against a database of labeled prototypes.

2.5.3 Bayesian Alignment

Bayesian model

The principle approach of Bayesian modeling is updating the degree of belief in a hypothesis given prior knowledge and given available evidence. Both prior knowledge and evidence are combined using Bayes' rule to obtain the posterior hypothesis [Volker, 2006]. Bayesian models are commonly applied to many scientific disciplines, including the field of natural language processing. In particular they have been applied successfully to word segmentation and bilingual word/character alignment. The manner of Bayesian models are constructed growing of model with few parameters, therefore they align consistently and tend to not overfit the data [Finch et al., 2011a].

Dirichlet process

In Bayesian alignment [Fukunishi et al., 2012], a Dirichlet process is a stochastic process that is defined over a set of all possible bilingual sequence-pairs of which sample path is a probability distribution on those bilingual sequence-pairs. In this way, bilingual sequence-pair generation processes are assigned as an instance of the Chinese Restaurant Process (CRP)[6]. The model includes two basic processes as follows:

The Base Measure

Another process assigns a probability to an outcome that has not yet been produced. A joint spelling model assigns probability to new sequence-pairs according to the joint distribution.

[6] The dish served at its table corresponds to every bilingual sequence-pair in potential infinite set of tables in Chinese restaurant. Each table is seated by the number of customers that represents the accumulative count of the bilingual sequence-pair. A new customer can take a seat at an occupied table with a probability proportional to the number of customers at that table, and allowed to eat that table's dish, or can take a seat at an unoccupied table with a probability proportional to a constant, in which case they must eat a dish (a bilingual sequence-pair) chosen by the chef (the base measure) [Fukunishi et al., 2012]

The Generative Model

A process for generating an outcome that has already been generated at least once before. In learning, the prior knowledge is formulated as a prior distribution over parameters and the evidence corresponds to the observed bilingual sequence-pairs. The probability distribution of bilingual sequence-pairs can be learned directly from unlabeled data by Bayesian inference of the hidden alignment of the corpus.

Chapter : 3

Methodology

Framework of methodology is built upon the edit distance measure techniques by integrating the knowledge of Cross-Language Sound Grouping (CLSG) to provide a structure for the edit costs. The development of phonetic similarity metrics attempts in two principle approaches: manual designing approach and learning approach. A method of classical edit distance (Levenshtein Distance) is initially attempted in manual designing approach. Mainly proposed method in learning approach is attempted within two different models based on stochastic edit distance: a method based on one-to-one Expectation Maximization (EM) alignment [Ristad & Yianilos, 1998] and a non-parametric Bayesian many-to-many alignment approach [Fukunishi et al., 2012]. Figure 3.1 depicts framework of methodology.

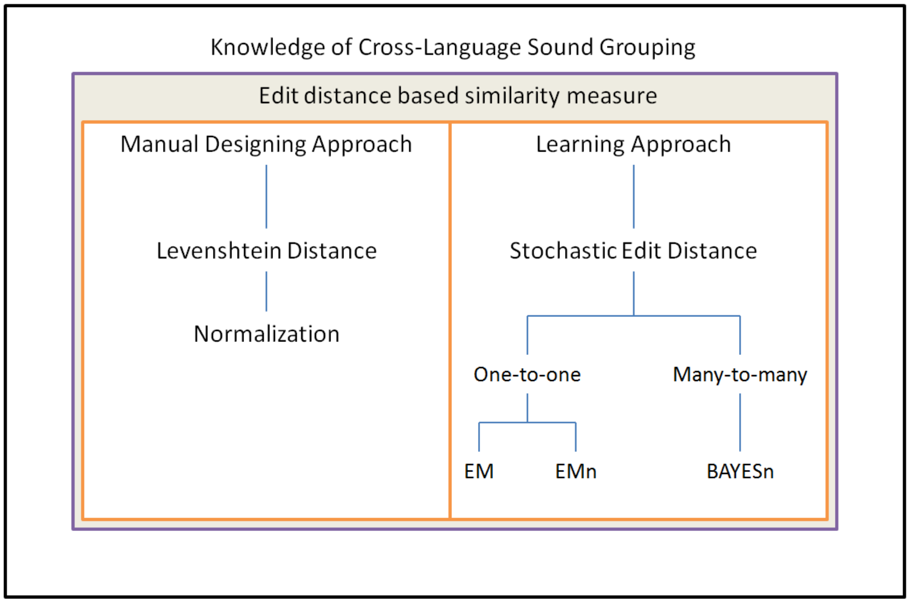


Figure 3.1: Framework of Methodology

3.1 EXTENSION OF SOUNDEX PHONETIC CODING FOR ASIAN LANGUAGES

Origin of Soundex

Cross-Language Sound Grouping (CLSG) was developed based on the originative approach of Soundex [R.C. Russell et al., 1922] and a concept of phonetics and phonology of Asian languages. Soundex is language dependent especially based on English phonemes. When assigning codes and phonetic categories to languages other than English, typically the characters cannot be directly mapped onto the original Soundex categories. The Soundex is usually adapted for use with other languages by taking into account the particular characteristics of each language. Therefore, the Soundex scheme is not enough to achieve this research purpose, a revised Soundex scheme is developed especially for Asian languages by adding particular sounds symbols and groups.

A brief description of Soundex

In the Soundex representation, names with the same pronunciation are encoded to the same string, therefore matching can occur despite minor differences in spelling like "Smith" and "Smythe". The original Soundex codes have four alphanumeric characters consisting of a letter and three numbers. The letter is always the first letter of the surname and then the numbers are assigned to the remaining letters of the surname. Moreover, all vowels and h, w, and y letters are dropped except those occurring in the first letter, because it is based on the phonological concept that vowels are not pronounced.

Requirement of measuring phonetic similarity in Asian languages

In order to measure the similarity between two words, the length of each character string can be more or less than four characters. In addition, many Asian languages use borrowed words not only from English, but also from Chinese, Sanskrit, Arabic, Portuguese, French, etc., in which some h, w, and y letter sounds are used as consonants (see in Table 3.1).

Vowel/Language	**a**	**e**	**i**	**o**	**u**	**y**	**w**	**h**
Japanese						や、ゆ、よ	わ、を	は
Korean								ㅎ
Malay						y	w	h
Myanmar	အ					ယ	ဝ	ဟ
Indonesian						y	w	h
Thai						ญ, ย	ว	ห, ฮ
Vietnamese								h

Table 3.1: Using English vowels as consonants in some Asian languages

Type of loanwords

Usually loanwords are affected by phonological adjustments within the borrowing languages that can be classified into four types of phonetic adjustment as following (refer to section 1.2.2):

i. Insertion of phoneme

ii. Deletion of phoneme

iii. Incorporation of phoneme of borrowed language

iv. Replacement of phoneme in the source language

Requirement for different types of loanwords

However, loanwords in the first and second types of phonological adjustment can be accorded to code using Soundex, loanwords in third and fourth types of phonological adjustment cannot be accorded. As mentioned before, this is because many Asian languages use borrowed words from various source languages. Furthermore, some consonants are formed with two letters for a single phonetic unit (See Table 3.6) or some other consonants are combined with affixes from the local sound (i.e., prefix or suffix) on an original word. In addition, non-Latin script languages such as Japanese, Korean, Myanmar, and Thai need to be Romanized. Therefore, a large amount of phonological adjustment has occurred between source and recipient words and they would encoded as far apart. The following tables 3.2~3.5 show some deviant examples of the encoding process using Soundex in each type of loanword.

	Word in character	After vowel deletion	Soundex coding
English	smart	smrt	2563
Japanese	sumaato	smt	2530

Table 3.2: A positive example of insertion of phoneme in Soundex

	Word in character	After vowel deletion	Soundex coding
English	department	dprtmnt	3163~~553~~
Japanese	depaato	dpt	3130

Table 3.3: A positive example of deletion of phoneme in Soundex

	Word in character	After vowel deletion	Soundex coding
English	oxygen	$xgn	$225
Myanmar	aauksijaang (အောက်စီဂျင်)	$ksjng	$222~~52~~

Table 3.4: A negative example of incorporation of phoneme of borrowed language in Soundex

	Word in character	After vowel deletion	Soundex coding
English	sulfur	slfr	2416
Malay	sulfuroksida	slfrksd	2416~~223~~
English	volt	vlt	1430
Thai	woow	w	$000

Table 3.5: A negative example of replacement of phoneme of source language in Soundex

Examples of Soundex in different types of loanwords

Regarding to the above tables, examples in table 3.2 and 3.3 shows the results for loanword type (i) and (ii) using Soundex coding that can measure phonetic similarity of languages. Whereas examples in table 3.4 and 3.5 show the results from loanword type (iii) and (iv) using Soundex that are not suited for incorporation of phoneme and sound changed in replacement of phoneme of source language.

Extension of Soundex

Therefore, the Soundex system is not enough to achieve the purpose of this research, an extension of the Soundex scheme is built for Asian languages by adding particular sounds symbols and groups. For examples, there is no /f/ consonant in Myanmar, but /ph/(ဖ) is the closest sound to /f/ for phonetic adjustment. Similarly, some English consonants, such as bilabial approximant /w/, r-sounds /r/ and palatal approximant /j/, do not occur in Vietnamese. Moreover, both /f/ and /v/ sounds do not exist in Japanese and Malay. There are no /v/, /z/, and /th/ (as in 'the' or 'there') sounds in Thai. Sanskrit has appeared as many loanwords in religious and scholarly terms in Asian countries, therefore /h/ and /y/ consonants are found in those languages. In order to cover all these Asian languages, the Soundex scheme is revised as follow:

1. In Soundex, /b, f, p, v/ consonants are grouped together in Labial place of articulation, but it is necessary to distinguish the phonetic variation between /b, p/ and /f, v/ for Asian languages and other non-English languages too. Each group is split based on manner of articulation such plosive and fricative.
2. However /h, w, y/ consonants in Soundex are discarded; these consonants are used in Asian languages and some non-English languages. Therefore, /w, y/ consonants are added in a new group as approximants and /h/ consonants are combined into Velar-Fricative.

Categorization of feature sounds in Cross-Language Sound Grouping ver.1

In this CLSG (Ver. 1) for Asian languages, a consonant of a language associates with three distinctive features such as place of articulation, manner of articulation, and voicing (See in Table 3.6). In fact, voicing is neglected and only place and manner of articulation features are included in this CLSG because voicing is not important in judging phonological similarity (Kodama, 2010). Therefore, phonetic alphabets categorize both voiceless consonants and voice consonants together into the same group under the articulation of place and manner features. Rhotic and lateral sounds are distinguished in the grouping because those sounds are born as functional roads in many languages. Nasals and approximants are grouped respectively based on the place of articulation. Plosives, affricates and fricatives are grouped according to their place of articulation first and then grouped by manner. But affricates and fricatives are categorized into the same group. Initial zero consonant is given for words that begin with vowels, because these beginning vowels behave as consonants. Then each group is assigned to one character that is applied to calculate the similarity.

Articulatory place	Articulatory manner	Symbol	IPA	Assigned Code
Labial	Plosive	p, b	p, b	1
	Fricative	f, P,v	f, (ɸ), v,(β)	2
Dental	Plosive	t, d	t, d	3
	Fricative	T, D, s z S Z C J, j	Θ, ɚ, s, z ʃ, (ʂ,ɕ) ʒ (ʐ,ʑ) ʧ, ʦ(c), ʤ, ʣ (j) ɟ	4
Velar	Plosive	k,c,q,x g	k, (q) g, (ɢ)	5
	Fricative	X h H Q	ç, x, (X) h (ɦ, ħ, ʕ) ɣ (ʝ, ɕ)	6
Other Sounds				
r-sounds		r R	r-sound Used when another r-sound in language	7
l-sounds		l L	l Other laterals	8
Nasals		m n G N	m n, (ɳ) ɧ Used when there is N uvular nasals	9
Approximants		y w Y W	j w (ɥ) (ʍ)	A
Initial zero consonant		$	Ø(ʔ)	B

Table 3.6: Cross-Language Sound Grouping (Ver.1) 2010

Using IPA symbols

The International Phonetic Alphabet (IPA) for phonetic transcription of speech sounds is used, but IPA includes many trivial distinctions, which are not necessary for this research, various IPAs are grouped into one Latin alphabet. Table 3.6 shows the relation of phonetic features, phonetic symbols, baseline IPAs, and the assigned codes.

Refine to CLSG to cover various source and target languages

After an initial attempt at methodology development in manual designing approach, the CLSG (Ver.1) was refined again to distinguish expeditiously between various source languages (European, Sanskrit, Chinese, etc.) and target Asian languages. The place of articulation in the alveolar group is split and some required phonetic symbols are added into the map.
Example: /j/ sound like [**j**ibun] or [go**j**ū] is in alveolar -plosive in Japanese.

Problems in two letter form of consonant

In addition, two letters forming initial consonants in Table 3.7, which are appeared in many Asian languages .
Examples:
<cupid> (En)→ กามเทพ/[**kh**iwpit](Thai)
ငပလီ/[ngpli] (Myanmar)

These types of consonants are included in CLSG Ver.2, for example, [c, s] single consonants are in different articulatory place and manner of alveolar fricative and velar plosive separately; whereas [ch, sy] two letters forming initial consonants are in alveolar fricative together.

	Japanese	Korean	Malay	Myanmar	Indonesian	Thai	Vietnamese
Labial Plosive	py (ぴょ) by (びょ)			ph(ဖ)		ph(ผ,พ,ภ)	ph
Dental Fricative				dh(ဓ)		th(ถ,ฐ,ท,ฑ,ธ,ฒ)	th
Alveolar Plosive							
Alveolar Fricative	ch (ちょ) sy (しょ)	ch(ㅊ)	sy	hc (ဆ)			ch
Nasals	ny (にょ) my (みょ)	ng(ㅇ)	ng,ny	ng(င)	ng,ny	ng(ง)	ng

Velar Plosive	ky
	(きょ)
	gy
	(ぎょ)
			kh	kh(ခ)	kh	kh(ข,ฃ,ค,ฅ,ฆ)
						ch(ฉ,ช,ฌ)	

r-Sound	ry (りょ)						

Table 3.7: Two letters forming a single phonetic unit in some Asian languages

Illustration of CLSG ver.2

The refined CLSG for Asian languages (Ver. 2) is illustrated in Table 3.8. It is used in the following attempt at prototype experiments in the learning approach.

Articulatory Place	Articulatory Manner	Symbols	IPA	Assigned Code
Labial	Plosive	p	p	1
		b	b, ɓ	
		V, ph, py, by	ʋ	
	Fricative	f	f, (ɸ)	2
		v	v, (β)	
Dental	Plosive	t	t (ʈ)	3
		d	d (ɖ)	
	Fricative	th	θ	4
		dh	ð	
Alveolar	Plosive	j	ɟ	5
	Fricative	s, sh	s, ʃ	6
		z	z	
		zh	ʒ	
		ch, sy	tʃ	
		ge	dʒ	
Velar	Plosive	k, ky	k	7
		c, kh		
		q	q	
		x	χ	
		g, gy	g	
	Fricative	X	x, (χ)	8
		h	h, ç	
		H	ɦ, ħ, ʕ	
		Q	ɣ, ɟ	

Other Sounds				
	Nasals	m	m	9
		my	ɱɣ	
		n	n	
		N,hn, hm	ɳ	
		ng	ŋ	
		ny	ɲ	
	r-Sounds	r	ʀ	A
		rr, ry	ɾ, ʁ, ɹ, ɻ, ʀ, ɽ, ɺ	
	l-Sounds	l	ʟ	B
		LY	ɬ, ɫ, ʟ, ɭ, ɮ	
	Approximants	y	j	C
		w	w	
		yy	ɥ	
		hw	ʍ	
Initial Vowel Sounds		a, e, i, o, u	ʔ	$
Vowel		a, e, i, o, u	a, æ	Delete

Table 3.8: Cross-Language Sound Grouping (Ver.2) 2012

3.2 DEVELOPMENT OF CROSS-LANGUAGE PHONETIC SIMILARITY METRICS

Description of methodology

An initial methodology development is attempted to build a cross-language phonetic similarity metrics based on classical Edit-Distance (Levenshtein distance) in manual designing approach. The method will be referred to as CLPSM. The CLPSM algorithm consists of five steps: Romanization, vowel deletion, simplification of similar sounds, calculation of Levenshtein distance, and normalization. Figure 3.2 depicts the steps of the methodology.

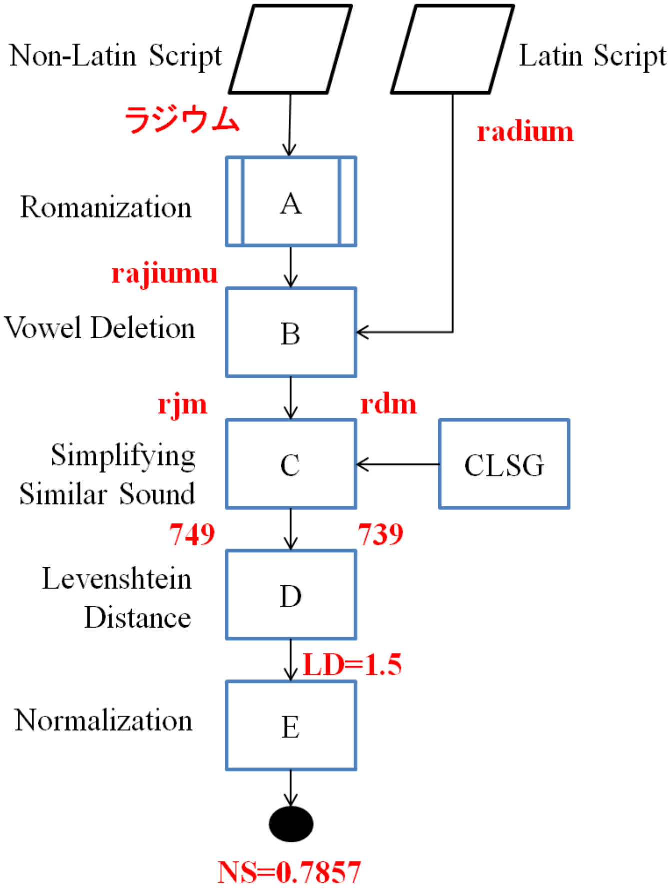


Figure 3.2: Process of Cross-language Similarity Metric (CLSM)

3.2.1 Romanization

Using rules and examples

Measuring phonetic similarity requires the character strings to have the same character set. Therefore, in the first step, the terms of non-Latin scripts are converted into Romanized alphabets during the pre-processing stage. Even though various Romanization rules exist for each non-Latin script language, an appropriate standard Romanization rule is selected in each non-Latin language. "99-SHIKI" Japanese Romanization system is applied for Japanese [社団法人日本ローマ字会, 1999-7], Revised Romanization of Korean for Korean [Korean language, 2000], Pinyin Romanization of Mandarin Chinese

(Pinyin) [Shibles, 1994] for Chinese, the Myanmar Language Commission Transcription System (MLCTS) [Naing, 2006] for Myanmar, and the Royal Thai General System of Transcription [Royal Thai, 2002] for Thai. Table 3.9 provides some examples of characters in the Romanization assignment of each target language.

Language	**Characters**				
Japanese	カ	キ	ク	ケ	コ
	ka	ki	ku	ke	ko
Korean	ㄱ	ㅋ	ㄴ	ㄷ	ㅌ
	g	k	n	d	t
Chinese	合	土	竹	小	用
	ge	tu	zhu	xiao	young
Myanmar	က	ဂ	င	◌ျ	◌ူ
	k	g	ng	-y-	ù
Thai	ก	ซ	ด	ต	บ
	k	s	d	t	b

Table 3.9: Example of Romanization characters in Japanese, Korean, Myanmar and Thai

3.2.2 Vowel Deletion

In this step, all vowels (i.e., a, e, i, o, u) from both the source and the target languages are eliminated. If a word (initial letter) begins with a vowel, it is not deleted (e.g. iodine → idn). This process complies with Soundex purposes.

3.2.3 Simplifying Similar Sounds

Simplifying language particular sound

Although the various language scripts are written in Latin scripts, the spelling does not always correspond directly to the pronunciation. Therefore, there is needed to transcribe spellings to phonetic notations [Kodama, 2010]. Simplifying similar sounds can be divided into two sub-steps: language-particular simplification and baseline simplification (Figure 3.2). A particular sound-mapping class is developed that corresponds to the native phoneme of each recipient language (i.e. language dependent). Table 3.10 shows some examples of the language-dependent sound mapping class in Myanmar [Naing, 2006], Thai [Royal Thai, 2002], and Vietnamese [Wii-vun, 2001].

Language	Sound change
Myanmar	/jh/="z", /kh/="k", /ch/= "s"
Thai	/ch/= "k", /dg/,="j", /qu/= "k"
Vietnamese	/x/= "s", /ng/= "n", /q/= "k"

Table 3.10: Part of language dependent sound mapping class

Simplifying baseline sound

The next step, baseline simplification is language independent and based on the IPA phonetic mapping table [Odden, 2005]. In principle, IPAs can represent phonetic transcription of speech sounds for all languages, but this research does not require such distinctions. Thus, some different IPAs are grouped into one symbol. Depending on articulation place and manner of sound, those symbols are grouped and each group has been assigned into a code. Table 3.7 specifies the relation of phonetic features, phonetic symbol, baseline IPAs, and the assigned codes used in this algorithm.

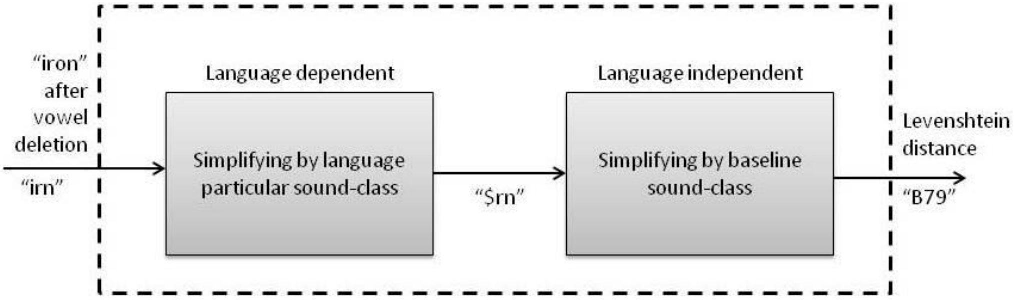


Figure 3.3: Steps of Simplifying Sound Process

Example

Figure 3.3 depicts the steps of simplifying sound process, after vowel deletion, a term such as "irn" (iron) in English is converted from symbol "$rn" to assigned code "B79".

3.2.4 Levenshtein Distance

Cross-Language Sound Grouping weighted formula in Levenshtein

Knowledge of CLSG is integrated into an edit cost of Levenshtein distance (this measurement will be referred to as LD) calculation as follow:

$$LD = D + I + \omega S \qquad (1)$$

where

D = the number of deletions

I = the number of insertions

S = the number of substitutions

ω = variable weight (i.e. referring to the group of sounds in a place of articulation and manner)

description of weight setting

In this calculation, each operation counts as 1, however, a variable weight ω is set up in substitution. In this process, if the relation of sound symbols belongs to the same place of articulation and manner, ω in each substitution sets 0.5. Even if the relation of sound symbols exists in the different place of articulation and manner, ω in each substitution sets 1. All these weight values are taken into account in the calculation of the Levenshtein distance. Results in LD, a score of zero represents a perfect similarity between two words.

3.2.5 Normalization

Normalized formula

Finally, LD is normalized in the process of calculation. The normalized similarity measurement [Bhagat et al., 2007] [Zobel et al., 1996] [Kodama, 2010], which is denoted here by NS is defined as:

$$NS = 1 - \frac{LD}{(L_1 + L_2)} \qquad (2)$$

where L1 and L2 are the length of the converted strings due to simplification of the sound process.

Purpose of normalization and conditional metric value setting

In fact, the normalization intends to eliminate the length of the string effect. LD is divided by the length of both strings to minimize the weight of a mismatched character in longer strings [Zobel et al., 1996] [Kodama, 2010] [Colin et al., 2008]. A score of 1 for NS (0<NS≤1) represents a perfect similarity between two words. Table 3.11 offers some examples of English and Japanese measuring results.

	Romanization	Vowel deletion	Simplify sounds	LD	NS
1	hydrogen	Hdgn	63759	4.5	0.3571
	suiso	Ss	44		
2	chromium	Crmm	4799	2	0.7143
	kuromu	Krm	579		
3	zinc	Znc	495	2	0.6
	aen	$n	B9		
4	radium	Rdm	739	0.5	0.9167
	rajiumu	Rjm	749		
5	terbium	Trbm	3719	0	1
	terubiumu	Trbm	3719		

Table 3.11: Examples of retrieved terms in Japanese

3.3 Development of Cross-Language Phonetic Similarity Metrics in Learning Approach

Phonetic based cost learning frameworks

The proposed method allows a human to provide a structure for the edit costs that are based around the phonetically-motivated model of phoneme sound groups (CLSG) described in Table 3.8 of previous section 3.1. The machine is able to determine precise values for these costs within three different frameworks: a stochastic string-edit distance [Ristad & Yianilos, 1998], stochastic sting-edit distance with noise model [Hassan et al., 2012] and a Bayesian alignment approach [Fukunishi et al., 2012] [Finch et al., 2010]. Figure 3.4 presents an overview of the experimental procedure used to evaluate the effectiveness of these frameworks which are described in the following sections.

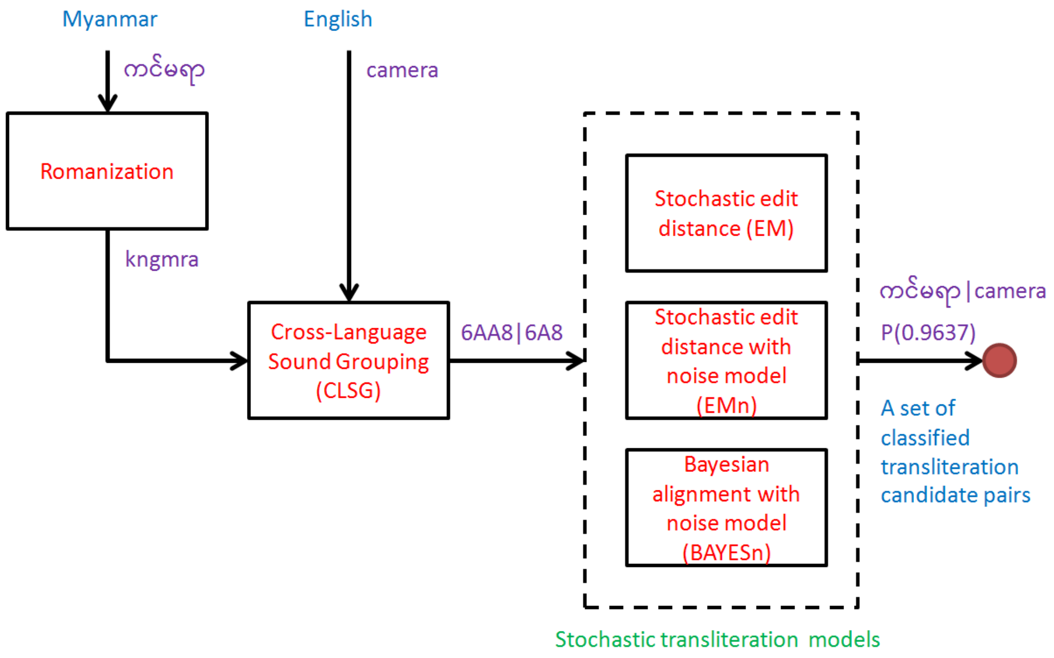

Myanmar
ကင်မရာ
Romanization
kngmra
English
camera
Cross-Language Sound Grouping (CLSG)
6AA8|6A8
Stochastic edit distance (EM)
Stochastic edit distance with noise model (EMn)
Bayesian alignment with noise model (BAYESn)
Stochastic transliteration models
ကင်မရာ|camera
P(0.9637)
A set of classified transliteration candidate pairs

Figure 3.4: Experimental framework

3.3.1 Stochastic Edit Distance

EM learning

String Edit Distance (Levenshtein distance) measures the dissimilarity between strings and is defined as the minimum number of edit operations needed to transform one string into the other, where an edit operation is an insertion, deletion, or substitution. Identity substitutions are defined to have zero cost. The edit distance can be calculated by a simple edit operation count, or each edit operation can be assigned its own cost. Often these edit costs are assigned plausible values by hand. As an example, a list of edit operations and their associated edit costs from a string which is represented as the sequence X=(a,b) to the string Y=(c,b) are given below:

cost((a,ϕ)) = 1 (deletion)

cost((ϕ,c)) = 1 (insertion)

cost((a,c)) = 1 (substitution)

cost((b,b)) = 0 (identity)

Using the above edit operations and costs, the Levenshtein distance is given by the minimum number of edits. In this example, only a single edit is required - a substitution of c for a - and the Levenshtein distance is therefore 1.

Defining edit distances

The standard Levenshtein distance for string sequences was given a stochastic interpretation by [Ristad & Yianilos, 1998], where a stochastic transducer was used to define two string edit distances: the Viterbi edit distance and the stochastic edit distance. The model is described in some detail below, as it forms the basis for the models in this research paper.

Process of stochastic distance

In a stochastic framework, the generation process is governed by a generative model that can assign a joint probability to a pair of strings using probabilities on edit operations. Under their interpretation, the joint probability is transformed into an edit cost by taking the negative logarithm. A given example illustrates visually how a stochastic edit distance is derived from the joint probability of strings X=(a,b) and Y=(c,b) generated by a memoryless stochastic transducer (all possible paths are shown in figure 3.5).

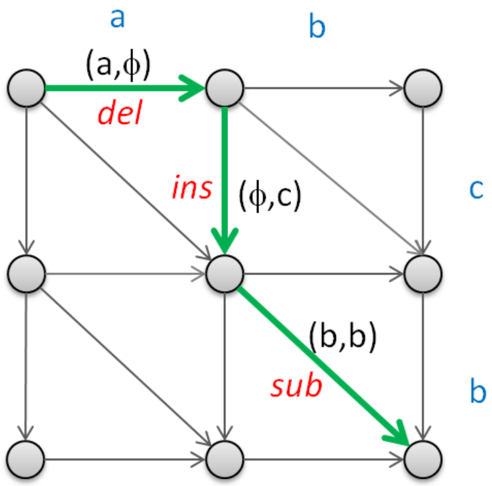


Figure 3.5: An edit path

Description of example learning

In the above figure, horizontal arcs represent deletions, vertical arcs represent insertions and diagonal arcs represent identity/substitution operations. One of the edit paths is shown in green and the corresponding sequence of edit operations is listed below:

(a, ϕ) = deletion

(ϕ, c) = insertion

(b, b) = identity/substitution

The probability of the cost of a single path s is the product of each associated edit cost belonging to the path and is defined as:

$$P(s) = \prod_{e \epsilon s} P(e) \qquad (3)$$

where e is an edit, and s =(e_1, e_2, e_3, ...e_n).is a sequence of edits (an edit path).

Viterbi edit distance

A source string *X* and a target string *Y* can be aligned in many ways, corresponding to multiple paths in the graph. The Viterbi edit distance is defined using the most likely edit sequence for the string pair <*X*,*Y*>.

$$P_v(X,Y) = max_{s\epsilon Z}P(s) \qquad (4)$$

where $Z = \{s_1, s_2, s_3, ..., s_j\}$ is the set of all edit operation sequences that can generate *X* and *Y*.

Stochastic edit distance

The stochastic edit distance $d_s(X,Y)$ is defined as the negative logarithm of the joint probability of the string pair $P(X,Y)$ according to a memoryless stochastic transducer [Ristad & Yianilos, 1998] [Hassan et al., 2012]. This is calculated by summing the derivation probabilities over all paths in the graph in Figure 3.4.

$$d_s(X,Y) = \sum_{s\epsilon Z}\sum_{e\epsilon s} -log(P(e)) \qquad (5)$$

Expectation maximization

The parameters of the memoryless stochastic transducer (which will be reinterpreted as the edit costs) are learned with the Expectation Maximization (EM) algorithm using a forward-backward dynamic programming technique for efficiency. The EM algorithm finds a locally optimal set of parameters in terms of the likelihood of the model given the data.

3.3.2 Stochastic Edit Distance with Noise Model

Adding noise model to EM

When applied to transliteration mining, the stochastic model of edit distance described in the previous section can be extended by adding a noise model (a non-transliteration sub-model) to the transliteration sub-model [Hassan et al., 2012]. The full transliteration mining model being an interpolation of both models:

$$P(X,Y) = (1-\lambda)P_t(X,Y) + \lambda P_n(X,Y) \qquad (6)$$

Where λ is the prior probability of the data being noise (a non-transliteration pair), P_t is the probability of transliteration sub-model, and P_n is the probability of noise model.

How do the models work

The transliteration sub-model aligns the characters to each other or to null using a unigram model. The EM training of the transliteration model is similar to that process described in the previous section. The noise model randomly generates characters using two independent unigram models, and is estimated once at the start of training from the training data. After EM training, the transliteration word pairs are expected to be assigned a high probability by the transliteration sub-model and a low probability by the noise model. This model will be referred as EMn (EM alignment with noise model).

3.3.3 Bayesian Alignment with Noise Model

Adding noise model to BAYESn

This model incorporates the essence of the ideas proposed in [Hassan et al., 2012] into a non-parametric Bayesian learning framework [Fukunishi et al., 2012]. It contains a similar explicit noise model, and to do so introduces an additional generative step that selects the type of word pair the model will generate. This technique will be referred to as

BAYESn (Bayesian alignment with noise model). The structure and characteristic of this model is described as follow:

3.3.3.a Overfitting

The motivation for extending the EM model to a Bayesian model is the desire to use many-to-many alignment. One-to-one alignment is useful when aligning two sequences of romanized characters, but usually cannot be used for non-roman scripts. A major limitation of maximum likelihood training when applied to bilingual alignment is its tendency to overfit the data. Assigning a large amount of probability mass to long sequence pairs in the data will produce a model with a high likelihood. In the most extreme case where there are no restrictions on the source and target sequence lengths in the many-to-many mapping, the most likely model will assign a probability of one to a single alignment of the entire source side of the corpus to the entire target side. Nonparametic Bayesian models discourage the addition of long pairs into the model, by assigning them a low probability and by rewarding the re-use of parameters in their models [Finch et al., 2011].

3.3.3.b Model Structure

Figure 2 shows the structure of the model, which consists of two square, graphs corresponding to the transliteration sub-model and the noise sub-model. The generative story for the model is a 2-step process as follows:

Step 1: Choose whether to generate a noise pair (with probability λ) or a transliteration pair (with probability 1-λ);

Step 2: Generate a pair of the chosen type using the appropriate model.

Model training

For details of how the transliteration sub-model is trained the reader is referred to [Fukunishi et al., 2012]. This model differs from theirs in that it performs clustering into two classes as it learns the probabilities for its model parameters. The transliteration sub-model is thereby trained using only those clean samples that fall into the transliteration class. The λ probability is updated based on a simple frequency count whenever a transliteration candidate is assigned a new class. It is possible either to train the noise model in the same manner as [Hassan et al., 2012] so that it is trained only on noisy data, or it can be trained once at the start of training from the whole of the training corpus. According to the observation of pilot experiments, both of these techniques are effective and gave approximately equal performance. The model is chosen to train once at the start of the training for the experiments in this paper. The classes for the candidate pairs were assigned randomly in the first iteration of training using a noise probability λ=0.5.

Properties of BAYESn

In [Finch et al., 2010], the aligner performed a forced many-to-many alignment in the spirit of the alpha/beta edit operations proposed in [Brill et al., 2001], but did not include the capability to make null alignments. In this work their model is extended to allow null alignments to multiple characters in both languages. In Figure 3.6 for simplicity only one-to-one alignment is illustrated, but in reality arcs that traverse greater distances in the graph are possible but are limited by parameters that control the maximum spans of an edit operation in terms of the number of source and target characters it can operate on.

Learning example

As shown in the figure below, a generative model generates X and Y clustering and aligning into two sub-models: transliteration model and noise model. An edit path of each sub-model is shown in green and the corresponding sequence of edit operations is listed on the right.

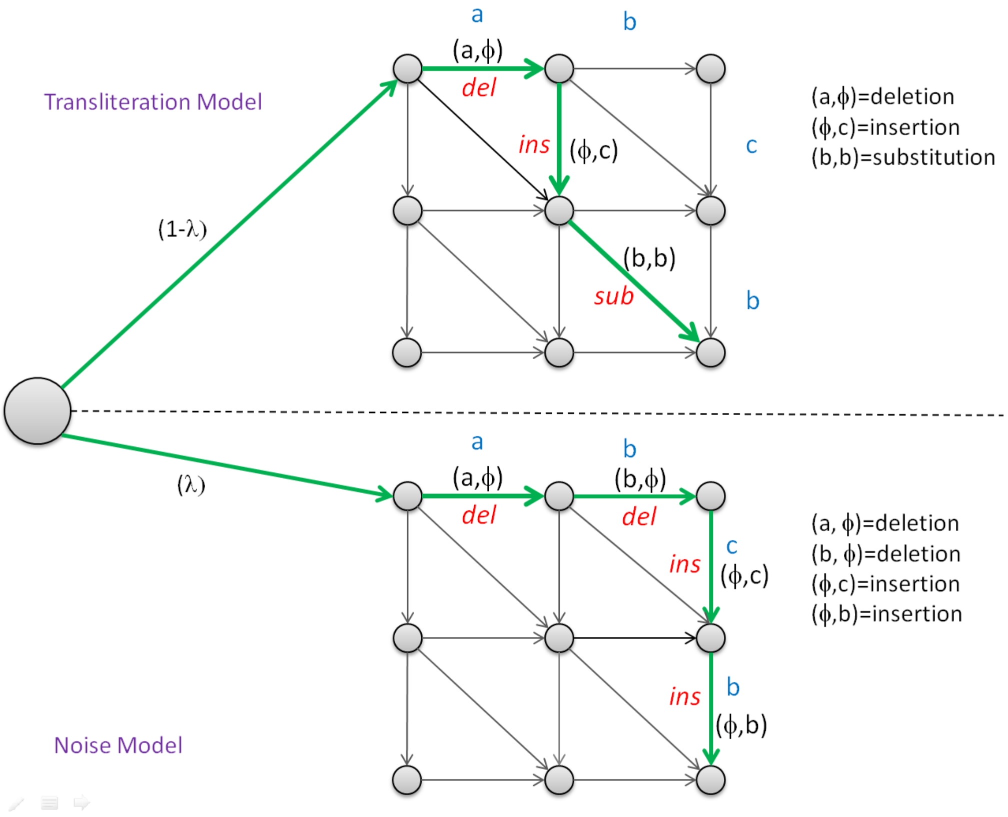


Figure 3.6: Sub-models of BAYESn

Chapter : 4

Experiment of Classical Edit Distance Measures

4.1 DATA

The names of ninety-two chemical terms from periodic table [Emsley, 2011] were deployed as the test dataset in this initial experiment. The test dataset was composed of eight language pairs: English-Japanese, English-Korean, English-Malay, English-Thai, English-Myanmar, English-Indonesian, English-Vietnamese, and English-Chinese. We chose these elements' language pairs because the etymological origins of these element names are already known in the recipient languages (i.e., whether the name of an element is a phonetic borrowing or semantic adopting from the source language). The terms in non-Latin script languages like Japanese, Korean, Myanmar, Chinese, and Thai are romanized with their own set of rules for pronunciation of romanized words (refer to section 3.2.1).

4.2 RESULTS

The results in two distance measures described in chapter 3, LD is a baseline Levenshtein distance measure with weights computed using Cross-Language Sound Grouping (CLSG) version 1. NS refers to LD with normalization. The scores of the distance and similarity in each language pair are partitioned into groups (i.e., phonetic borrowing and semantic adopting groups) by allocating a threshold. Then the results are compared and explanations for the results will be suggested in following section 4.3.

4.2.1 Retrieving Terms in Languages

Retrieval results from LD

The allocation of threshold (i.e., refers to section 4.3.1), it is assumed that the scores of distance in LD are equal to or less than the threshold suggests to the phonetic borrowing word-pairs and above the threshold indicates to the semantic adopting words. Table 4.1 describes the number of identified phonetic borrowing words by assigning a threshold (T_{LD}) in LD for each language measurement.

Source: English	**Threshold (T_{LD})**	**Origin of terms**		**Identified results**	
		Phonetic	**Else**	**Phonetic**	**Else**
Japanese	3	72	20	79	13
Korean	3	73	19	80	12
Malay	3	87	5	87	5
Thai	2.5	79	13	74	18
Myanmar	2.75	81	11	83	9
Indonesian	3	84	8	88	4
Vietnamese	3.25	82	10	87	5
Chinese	3.5	73	19	55	37

Table 4.1: Thresholds and number of retrieved terms in LD results

Retrieval results from NS

The scores of similarity in NS are equal to or greater than the threshold to the phonetic borrowing word-pairs indicated and below the threshold suggested for the semantic adopting word-pairs. Table 4.2 reveals the number of identified phonetic borrowing words by assigning a threshold (T_{NS}) in NS for each language measurement.

Source: English	Threshold	Origin of terms		Identified results	
	(T_{NS})	Phonetic	Else	Phonetic	Else
Japanese	0.5950	72	20	73	19
Korean	0.6333	73	19	75	17
Malay	0.6384	87	5	86	6
Thai	0.6071	79	13	79	13
Myanmar	0.6389	81	11	81	11
Indonesian	0.6384	84	8	84	8
Vietnamese	0.5428	82	10	83	9
Chinese	0.4714	73	19	31	61

Table 4.2: Thresholds and number of retrieved terms in NS results

4.2.2 Distribution of distance and similarity of word pairs in each language

English-Japanese

From an examination of LD results in English-Japanese, a proposed threshold at distance unit 3 cannot distinguish between phonetic borrowing and semantic adopting word-pairs completely (see in Figure 4.1), whereas, the distribution of terms based on the results of NS in English-Japanese (see in Figure 4.2), the word pairs above the 0.634 threshold differentiate phonetic borrowing word-pairs, while those below identify semantic adopting word-pairs (0<NS≤1). However, among the phonetic borrowing terms, potassium (K) is below the threshold, because kariumu (K) in Japanese derives from kalium in German. Zinc (Zn) and gold (Au) are above the threshold of NS. .

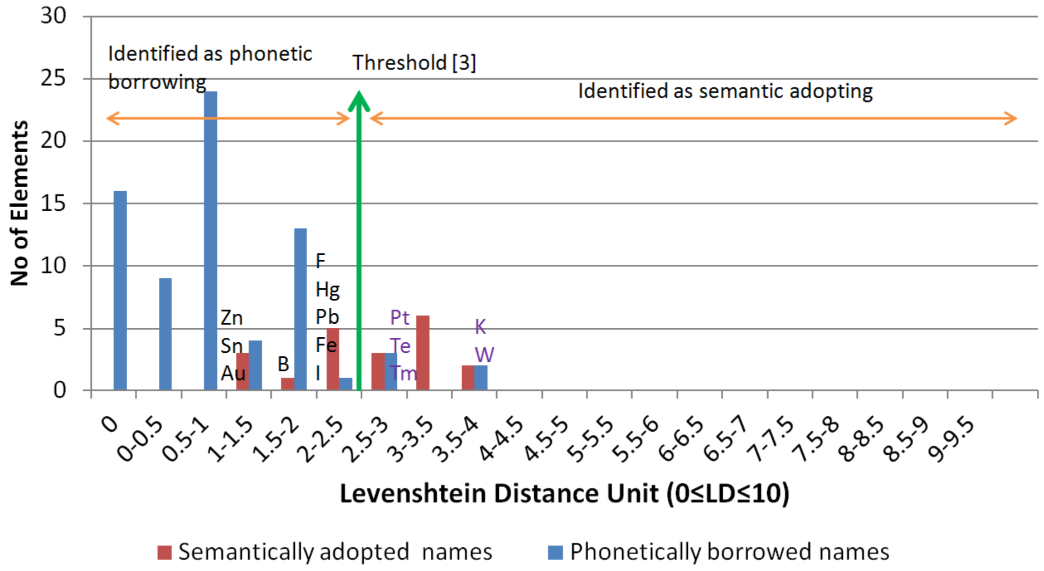


Figure 4.1: Distribution of phonetic and semantic borrowing words in Japanese (LD)

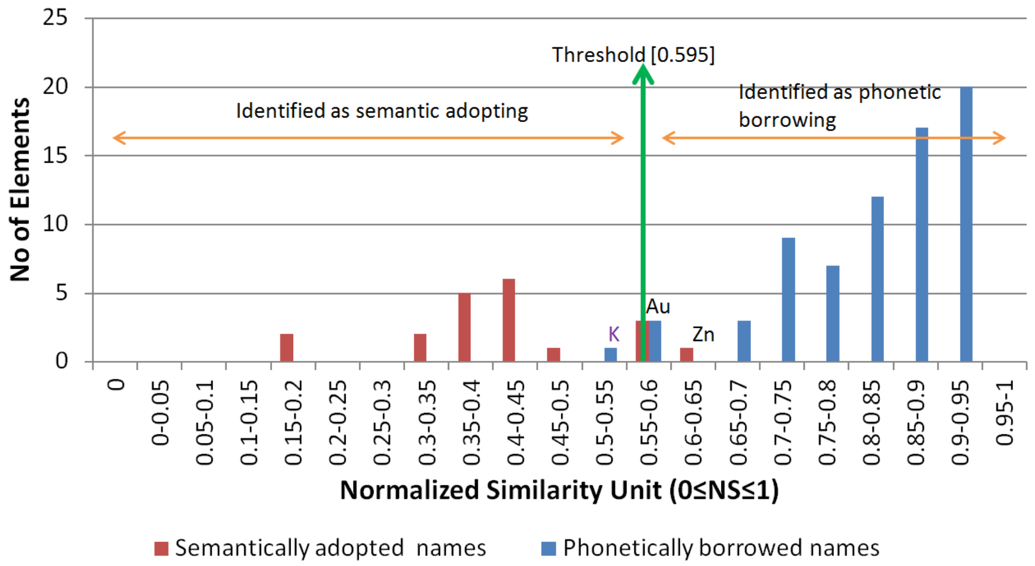


Figure 4.2: Distribution of phonetic and semantic borrowing words in Japanese (NS)

English-Korean

However, the distribution of elements in the LD result cannot differentiate between phonetic borrowing and semantic adopting words (Figure 4.3), allocation of threshold in NS result can differentiate (Figure 4.4), except in the case of a phonetic borrowing element name 볼프람/[bolpeulam] (W) from German, which is below the threshold, and some semantic adopting element names (boron (B), copper (Cu), and platinum (Pt)) are above the threshold, but their score are very closed to the threshold.

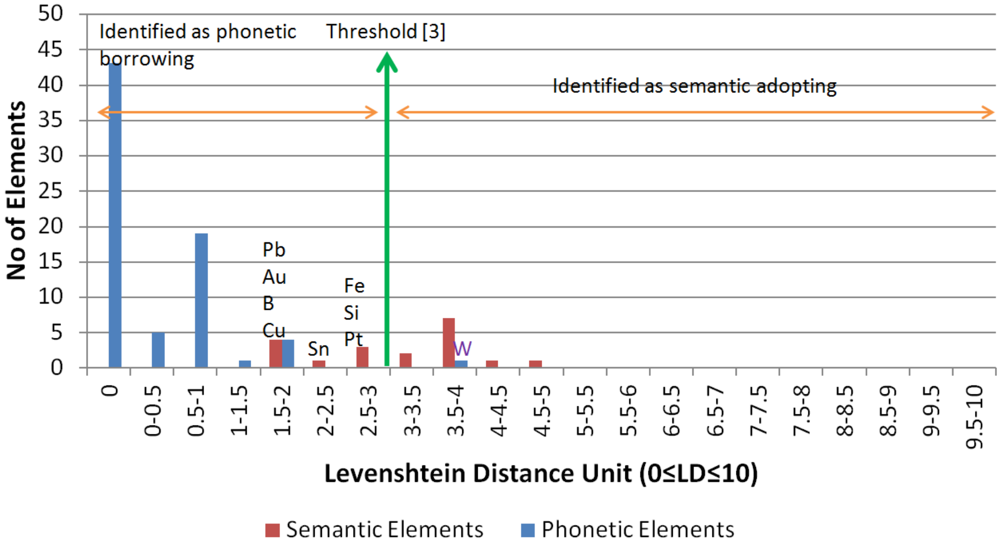


Figure 4.3: Distribution of phonetic and semantic borrowing words in Korean (LD)

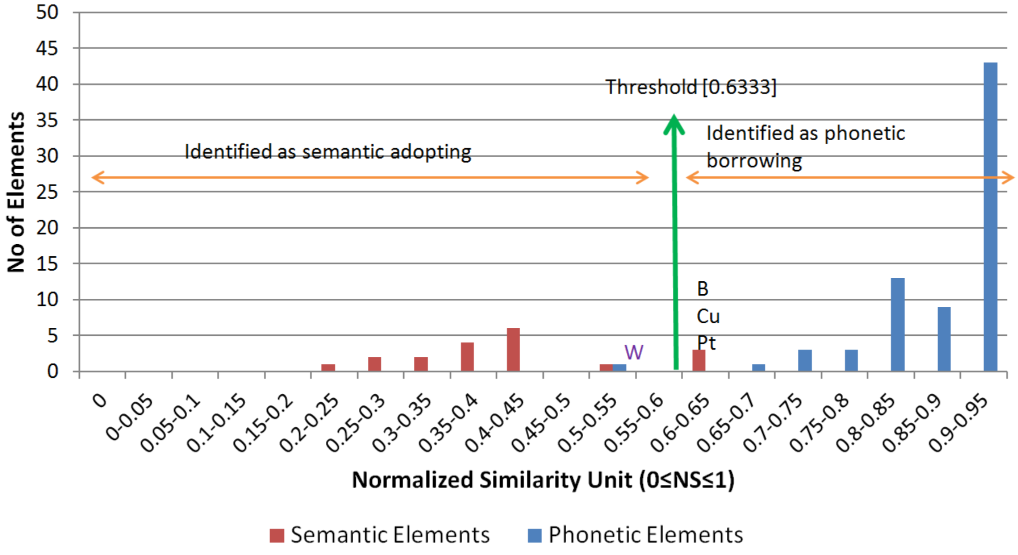


Figure 4.4: Distribution of phonetic and semantic borrowing words in Korean (NS)

English-Malay

Similarly, a threshold of 3 in LD cannot distinguish clearly between phonetic borrowing and semantic adopting word pairs (See in Figure 4.5). Below the threshold in NS, two elements (i.e., potassium and tungsten in English) have fallen into the semantic group; this is because those terms are borrowed from German (i.e., wolfram and kalium in Malay). Also a semantic adopting element word pair "mercury | raksa" (Hg) has fallen into the phonetic group (See in Figure 4.6).

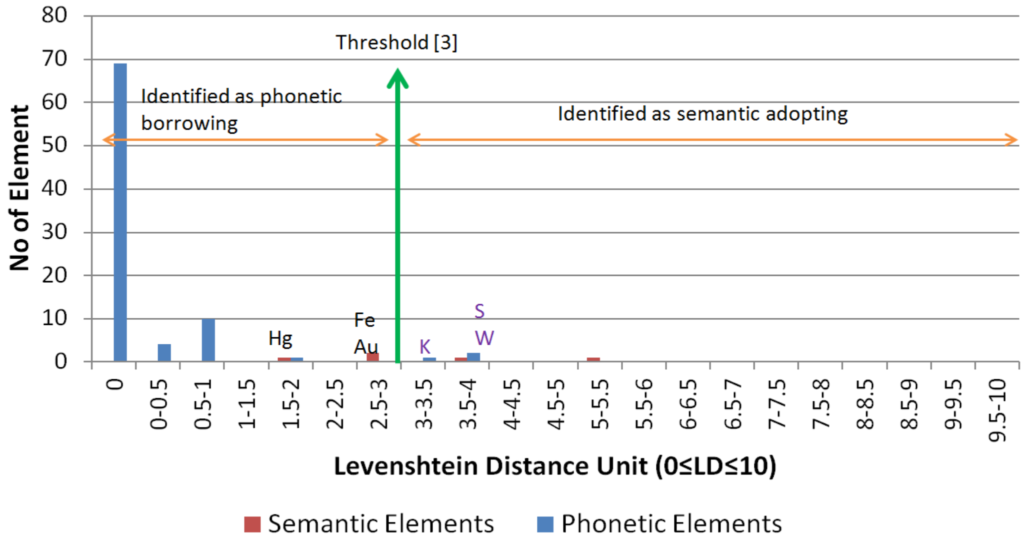


Figure 4.5: Distribution of phonetic and semantic borrowing words in Malay (LD)

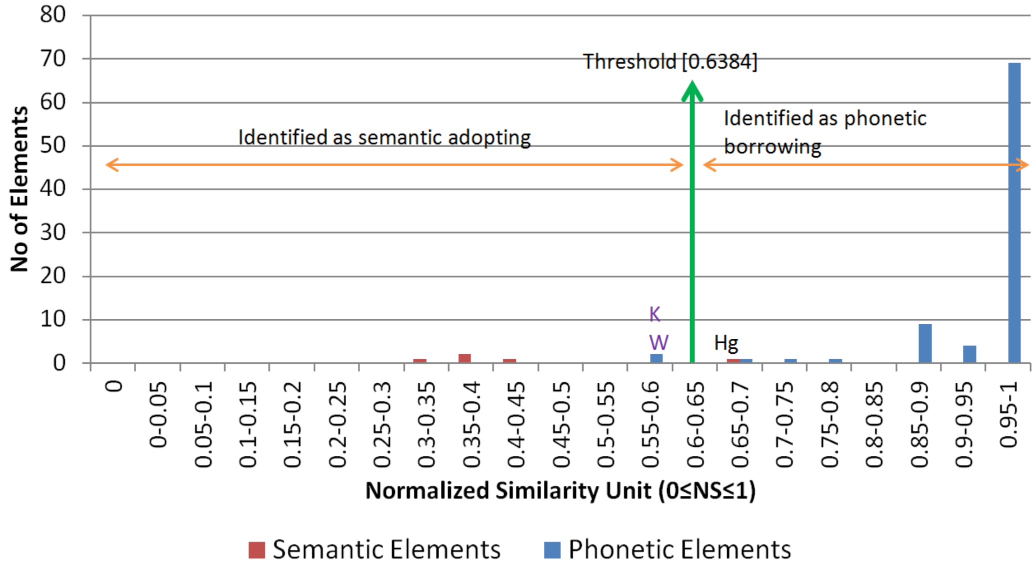


Figure 4.6: Distribution of phonetic and semantic borrowing words in Malay (NS)

English-Thai

Likewise, results in LD cannot distinguish between phonetic and semantic groups of elements, whereas NS does (See in Figure 4.7 and 4.8). Tungsten (W) is borrowed as วุลแฟรม/[Wulferm] from German. Only one element, semantic adopting word pairs "arsenic | สารหนู/[sannuk]" (As) has fallen into phonetic borrowing group.

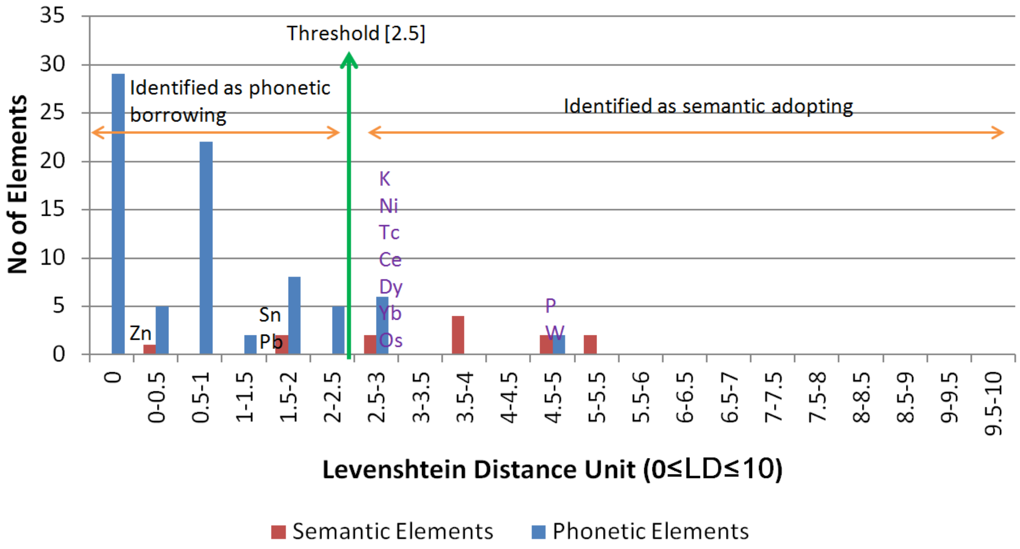


Figure 4.7: Distribution of phonetic and semantic borrowing words in Thai (LD)

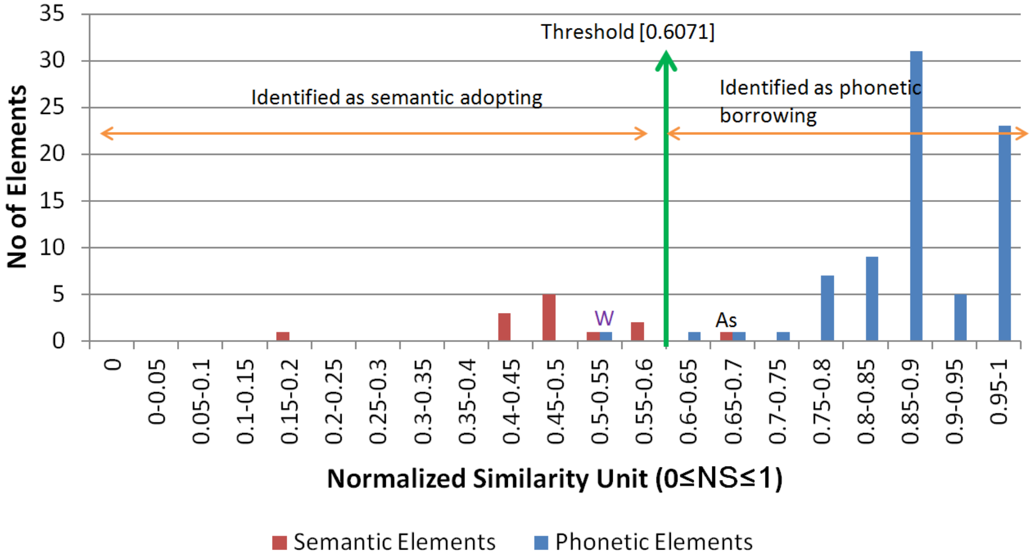


Figure 4.8: Distribution of phonetic and semantic borrowing words in Thai (NS)

English-Myanmar

Figure 4.9 and 4.10 show the results in LD and NS for measurement in Myanmar. LD cannot differentiate between phonetic borrowing and semantic adopting groups. A phonetic borrowing element (i.e., sulfur) was below the threshold of NS and two semantic adopting elements (zinc and copper) were above the threshold.

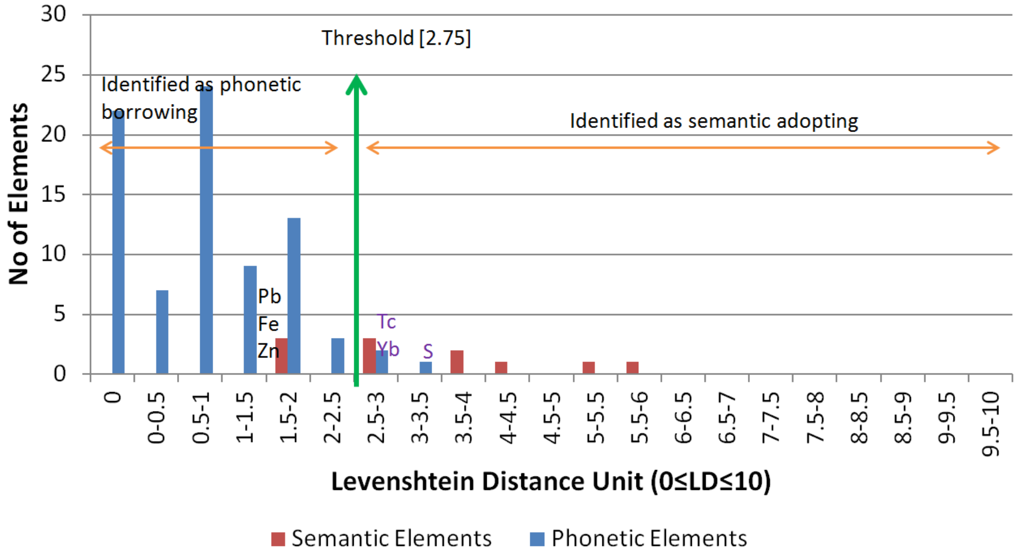


Figure 4.9: Distribution of phonetic and semantic borrowing words in Myanmar(LD)

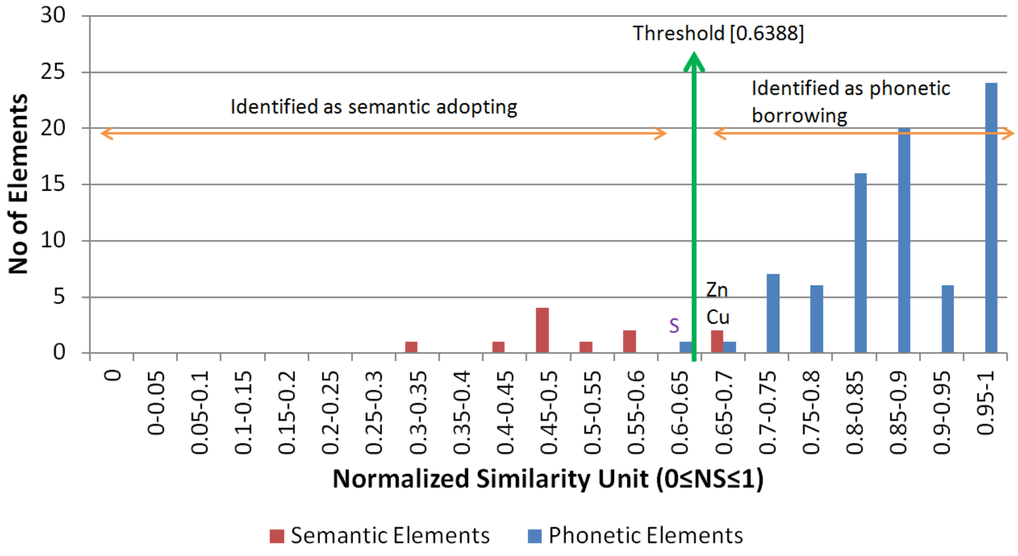


Figure 4.10: Distribution of phonetic and semantic borrowing words in Myanmar (NS)

English-Indonesian

Figure 4.11 and 4.12 show the results and allocation of thresholds in LD and NS for measurement in Indonesian. LD cannot distinguish completely. Two phonetic borrowing element names from German (i.e., potassium (K) and tungsten (W)) are below the threshold and two semantic adopting element names are above the threshold.

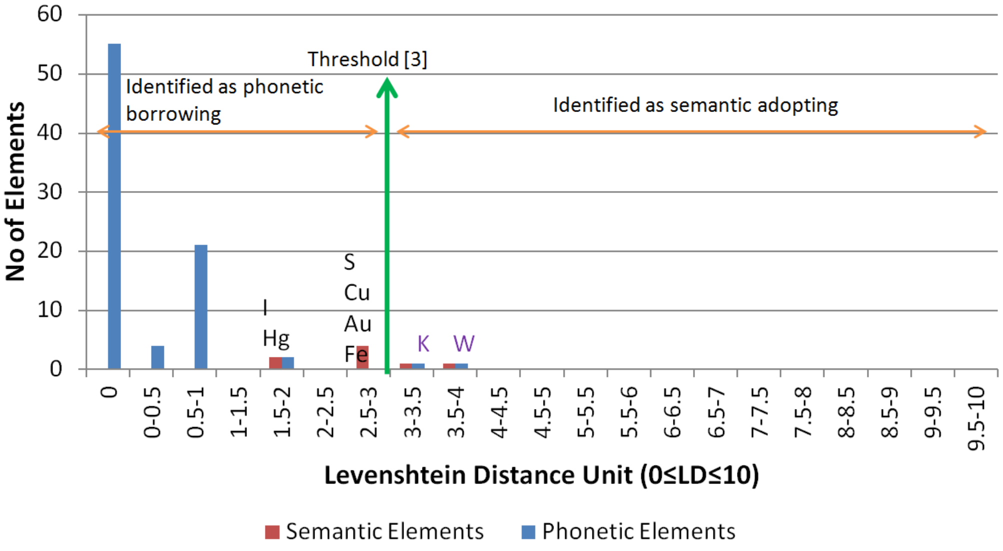


Figure 4.11: Distribution of phonetic and semantic borrowing words in Indonesian (LD)

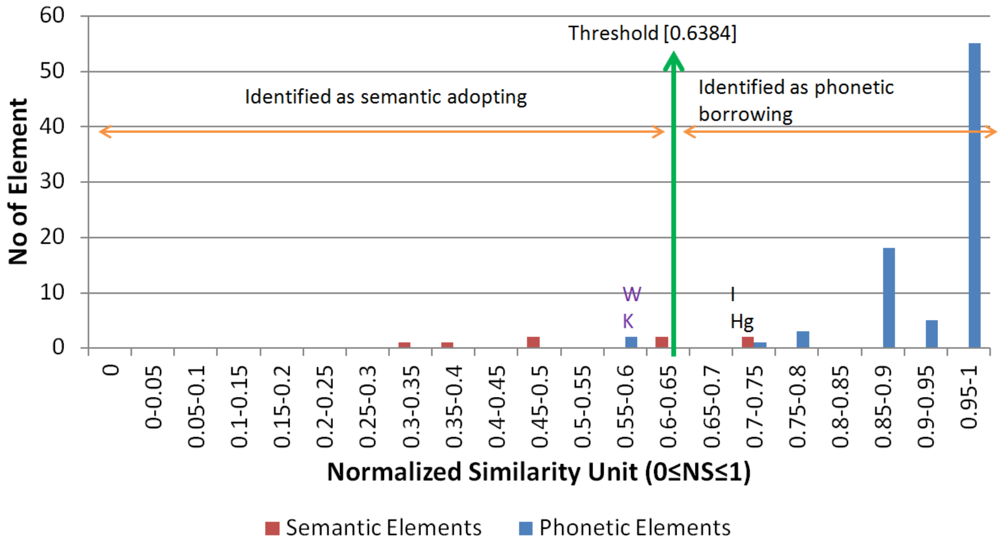


Figure 4.12: Distribution of phonetic and semantic borrowing words in Indonesian (NS)

English-Vietnamese

Results in LD cannot differentiate between phonetic and semantic borrowing word pairs in Vietnamese (Figure 4.13).

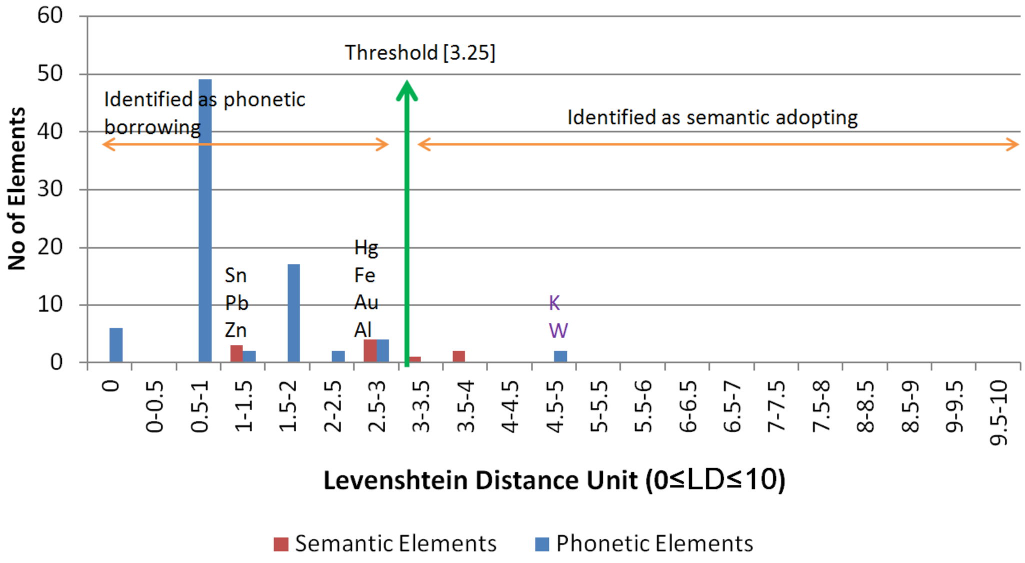


Figure 4.13: Distribution of phonetic and semantic borrowing words in Vietnamese (LD)

The results in NS for Vietnamese are slightly diminished (See in Figure 4.14). The lower similarity score in some word-pairs (e.g., “chlorine” with “clo”) failed since only four of eight characters match in a word in the recipient language (as in the Chinese case; see Figure. 4.3). Also some semantic adopting names of word pairs (e.g., tin (Sn), aluminum (Al), mercury (Hg), and zinc (Zn)) are above the threshold. In fact, Vietnamese was originally written in a Siniform script known as Chữ-nôm or Nôm (similar to the Chinese writing system) and it is still used in many Chinese loanwords [Wii-vun, 2001].

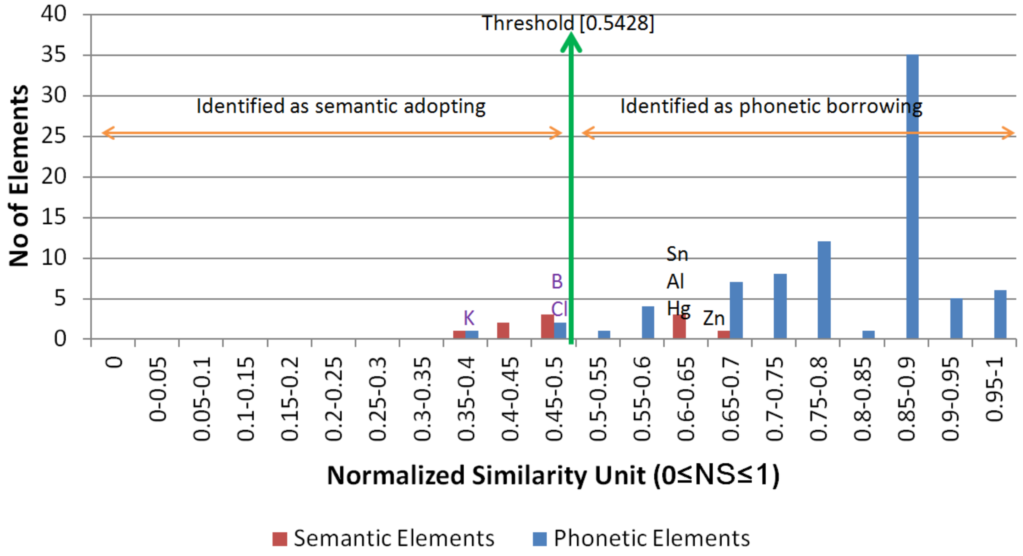


Figure 4.14: Distribution of phonetic and semantic borrowing words in Vietnamese (NS)

English-Chinese

The results in Chinese cannot differentiate between phonetic borrowing word-pairs and semantic adopting word-pairs. The scores of LD and NS cannot identify the allocation of threshold. Since the name of the chemical elements in Chinese are pronounced as an initial sound of the term of the source language (i.e., helium: 氦/[hai], titanium: 钛/[tai]), the length of a word does not longer than four characters (Table. 4.3). In case of borrowing the science and technology terms in Chinese, mostly terms are adopted as semantic adopting words (i.e., a type of loan creation in Semantic adoption) and some other types of creation as semi-phonemic and semi-semantic loanwords [Zhiwei, 2004]. Therefore, most of their distance scores were high and similarity score were low in the results. Therefore, Chinese cannot be measured using this methodology. It the terms are phonetically adopted, it works well. For example, an famous information theory and name "shannon" is borrowed as 香農/[xiang nong], after measuring, LD scores 1 and NS scores 0.875 accurately.

	Romanization	Vowel deletion	Simplify sound	LD	NS
1	iron	$rn	B68	3	0.2500
	tie	T	3		
2	phosphorus	PsPrs	24264	5	0.2857
	lin	Ln	78		
3	bismuth	bsmT	1484	3	0.4000
	bi	B	1		
4	helium	Hlm	978	2	0.5000
	hai	H	9		
5	titanium	Ttnm	3388	3	0.4000
	tai	T	3		
6	tin	tn	38	2	0.3333
	xi	X	5		

Table 4.3: Example of results for Chinese terms

4.3 Evaluation

4.3.1 Allocation of Thresholds

Defining threshold for LD

Retrieving the phonetic borrowing terms in a recipient language from each result of LD and NS, it is necessary to set up an appropriate threshold between the groups of phonetic and semantic borrowing word-pairs. Following this, allocation of a threshold in LD is calculated from the mean (average) value between a minimum distance of semantic borrowing word-pair and a maximum distance of phonetic borrowing word-pair. Threshold in LD is defined as:

$$T_{LD} = \frac{S_{min} + P_{max}}{2}$$

where

T_{LD} = Proposed threshold for LD

S_{min} = A minimum distance value of semantic adopting word-pair

P_{max} = A maximum distance value of phonetic borrowing word-pair

Defining threshold for NS

Allocation of threshold in NS is calculated from the average of two values, a maximum similarity of semantic adopting word-pair and a minimum similarity of phonetic borrowing word-pair. Threshold in NS is defined as:

$$T_{NS} = \frac{S_{max} + P_{min}}{2}$$

where

T_{NS} = Proposed threshold for NS

S_{max} = A maximum similarity value of semantic adopting word-pair

P_{min} = A minimum similarity value of phonetic borrowing word-pair

4.3.2 Retrieval Standard

Defining precision and recall

Precision and recall are standard evaluation strategy for information retrieval. They are also used extensively in the information retrieval literature. In particular this evaluation strategy assesses quantitatively both the quality of the overall answer set (results) and the breadth of the retrieval algorithm [Ricardo et al., 1999]. Thus, the quality of measuring results and comprehensiveness of retrieval results of cross-language phonetic similarity metric are determined by using recall and precision evaluation. Essentially, recall and precision are binary measurements, where an object is either relevant or non-relevant (true or false) [Ibrahiem et al., 2005]. In the case of the CLPSM retrieval process, the binary measurement can be determined by the following possible states (Table 4.4).

In a resulting data set of LD/NS		**number of phonetically similar word-pairs identified by a threshold**	
actual number of phonetically similar word-pairs in the data		Retrieved	not retrieved
	relevant	true positive (tp)	false negative (fn)
	not relevant	false positive (fp)	true negative (tn)

Table 4.4: States of the binary classification in a measurement

The correctness of measuring phonetic borrowing word-pairs can be evaluated by calculating the number of correctly retrieved phonetic borrowing word-pairs (true positives), the number of correctly retrieved word-pairs that do not belong to the phonetic borrowing word group (true negatives), and word-pairs that either were incorrectly assigned to the phonetic borrowing word group (false positives) or that were not retrieved as phonetic borrowing word-pairs (false negatives) by assigning a threshold. These four states constitute a matrix shown in Table 4.4 for the case of binary classification.

Calculating precision

To obtain precision, we calculate the number of correctly identified phonetic borrowing word-pairs divided by the number of word-pairs retrieved by the CLPSM as positive. Then precision (P) for each language measurement can be obtained as follow [Marina et al., 2009]:

$$P = \frac{tp}{tp + fp}$$

where

P = the quality of measuring phonetic similarity resulting from CLS

tp = the number of relevant phonetic borrowing word-pairs in retrieving data set

fp = the number of irrelevant semantic borrowing word-pairs in retrieving data by allocating of threshold.

Calculating recall

Recall measures the number of correctly identified phonetic borrowing word-pairs divided by the number of actual phonetic borrowing word-pairs (positive) in the data set. We can obtain recall (R) of each language measurement as follow [Marina et al., 2009]:

$$R = \frac{tp}{tp + fn}$$

where

R = the comprehensiveness of retrieving phonetic borrowing word-pairs resulting from CLS

tp = the number of relevant phonetic borrowing word-pairs in retrieving data set

fn = the number of relevant phonetic borrowing word-pairs out of retrieving data by allocating of threshold.

Calculating average precision and recall

The average precision for a retrieval of phonetic borrowing word-pairs in the target languages is given by

$$P_{av} = \frac{1}{n}\sum_{i=1}^{n} P_i$$

where the number of target language measurement is i = 1, 2, ..., n.

Similarly, average recall is given by

$$R_{av} = \frac{1}{n}\sum_{i=1}^{n} R_i$$

Calculating F-measure

The F-measure is a measure of a test’s accuracy, and it is basically the harmonic mean of precision and recall. The harmonic mean is more intuitive than the arithmetic mean (average) when computing a mean of ratios [Sasaki, 2007]. In addition, to get a better picture of the performance of CLPSM, the results are evaluated by the F-measure, which is calculated using the following equation:

$$F = \frac{2PR}{(P + R)}$$

where:

P = Precision value of a measurment using formula

R = Recall value of a measurment using formula

4.3.3 Retrieval Accuracy

Performance Analysis of LD & NS Metrics (CLSG ver.1)

Table 4.5 describe the ratio of precision, recall, and F-measure by LD and NS metrics. The results show that NS achieves high performance in all languages (i.e., using CLSG ver.1).

	LD			NS		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Japanese	0.8861	0.9722	0.9272	0.9726	0.9595	0.9660
Korean	0.9000	0.9863	0.9412	0.9600	0.9863	0.9730
Malay	0.9655	0.9655	0.9655	0.9885	0.9773	0.9829
Myanmar	0.9398	0.9630	0.9512	0.9756	0.9877	0.9816
Thai	0.9595	0.8875	0.9221	0.9747	0.9625	0.9686
Indonesian	0.9318	0.9762	0.9535	0.9762	0.9762	0.9762
Vietnamese	0.9195	0.9756	0.9467	0.9518	0.9634	0.9576
Chinese	0.7818	0.5890	0.6719	0.8065	0.3906	0.5263

Table 4.5: Performance of LD and NS by using CLSG ver.1

Performance Analysis of LD & NS Metrics (CLSG ver.2)

Table 4.6 describe the ratio of precision, recall, and F-measure by LD and NS metrics. The results similarly show that NS achieves better performance than LD in all languages (i.e., using CLSG ver.2).

	LD			NS		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Japanese	0.8875	0.9861	0.9342	0.9733	0.9865	0.9799
Korean	0.9000	0.9863	0.9412	0.9730	0.9863	0.9796
Malay	0.9663	0.9885	0.9773	0.9885	0.9885	0.9885
Myanmar	0.9500	0.9383	0.9441	0.9759	1.0000	0.9878
Thai	0.9157	0.9620	0.9383	0.9747	0.9747	0.9747
Indonesian	0.9765	0.9881	0.9822	0.9765	0.9881	0.9822
Vietnamese	0.9091	0.9756	0.9412	0.9759	0.9759	0.9759
Chinese	0.8333	0.4225	0.5607	0.7097	0.3099	0.4314

Table 4.6: Performance of LD and NS by using CLSG ver.2

Chapter: 5

Experiment of Stochastic Edit Distance Measures

5.1 DATA

In order to measure in learning approach, the experiments used a Myanmar-English bilingual training corpus consisting of 1,729 word pairs from the titles of Wikipedia articles linked by inter-language links[7]. The corpus additionally included 11,462 word pairs extracted from a Multilingual Terminology Dictionary [MTD, 2009]. From this complete data set, two derivative training sets are constructed. The first consisted of 3,100 single word pairs that it is called 'clean data'; this type of data is typical of the kind of data one might encounter in a shared task on transliteration mining such as the NEWS workshop [Kumaran, 2010]. The second consisted of 14,891 multiple word pairs in which a pair may consist of more than one word in English; this data is called 'realistic data' as for this task, where resources are low, we believe it is necessary to mine data in a less than a ideal format.

5.1.1 Data Annotation

The testing data were manually annotated as transliteration (phonetic borrowing) and non-transliteration (semantic adopting) pairs by a bilingual human annotator. Of the 3,100 bilingual pairs sampled, 1,291 pairs were transliterations and 1,809 pairs were non-transliterations.

[7] http://dumps.wikimedia.org/mywiki/20120824/

5.1.2 Pre-processing

5.1.2.1 Data Cleaning

The source data (Myanmar) contained a lot of noise such as spelling errors and issues with mixed encoding. Moreover, there are many technical problems in Myanmar Unicode characters, for example, the case of U+200C (Zero-width non-joiner) is a non-printing character used in the writing system, that needs to be eliminated. Some other usages like "။", "၊", white spaces, and the Unicode born were also eliminated.

5.1.2.2 Segmentation

The Myanmar language is syllable-timed; therefore a preprocessing task was required to break it into syllables (syllable segmentation). As an example, according to the Unicode encoding standard, the syllable ('ကျောင်း') is encoded as follows:

က + ျ + ေ + ာ + င + ် + း = ကျောင်း

The syllable ('ကျောင်း') basically consists of an initial consonant ('က') with optional medials ('ျ'), dependent vowels ('ေ','ာ'), dependent signs ('း'), and more than one consonant may appear together with the devoweliser (the killer character Asat '်').

A segmentation process consisting of 3 rules is necessary to segment the Myanmar grapheme sequences into sequence of syllables (this experiment used a syllable breaking tool that is developed by Ye Kyaw Thu, NICT (2012)).

Rule-1: Break in front of consonant, independent vowel, number and symbol characters" and is the first step for syllable breaking. But there is an exception for Kinzi ('င်္'), i.e., a combination of a conjunct (U+1004) with Myanmar letter ('င') preceding the consonant. For example, the Myanmar word "ကွန်ပျူတာ" can be segmented into syllables as "|ကွန်|ပျူ|တာ|".

Rule-2: Remove the breaking point in front of a subscript consonant (i.e., PadSint). For example, "မိတ္တူ" breaks as |မိ|တ္|တူ| and replaces with Asat |တ်| and finally combines with a front segmented letter "|မိတ်|တူ|".

Rule-3: Break in front of Kinzi character ('င်္'). For example, "အင်္ဂလိပ်" breaks as "|အင်|ဂ|လိပ်|".

5.1.2.3 Romanization

Furthermore, the Myanmar script is non-alphabetic, and therefore two different romanization schemes are applied to convert Myanmar into the Latin alphabet in order to study the effect of the using differing romanization schemes on mining performance. This experiment used two romanization systems: the Myanmar Language Commission transcription system [MLCTS, 1980] and the University of Foreign Language (UFL) pronunciation system [UFL, 2005] that is an extended version of MLCTS, but is significantly different in character. Table 5.1 shows three examples of the two romanization schemes in use. Romanized words in UFL appear to be more similar to the spelling of the word in English.

English	Myanmar	MLCTS	UFL
vitamin	ဗီတာမင်	bitamang	bi.ta.min
motorcar	မော်တော်ကား	mautauka:	mo.to.ka
platinum	ပလက်တီနမ်	paktinam	ple'ti.nam

Table 5.1: Romanization schemes

5.1.2.4 Phonetic Coding

Training data in both languages were mapped into the phonetic code strings using our technique (CLSG ver.2) for phonetic similarity grouping (i.e., referred to the

Table3.8 of Section 3). Table 5.2 shows three word pairs transcribed into phonetic code sequences. Note that although the spellings of each of the romanized forms differ significantly, the phonetic coding sequences are all identical.

	English	MLCTS	UFL
Romanization	vitamin	bitamang	bitamin
Phonetic coding	1030A0A	1030A0A	1030A0A

Table 5.2: Romanization & Phonetic Coding

5.2 RESULTS

5.2.1 EM Training

The EM model was trained on the 3,100 single words of the clean data set that were both romanized and phonetically coded. In order to eliminate the noise model, the EMn model is trained by fixing the parameter λ (the probability of non-transliteration) at zero. The EM model could not learn a good model from both MLCTS and UFL training data, both of which gave poor mining performance (See Figures 5.1 and 5.2).

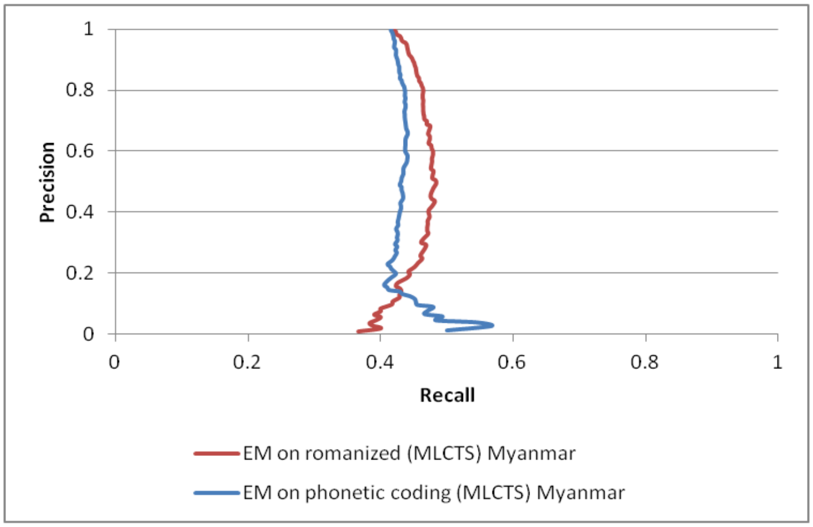


Figure 5.1: Comparing MLCTS romanization to phonetic coding (CLSG ver.2) with the EM model

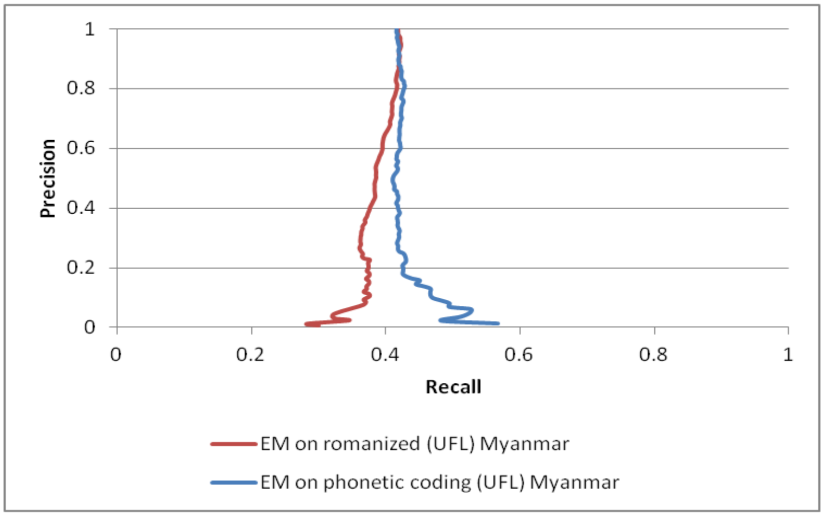


Figure 5.2: Comparing UFL romanization to phonetic coding (CLSG ver.2) with the EM model

5.2.2 EMn Training

The EMn model showed improved performance on phonetically coded data (CLSG ver.2) over the models trained on romanized data (see Figure 5.3). The prior probability of λ was initially set to a value between 0 and 1. We found that performance was dependent on the romanization scheme used, and that the results in UFL were better than MLCTS data (See Figures 5.3 and 5.4).

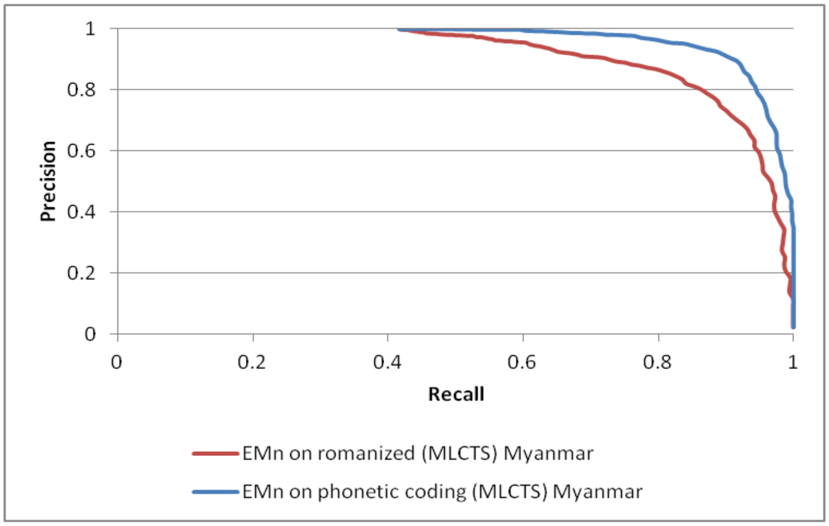


Figure 5.3: Comparing MLCTS romanization to phonetic coding (CLSG ver.2) with the EMn model

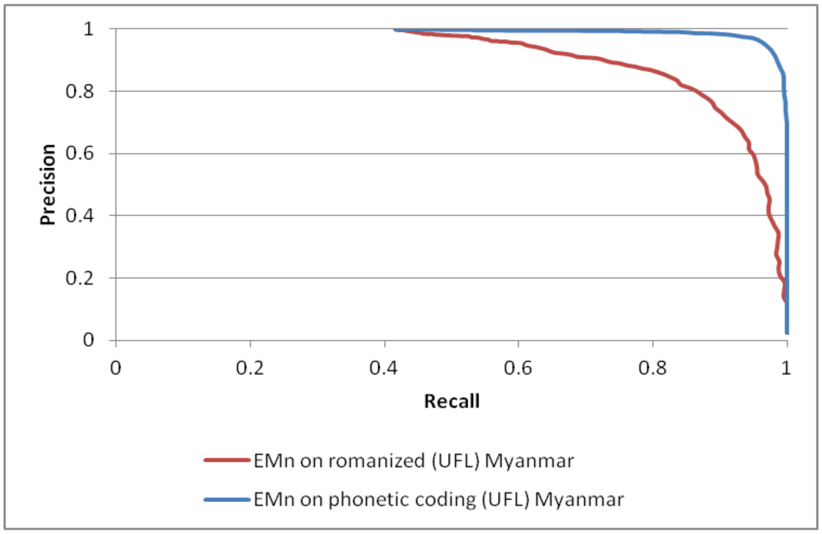


Figure 5.4: Comparing UFL romanization to phonetic coding (CLSG ver2) with the EMn model

5.2.3 BAYESn Training (Clean Data)

The BAYESn model learned models with comparable performance to the EMn model on the clean data set for both romanized and phonetically coded data. Again it is found that using the phonetic coding (CLSG ver.2) gave a much better performance. The results are shown in Figures 5.5 and 5.6.

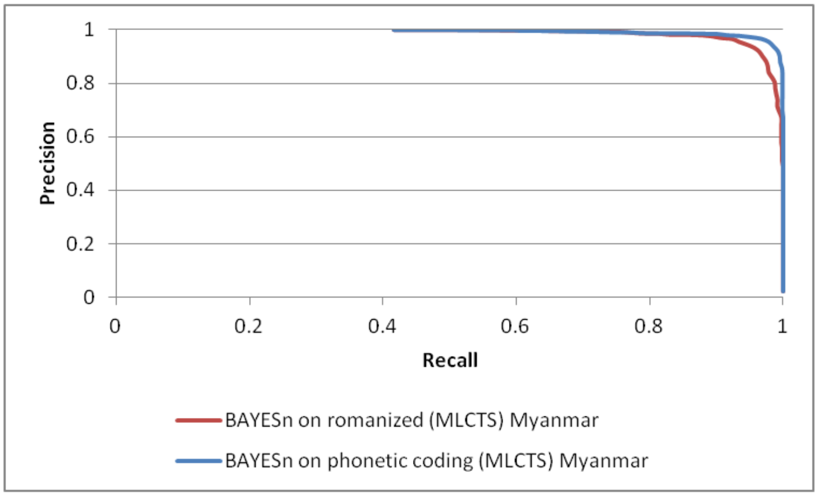


Figure 5.5: Comparing MLCTS romanization to phonetic coding (CLSG ver.2) with the BAYESn model

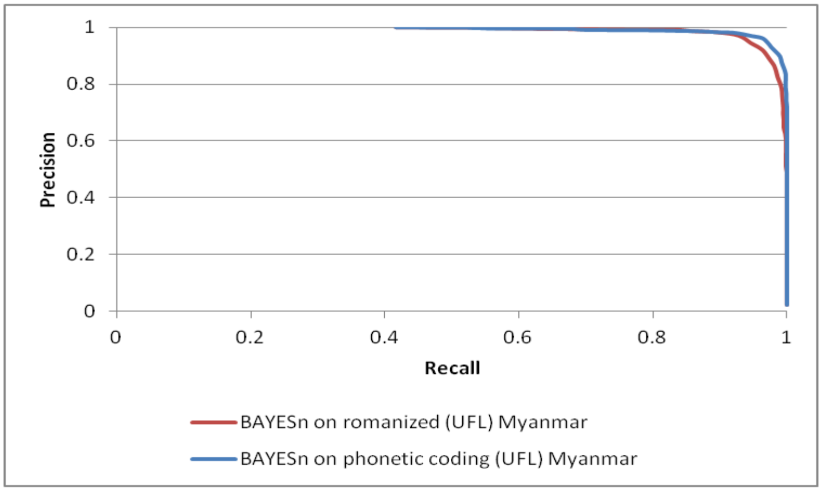


Figure 5.6: Comparing UFL romanization to phonetic coding (CLSG ver.2) with the BAYESn model

5.2.4 BAYESn Training (Realistic Data)

The results from the BAYESn model trained on the 14,891-sample realistic dataset show similar characteristics to the results from clean data: a good-performing model is learned in both cases, but there is higher performance from the phonetically coded data with respect to the romanized data. The results are shown in Figure 5.7.

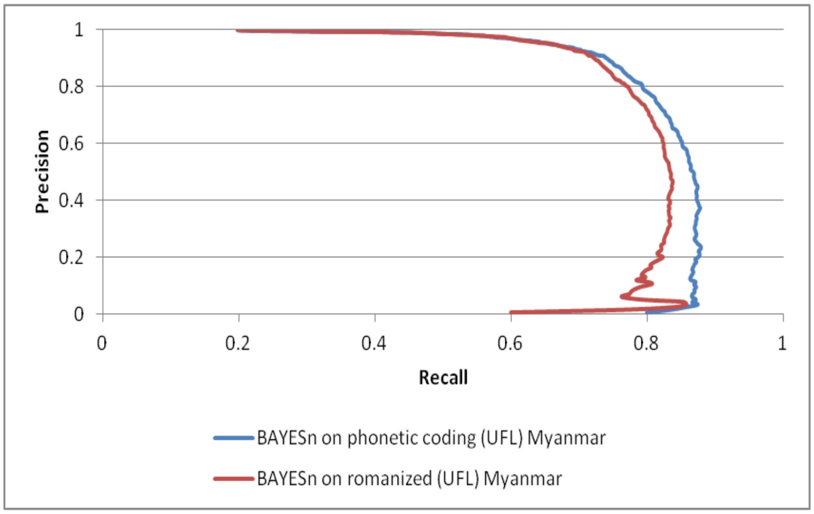


Figure 5.7: Comparing UFL romanization to phonetic coding (CLSG ver.2) with the BAYESn model using realistic data

5.3 Evaluation

5.3.1 Evaluation Criteria

In order to evaluate the performance of all models, the standard evaluation metrics are used to describe as follows: precision, recall, and f-score. Where TP is the number of correct pairs (transliteration pairs) that were labeled as correct (true positive), FP is the number of incorrect pairs (non-transliteration pairs) that were labeled as correct (false positive), and FN is the number of correct pairs (transliteration pairs) that were labeled as incorrect (false negative). When mining real data, the data may not necessarily be mined at the optical F-score; an appropriate trade-off between precision and recall may need to be selected to fit the specific application. For this reason the results are presented in the form of graphs of the complete precision/recall curves (shown in Figures 5.1-5.6).

5.3.2 Evaluation of Training Result

A summary of the results are provided at optimal F-score in Tables 5.3 and 5.4 for both UFL and MLCTS Romanization.

	Data type	Precision	Recall	F-measure
EM	Romanization	0.4241	0.8574	0.5675
	Phonetic coding	0.4236	0.8466	0.5647
EMn	Romanization	0.9601	0.9147	0.9369
	Phonetic coding	0.9825	0.9132	0.9466
BAYESn	Romanization	0.9642	0.9186	0.9408
	Phonetic coding	0.9785	0.955	0.9666

Table 5.3: Mining performance using UFL romanization

	Data type	Precision	Recall	F-measure
EM	Romanization	0.4637	0.7327	0.5680
	Phonetic coding	0.4364	0.7606	0.5546
EMn	Romanization	0.4419	0.9140	0.5958
	Phonetic coding	0.9166	0.8946	0.9055
BAYESn	Romanization	0.9555	0.9326	0.9439
	Phonetic coding	0.9706	0.9473	0.9588

Table 5.4: Mining performance using MLCTS romanization

Despite learning in the 3,100 pair clean data set, EM model did not learn accurately and precisions of both UFL and MLCTS romanization data are very low. However, EMn model with phonetic coding in MLCTS raised up the precision, but using UFL gave high precision in both romanization and phonetic coding data. BAYESn learned well in all types of data and gave high precision, recall, and F-measure.

	Data type	Precision	Recall
BAYESn	Romanization	0.79	0.72
	Phonetic coding	0.81	0.74

Table 5.5: The performance of BAYESn in multiple word data

Table 5.5 presents the results of BAYESn experiment using UFL Romanization and many-to-many alignment. Though cross-language sound grouping knowledge helps to learn to obtain better precision and recall on this dataset; the precision will discuss this in detail in the next chapter.

5.3.3 Learning Results in Asian languages

Table 5.6 shows the precision, recall and F-measure of EMn learning in 92 element names of word pairs in each Asian language. In this learning method, threshold is calculated as an average of all probability score of word pairs as below:

$$T_{EMn} = \bar{P}_n$$

where T_{EMn} is a threshold value and $\bar{P}_n$ is an average probability of n number of bilingual word pairs.

If the scores of probability of word pair are equal to or greater than T_{EMn} suggests phonetic borrowing words and below the threshold indicates semantic borrowing words.

Table 5.6 shows the retrieval results obtained from the EMn algorithm that was tested in (92) element names in eight Asian languages.

Source: English	Threshold	Origin of terms		Identified results	
	(T_{EMn})	Phonetic	Else	Phonetic	Else
Japanese	0.7639	72	20	71	21
Korean	0.7851	73	19	72	20
Malay	0.9422	87	5	86	6
Thai	0.8450	79	13	79	13
Myanmar	0.8844	81	11	81	11
Indonesian	0.9102	84	8	83	9
Vietnamese	0.8428	82	10	78	14
Chinese	0.3786	73	19	40	52

Table 5.6: Thresholds and retrieved terms in EMn results

Learning in English-Japanese (CLSG ver.2)

Figure 5.8 shows the results in English-Japanese, a proposed threshold differentiates between phonetic borrowing and semantic adopting word pairs accurately. However, sodium| natoriumu (Na) and potassium| kariumu (K) word pairs are below the threshold, these element names are derived from German. Only iodine| youso (I) has fallen into phonetic borrowing group.

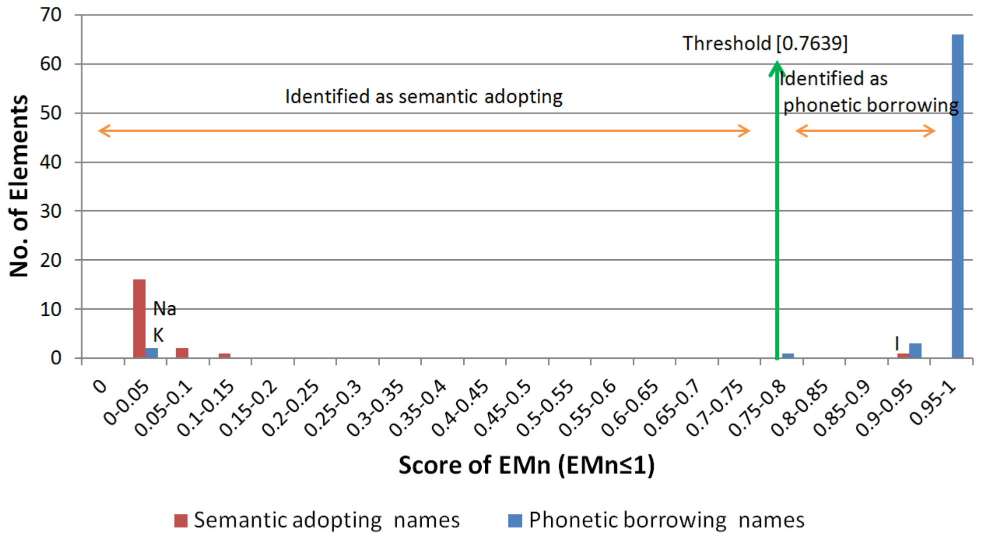


Figure 5.8: Distribution of phonetic and semantic borrowed word-pairs in Japanese (EMn) by using CLSG ver.2

Learning in English-Vietnamese by NS (CLSG ver.2)

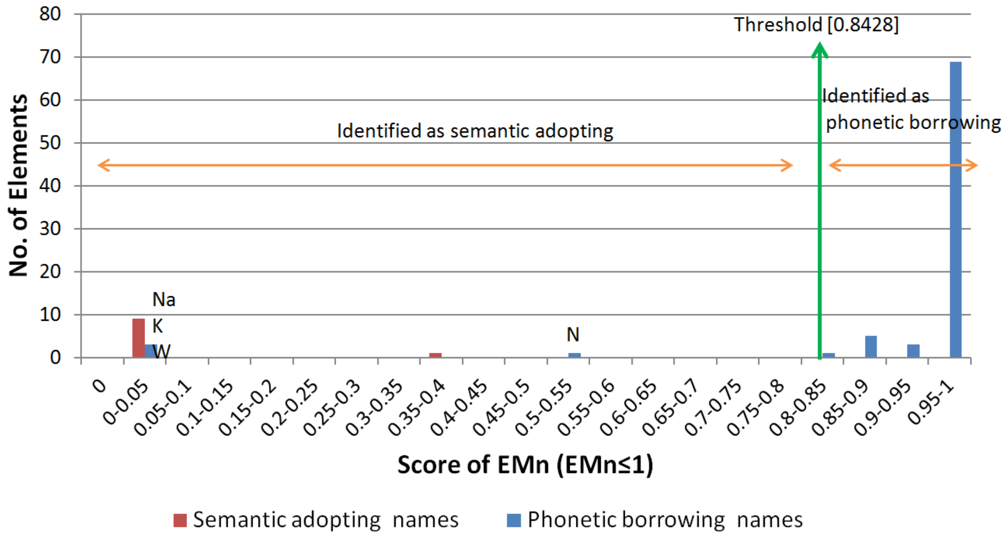


Figure 5.9: Distribution of phonetic and semantic borrowed word-pairs in Vietnamese (EMn) by using CLSG ver.2

Figure 5.9 shows the distribution of phonetic borrowing and semantic adopting word pairs in Vietnamese. Among the elements, "sodium|natri" (Na), "potassium|kali"

(K), "tungsten|volfram" (W) are derived from German and only "nitrogen|nito" are below the threshold (0.8428).

Performance of EMn in Asian languages (CLSG ver.2)

Tables 5.7 reports the performance results for eight Asian languages. As shown in this table, the recall for English-Chinese was dismal and suggests problems in dataset (i.e., refer to section 6.1.1).

EMn	Precision	Recall	F-measure
Japanese	0.9859	0.9722	0.9790
Korean	1.0000	0.9863	0.9931
Malay	1.0000	0.9885	0.9942
Myanmar	1.0000	1.0000	1.0000
Thai	0.9873	0.9873	0.9873
Indonesian	0.9880	0.9762	0.9820
Vietnamese	1.0000	0.9512	0.9750
Chinese	0.9250	0.5068	0.6549

Table 5.7: The performance of EMn in 92 chemical element name of word pairs in each Asian language

Chapter : 6

Discussion

6.1 MANUAL DESIGNING APPROACH MEASURES

6.1.1 Extension of Soundex (CLSG) for Asian Languages

Reasoning of Soundex extension

Soundexing is the concept of indexing information by how it sounds rather than its spelling. Early Soundex was developed based on English phoneme by Russell [Russell et al., 1922], in his patent he assumed that "*there are certain sounds which form the nucleus of the English language, and those sounds are inadequately represented merely by the letters of the alphabet, as one sound may sometimes be represented by more than one letter or combination of letters, and one letter or combination of letters may represent two or more sounds*". Based on this, similar sounding words are categorized with the same or similar codes. However, Soundex does not cover for all languages, but it is the most useful system for phonetic matching and it is adapted to use in other languages by modifying the particular characteristics of each of their languages. We found some European languages adopted incorporating some adjustment into Russell Soundex and American Soundex algorithm, for example, Daitch Mokotoff Soundex [Steuart, 1994] is a refinement of the Soundex algorithm for Slavic and Yiddish languages.

Features of extension Soundex

For the purpose of this research, the CLSG is extended based on the Soundex system, especially for Asian languages. Basically, a speech sound of consonant evolves with three phonetic features: place of articulation, manner of articulation, and voicing.

This CLSG is based on the place and manner of articulation (refer to section 3.1). The key features of extension of Soundex (CLSG) are summarized as follows:

- Allow more or less than four characters strings to measure
- Extend from six articulatory features (labials, gutturals & sibilants, dental, long liquid, nasal, short liquid) to eleven (primary ver.1: labial-plosive, labial-fricative, dental-plosive, dental-fricative, velar-plosive, velar-fricative, r-sound, l-sound, nasals, approximants, and initial zero consonant) and thirteen (secondary ver. 2: labial-plosive, labial-fricative, dental-plosive, dental-fricative, alveolar-plosive, alveolar-fricative, velar-plosive, velar-fricative, r-sound, l-sound, nasals, approximants, and initial zero consonant) articulatory features
    - /h/ and /w, y/ consonants are grouped into velar-fricative and approximants
    - Include two letter form of initial consonants: /dh/, /kh/, /ng/, /ny/, /ph/, /hm/, /hn/, etc.
- Adding IPA notation for various languages
- Allow modification or extension to particular language

Characteristic of tested loanwords in Asian languages

In addition, understanding the characteristics of loanwords in a language is very import in this research. A borrowing word can be classified into many types, for examples, direct loan, loan ship, loan creation, loan blend, etc. [Halvor, 2005]. This research is based on two fundamental types: semantic adopting and phonetic borrowing, developed the metrics to identify phonetic borrowing words (i.e., a kind of direct loan). The features of the borrowed word can be a single word, a compound, or a phrasal expression. Many words are shared in different languages a result of linguistic borrowing in various semantic fields: agriculture, food and drink, entertainment, animals, science and technology, economic, political, etc. Most of the Asian languages import the words in phonetic borrowing from European languages especially for science and engineering terminologies.

Analysis of phonetic borrowing in 92 elements

Among the 92 elements in each language, over three-quarters of element names are loanwords. The percentage of phonetic borrowing words in each language is shown in Table 6.1.

Language	**Percentage of phonetic borrowing words**
Japanese	78%
Korean	79%
Malay	95%
Thai	85%
Myanmar	88%
Indonesian	91%
Vietnamese	89%

Table 6.1: Phonetic borrowing rate in Asian languages

Semantic borrowing in Asian languages

It seems likely that there is a high rate of phonetic borrowing in chemical terminology in Asian languages. Only elements found early borrow semantically, for example, 'gold' in Japanese as '金' /kin/, in Malay and Indonesian as 'emas' /emas/, in Myanmar as 'ရွှေ' /rewh/.

Chinese terminology adoption method

However, the type of loanwords used in Chinese need to be considered as an exception, elements are mostly adopted as semantic borrowing words and new type of loanwords have been created (i.e., a type of loan creation in Semantic adoption). In the creation of a term using phonetic borrowing, a term is created in Chinese characters which either fit graphic classifiers or is a selection of existing characters that is based on the phonetic rendition of the source term [Martin et al., 2009]. In other words, they took from the prefix of pronunciation of the original word or person name that was found in those elements or, then assigned a Chinese character that is close to the prefix or suffix sound of the source language. For example, Radon (86) in Chinese is 'dong' (氡) that came from part of the scientist's name (Friedrich Ernst Dorn) who found it in 1900 [Zhiwei, 2004]. The first part of the Chinese character represents the nature of element (i.e., air, water, soil, etc.) and the second part of Chinese character represents the prefix or suffix of the element name (See an example in Table 6.2).

Element	**New character**	**Phonetic Form**	**Graphic Classifier**	**Definition of Graphic Classifier**
Silicon	矽	xi	石	Stone

Table 6.2: An example of new character creation in Chinese elements

Some other elements in semantic borrowing are from their ancient/local name (e.g. Copper 'tong' 銅 and silver 'yin' 銀). Another example is a word with the same pronunciation but in different Chinese characters (e.g., Aluminium (lu 鋁) and Chlorine (lu 氯)). Likewise, some elements in Vietnamese are very similar to Chinese. For example, ‘chlorine’ as ‘clo’, and ‘boron’ as ‘bo’. This is because Vietnamese was originally written in a Siniform script known as Chữ-nôm or Nôm (similar to Chinese writing system) and it is still used in many Chinese loanwords [Wii-vun, 2001].

Most frequent used adoption form

The most frequently found case of loanword adaption is sound changed in the recipient language. As an example, the voiced labio-dental fricative /v/ in Thai language does not have, it is replaced with /w/ as a closer sound [Kenstowicz, 2006]. Although the various language scripts are written in Latin scripts, the spelling does not always correspond directly to the pronunciation. Each language is required to simplify depending on its phonemic nature (referred to section 3.2.3).

Coverage of language family by CLSG

Our experimental languages are descended from various types of language family such as Japanese and Korean from Language Isolate, Chinese and Myanmar from Sino-Tibetan, Indonesian and Malay from Austronesian, Thai from Tai-Kadai, and Vietnamese from Austro-Asiatic. Therefore, CLSG can be covered to be applicable to other languages among these five types of language family.

6.1.2 Classical Edit Distance Measures

Retrieving phonetic borrowing words from a bilingual dataset, our methodology needed to deal with three forms of loanwords (i.e., a single word, a compound, or a phrasal expression) and its particular features in each language. Using classical edit distance in manual designing approach performed accurately to measure loanwords in single word form. The standard dynamic-programming solution is used to compute the phonetic similarity measure.

Romanization

In pre-processing, non-Latin scripts are converted into romanized alphabets using each language specific romanization standard such as "99-SHIKI" Japanese Romanization [社団法人日本ローマ字会, 1999-7], Revised Romanization of Korean [Korean language, 2000], Pinyin Romanization of Mandarin Chinese[8] (Pinyin), the Myanmar Language Commission Transcription System (MLCTS) [Naing, 2006], and the Royal Thai General System of Transcription [Royal Thai, 2002].

Simplifying sound process

In this process (referred in section 3.2.3), the first step proceeds to simplify based on a concept of sound change between source and recipient languages that is language dependent. Then following process, the baseline sound-class is carried out to simplify with phonetic coding using CLSG (i.e., extension of Soundex) independently from both source and recipient languages.

Levenshtein Distance

In this baseline measure (LD), calculation of edit cost (weight) in deletion and insertion is set to 1, and edit cost for substitution operation is dependent on the knowledge of CLSG, if it is between in the same group, variable weight (ώ) is set to 0.5, and otherwise is set to 1. But the result in LD could not differentiate between semantic and phonetic borrowing words in all Asian languages (referred to in section 4.2.2).

Normalization

In order to eliminate the length of string effect, normalized similarity (NS) is calculated as the baseline measure value of Levenshtein distance (LD) divided by the total length of the two strings (i.e., source and recipient strings) and it is a metric valued

[8] Romanization of Mandarin Chinese (Pinyin): http://www.mandarintools.com/

in [0,1] under the condition of cross-language sound grouping. The results in NS managed successfully to retrieve phonetic borrowing word pairs in Asian languages, except in Chinese (referred to in the previous section). In addition, it is found that if either the particular string in the source or recipient language is less than four characters. For example, even though it is a phonetic borrowing word pair, the measure between 'oxygen' and 'oxy' in Vietnamese yielded below the threshold (0.562) in a NS score of 0.5. The normalization helps the length of string effect in this task, but the fact that it was needed to take into account the word lengths.

Performance Analysis of CLPSM using CLSG ver. 1 & 2 in LD Result

Table 6.3 describes the performance of LD by using CLSG ver.1 and ver.2. On average, CLSG ver.2 achieves 0.0407% higher in precision, 1.4088 % higher in recall, and 0.7297% higher in F-measure than CLSG ver.1 respectively (i.e., excluding Chinese).

LD	CLSG ver.1			CLSG ver.2		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Japanese	0.8861	0.9722	0.9272	0.8875	0.9861	0.9342
Korean	0.9000	0.9863	0.9412	0.9000	0.9863	0.9412
Malay	0.9655	0.9655	0.9655	0.9663	0.9885	0.9773
Myanmar	0.9398	0.9630	0.9512	0.9500	0.9383	0.9441
Thai	0.9595	0.8875	0.9221	0.9157	0.9620	0.9383
Indonesian	0.9318	0.9762	0.9535	0.9765	0.9881	0.9822
Vietnamese	0.9195	0.9756	0.9467	0.9091	0.9756	0.9412
Chinese	0.7818	0.5890	0.6719	0.7818	0.5890	0.6719

Table 6.3: Performance results by LD using CLSG ver.1 and ver.2

Performance Analysis of CLPSM using CLSG ver. 1 & 2 in NS Result

Table 6.4 shows the performance of NS by using CLSG ver.1 and ver.2. Likewise in LD, using CLSG ver.2 yields on average precision of 0.9768, recall of 0.9857, and F-measure of 0.9812 which is 0.5482%, 1.2455%, and 0.8982% better than CLSG ver.1 (i.e., excluding Chinese) receptively.

NS	CLSG ver.1			CLSG ver.2		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Japanese	0.9726	0.9595	0.9660	0.9726	0.9861	0.9793
Korean	0.9600	0.9863	0.9730	0.9730	0.9863	0.9796
Malay	0.9885	0.9773	0.9829	0.9885	0.9885	0.9885
Myanmar	0.9756	0.9877	0.9816	0.9759	1.0000	0.9878
Thai	0.9747	0.9625	0.9686	0.9747	0.9747	0.9747
Indonesian	0.9762	0.9762	0.9762	0.9765	0.9881	0.9822
Vietnamese	0.9518	0.9634	0.9576	0.9759	0.9759	0.9759
Chinese	0.8065	0.3378	0.4762	0.7576	0.3378	0.4673

Table 6.4: Performance results by NS using CLSG ver.1 and ver.2

Performance Analysis of CLSG ver.1 and ver.2 in LD and NS

Table 6.5 shows the overall average performance (i.e., all languages excluding Chinese) between LD and NS by using CLSG ver.1 and ver.2. NS improved the F-measure by 0.7297% in CLSG ver.1 and 0.8982% in CLSG ver.2 receptively.

	LD			NS		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
CLSG ver.1	0.9289	0.9609	0.9439	0.9713	0.9733	0.9722
CLSG ver.2	0.9293	0.9750	0.9512	0.9768	0.9857	0.9812

Table 6.5: Overall average performance of LD and NS results by using CLSG ver.1 and ver.2

The results show that NS achieves better recall and precision than LD in all languages by using both CLSG ver.1 and CLSG ver.2. The allocation of threshold T_{LD} and T_{NS} affect performance of CLPSM algorithm regarding to precision, recall, and F-measure. However, some terms in Asian languages derive from the German, Dutch and some other languages. For examples, Natoriumu (Na) and Kariumu (K) in Japanese borrowed from Kalium (K) and Natrium (Na) in German, Yodium (I) in Indonesia borrowed from Jodium (I) in Dutch (i.e., /j/ sound using in early time of Indonesian language had changed to /y/ after 1972) [Jones, 2007]. Therefore, if these terms are measured again with the terms from the source in German and Dutch, threshold can differentiate all phonetic borrowing words. We found a few number of semantic adopting element names are similar to original source character sequence, thus some of those word pairs are fallen into phonetic borrowing group.

6.2 LEARNING APPROACH MEASURES

In this learning approach framework, transliteration mining systems have to face the problem of scripts in which:

1. the languages have similar phonemes but varied scripts (for example, "ဗီတာမင်"(bi.ta.min) and "Vitamin" in Myanmar-English);
2. the languages have different phonemes and similar scripts (for example, "加速度"(Kasokudo) in Japanese and "加速度"(Jiāsùdù) in Chinese);
3. the languages have similar phonemes and similar scripts (for example, "атом"(atom) and "ATOM" in Russian-English);
4. the languages have different phonemes and different scripts (for example, "加速度" (Jiāsùdù) and "acceleration" in Chinese-English).

However, the aim of the current research is to extract phonetic borrowing word pairs/transliterated word pairs (i.e., case 1 in the above list) from parallel or comparable corpora automatically. Moreover, retrieving the phonetic borrowing word pairs, the methodology need to deal not only with the one-to-one form of bilingual word pairs, but also with the many-to-many form of parallel corpus. Therefore this second part of the methodology was attempted based on stochastic edit distance in learning approach, EM and EMn models conducted for a one-to-one form dataset and BAYESn for both one-to-one and many-to-many form datasets.

6.2.1 Learning the Noise Prior

The EMn model is trained using an initial value for λ (the prior probability of the candidate pair being noise) of 0.6. After training had been completed the model arrived at a value of 0.6152 for lambda. The value of lambda is estimated to be 0.5835 from human-assigned labels of all of the 3,100 pair clean data set. The value learned during the EMn training is commensurate with the true value.

The BAYESn model learns a similar parameter λ, which also represents the prior probability of the candidate pair being noise. This model can not only provide an estimate for the parameter, but also a distribution indicating its uncertainty. The convergence of the value of the parameter during training is shown in Figure 5.8. It can be seen from the graph that the value of this parameter converges after about 10 training iterations. The training is continued until iteration 50, and then sampled its value over the next 500 iterations. A histogram of representing the distribution of the value of λ for this 500-iteration sample is shown in Figure 5.9. This distribution has a mean of 0.580 which is extremely close to the ground truth value of 0.5835.

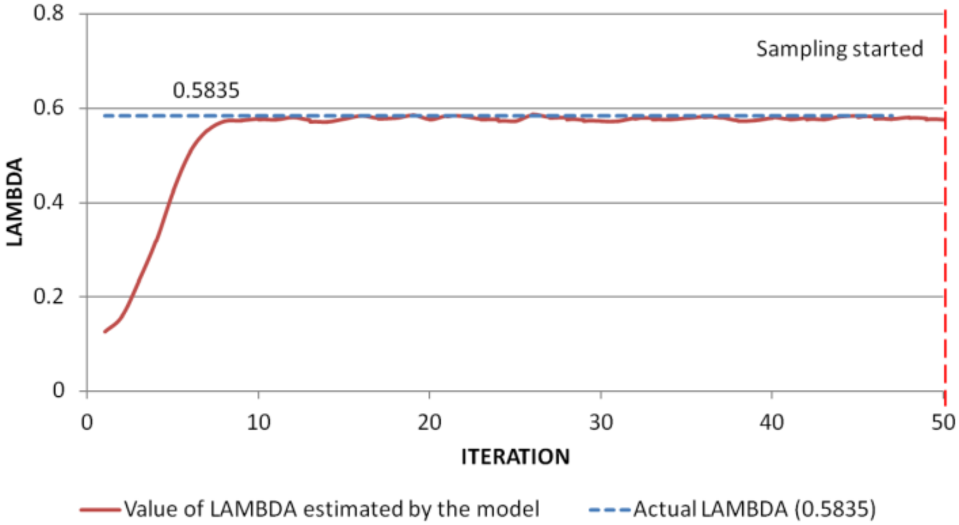


Figure 6.1: Convergence of the noise prior (λ)

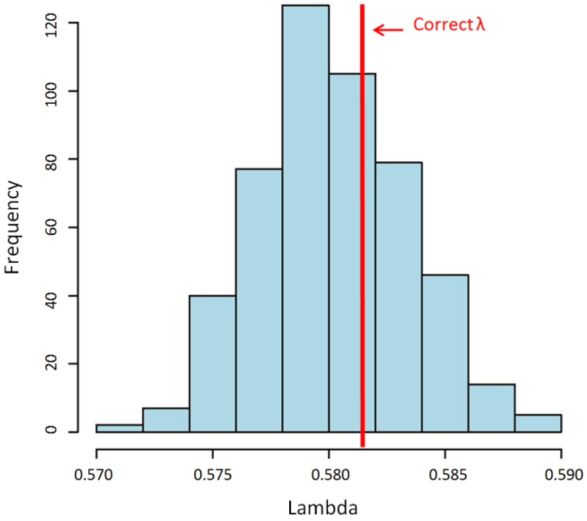


Figure 6.2: The distribution of λ from 500 samples taken after training completion

6.2.2 Using Different Romanization Systems

There are a variety of systems of romanizing Myanmar in use today. Some systems place emphasis on the orthography of the Myanmar script (transliteration), and other focus on the pronunciation of the Myanmar words (transcriptions) [Dictionary-

2008]. Our experiments used the Myanmar Language Commission Transcription System (MLCTS) and the University of Foreign Language (UFL) pronunciation system.

The results (presented in Figures 5.3, 5.4, 5.5 and 5.6) consistently show that for phonetic similarity comparison, the choice of romanization system matters, and in the case of these experiments, the UFL scheme always gave better results than using the MLCTS system. The UFL system is primarily intended for use by foreigners to pronounce the Myanmar language and this should make it more similar to the English than the MLCTS system which, as just one example, includes characters in the romanized form that represent non-articulated consonants.

To illustrate this point, consider the differences in how the word "platinum" is romanized by the two systems. In MLCTS it is romanized as "plaktinam" whereas in UFL it is romanized as "ple'ti.nan". Phonetically coding the ULF romanized form gives rise to the same coding as the original English coding (See in Table 6.6).

	English	**MLCTS**	**UFL**
Romanization	platinum	pla**k**tinam	ple'ti.nam
Phonetic Coding	19030A0A	190**6**30A0A	19030A0A

Table 6.6: An example difference between MLCTS and UFL

Moreover, the nature of Myanmar language is syllable-timed, whereas English is stress-timed, and consequently Myanmar words do not have final consonant [Dictionary-, 2008]. In the example above for the word platinum, written " ပလက်တီနမ်", the syllable "လက်" has a consonant "က" at the end, the "k" stands for "က" in the romanization but is not articulated. In pronunciation, it is easy to distinguish between a consonant and a non-articulated consonant, whereas in the transcription from Myanmar syllables to the Latin alphabet, how these consonants are represented will depend on the romanization scheme.

6.2.3 Using a Noise Model

Adding an explicit noise model [Hassan et al., 2012] to the EM [Ristad & Yianilos., 1998] and BAYES [Fukunishi et al., 2012] models substantially improved their mining performance on both clean data and realistic data in our experiments (see the results in Figures 5.3~5.7). This is at least in part due to the character of the data were using. This experimental data included word pairs extracted from a bilingual dictionary. This type of data contains multiple forms derived from the same root in which the prefixes of the words in both source and target language are identical across the word forms. An example a set of word forms and their translations is given below:

a. (စုပ်ယူနိုင်စွမ်းရည်|absorbance)

b. (စုပ်ယူနိုင်စွမ်းရည်|absorbing)

c. (စုပ်ယူအားတိုင်း|absorbing)

This set is composed entirely of non-transliterations, yet the prefixes of both source and target words are identical for all members of the set (i.e., စုပ်ယူ|absorb). The bilingual alignment models may learn erroneous features from this data since these examples will support each other when it comes to the alignments of their prefixes. Even though there is little or no support from the rest of the corpus, the mutual support from the members of the set may be enough to cause these pairs to be assigned a high enough probability to be mined as correct pairs. Using noise model can be reduced this type of problem, but it did not eliminate all from the alignments.

6.2.4 Using Phonetic Grouping (CLSG)

Using this proposed method for cross-language sound grouping (CLSG) gave the best overall mining performance in these experiments and at the same time speeded up the learning tasks. The processing time for methods are measured with and without CLSG and found that using CLSG can reduce the execution time by about 60%~70%. All attempted experiments measured 200-iterations of training on the clean dataset, and the results are shown in Table 6.7.

Model	Data	Execution Time (minutes)
EM	MLC with phonetic coding	7.13
	MLC without phonetic coding	25.54
	UFL with phonetic coding	4.86
	UFL without phonetic coding	16.17
EMn	MLC with phonetic coding	8.77
	MLC without phonetic coding	27.93
	UFL with phonetic coding	5.04
	UFL without phonetic coding	16.42

Table 6.7: Execution Time

In all these experiments, performance in transliteration mining was increased by using CLSG (see in Figures 5.3 & 5.4). The EMn method gave improved precision using phonetic grouping compared to the results obtained by using romanization. For example, EMn erroneously labeled the romanized bilingual pair (2011|nhshtnghsyti) as a true transliteration pair with probability 0.993203, whereas EMn correctly rejected this pair (the probability of it being a transliteration/phonetic borrowing word pair was 0.001139) when phonetic grouping was employed. This is due to the fact that this cross-language sound grouping approach improves to the performance of model.

From the experimental results with clean data, the performance of this approach on Myanmar data using phonetic coding (CLSG ver.2) was sufficiently high to make this technique useful in a real-world mining application: the BAYESn approach model achieved scores of 0.97 in precision, 0.95 in recall, and 0.96 in F-measure. In addition, when mining with realistic data, not only did the BAYESn model learn a strongly performing model within cross-language sound grouping (CLSG) like human expert knowledge framework, but it also speeded up the learning procedure.

6.2.5 Aligning without Romanization

In this experiment, it is aligned directly from Myanmar syllables to English characters in a many-to-many manner. This has the advantage of removing any requirement for a romanization system, thereby making it far more generally applicable. The total number of 13,483 multiple word pairs without romanization are used in these experiments, and before aligning the data, pre-processing step segmented firstly it into syllables using the procedure described in Section 5.1.2.2.

In the many-to-many alignment procedure it is possible to constrain the maximum source and target character sequence sizes. The experiments conducted in 3 different settings of these parameters to investigate their effect, constraining both source and target sequences to be of length 4, 8 and 12 tokens. The results are shown in Table 5.8, and indicate that the model tends to improve with fewer restrictions on the sequence lengths. This is consummate with the hypothesis that the model does not have a tendency to overfit the data.

Parameter	Model Parameters	Accuracy
4-4	9927	69.93%
8-8	11625	76.28%
12-12	11717	77.58%

Table 6.8: Number of model parameters and accuracy

Although the results achieved a somewhat satisfactory level, the performance was considerably lower than the systems that used romanized data. This is due to the nature of the Myanmar language itself. There are around 1,880 unique syllables in the language [Myanmar spelling book, 2003], some much rarer than others, and given the small size of the data sets available for Myanmar that was simply insufficient to train a good alignment model given the number of parameters in the model. Visual inspection of the alignment data showed that some of the rarer syllables had very little or no data to train from. Romanization is an effective way to overcome this problem since the vocabulary size can be vastly reduced, and even the rarer syllables can be represented accurately. Phonetic grouping over the romanization is for the same reason even more effective when only small amounts of data are available, as these experiments have shown in chapter 5.

6.2.6 Learning in Asian Languages

In order to analyze clearly, Table 6.9 presents the summarization of precision, recall, and F-measure of LD, NS, and EMn by using CLSG ver.2. In the Asian language dataset, EMn extracts phonetic borrowing names accurately like NS. The average performance of EMn is slightly improved than NS and significantly better than LD. Some semantic adopting names are similar to the phonetic borrowing names, which differ by only one or two-character sequence and belongs to the same articulatory group (e.g., "zinc|seng" (Zn) in Indonesian).

CLSG ver.2	En-Jp	En-Kr	En-Ml	En-Mm	En-Th	En-In	En-Vn	En-Ch
LD								
Precision	0.8875	0.9000	0.9663	0.9500	0.9157	0.9765	0.9091	0.8065
Recall	0.9861	0.9863	0.9885	0.9383	0.9620	0.9881	0.9756	0.3906
F-measure	0.9342	0.9412	0.9773	0.9441	0.9383	0.9822	0.9412	0.5263
NS								
Precision	0.9726	0.9730	0.9885	0.9759	0.9747	0.9765	0.9759	0.7576
Recall	0.9861	0.9863	0.9885	1.0000	0.9747	0.9881	0.9759	0.3378
F-measure	0.9793	0.9796	0.9885	0.9878	0.9747	0.9822	0.9759	0.4673
EMn								
Precision	0.9859	1.0000	1.0000	1.0000	0.9873	0.9880	1.0000	0.9250
Recall	0.9722	0.9863	0.9885	1.0000	0.9873	0.9762	0.9512	0.5068
F-measure	0.9790	0.9931	0.9942	1.0000	0.9873	0.9820	0.9750	0.6549

Table 6.9: Summarization of precision, recall, and F-measure of LD, NS, and EMn

6.3 Pros and Cons of Approaches

In this section, pros and cons of methodologies in two approaches based around the knowledge of phonetic similarity categorization are presented.

6.3.1 Pros of classical edit distance in manual designing approach

- It can be applied rapidly to diverse tasks.
- It does not require any training datasets after it has been developed.
- It can be measured between one-to-one form of bilingual word pairs
- It is applicable to CLIR and various linguistic tic research

6.3.2 Cons of classical edit distance in manual designing approach

- It is somewhat arbitrary.
- If either source or target recipient string length is less than four characters, it cannot distinguish correctly between semantic and phonetic borrowing word pairs.
- non-Latin script languages depend on romanization scheme.
- It was only enable to measure the loanword in single word form (i.e., one-to-one)

6.3.3 Pros of stochastic edit distance in learning approach

- It is easily adaptable to various tasks.
- It is useful for low-resource language datasets, where phonetic similarity categorization knowledge can compensate for a lack of data resources (i.e., it does not required large amount of training datasets).
- EM and EMn can be mined one-to-one form of bilingual word pairs whereas BAYESn can mine both one-to-one and many-to-many form of datasets
- Using a knowledge of phonetic similarity categorization can be reduced the execution time by about 60% to 70%.
- It helps to improve mining performance
- Algorithms in learning approach can be applied the applications such as mining training data for transliterations, improving lexical coverage for MT, CLIR via translation resource expansion.

6.3.4 Cons of stochastic edit distance in learning approach

- EM model learning failed in both romanized and CLSG phonetic coding datasets

- Training in EM and EMn was not able to perform in un-romanized data and many-to-many form of data
- It is also depended on romanization scheme

Chapter : 7

Conclusion and Future Research

This chapter presents an overview of work completed, important findings and lessons learned. It also concludes some brief discussions on the limitations of the proposed techniques, and presents the future direction of work.

7.1 SUMMARY

With the objective of developing a methodology that can differentiate between semantic adopting and phonetic borrowing word pairs across the Asian languages, this dissertation reports the key feature of Cross-Language Sound Grouping (CLSG) and edit distance based similarity measure techniques in two approaches: the manual designing approach and the learning approach. However, this research focuses on stochastic edit distance measure in the learning approach; Levenshtein distance is used to apply an initial development task by integrating a primitive knowledge of CLSG.

In this first attempt, we conducted this study knowing the etymological origins of the chemical names in the target languages (i.e., we know which terms are either phonetically borrowed or semantically adopted from the source language). The results prove that our algorithm accurately retrieves the phonetic borrowing terms. Interestingly, simplifying sounds in a specific language and normalization turned out to work best for all languages. On average, by using CLSG ver.2, our results achieved 0.9767 precision, 0.9857 recall, and 0.9811 for the F-measure of normalization (NS) in all the Asian languages, excluding the Chinese (i.e., referred to Chapter 6: Tables 6.3, 6.4, 6.5). Because of the nature of language adopting method, Chinese cannot be differentiated between phonetic and semantic borrowing groups using this methodology. Therefore, we conclude that if the words in a language are borrowed phonetically from other languages, our algorithm can distinguish between phonetic borrowing and semantic adopting word pairs accurately. Moreover, CLSG can be applicable to the same language family of experimental Asian languages (i.e., referred to Chapter 6).

In the main development of our work, a novel technique is developed based stochastic models of string similarity (EM, EMn, & BAYESn) (i.e., referred to Chapter 5). The CLSG-like human knowledge is integrated into the stochastic models. Moreover, adding noise model into stochastic edit distance and Bayesian alignment gives the best performance results (i.e., increasing precision). This approach is expected to be most useful when working with low resource languages as human knowledge can be used to mitigate issues of data sparseness resulting from lack of data; the number of edit costs that need to be learned can cause problems in low resource languages where there may not be enough data available to learn them accurately. Therefore, the proposed approach uses CLSG like human expert knowledge to provide a framework within which a machine can learn the edit costs by effectively tying parameters in a linguistically-motivated manner in order to reduce the number of parameters to be learned, thereby simplifying and speeding up the learning task. In addition, EMn training in Asian languages achieves the best performance of 1.4999% higher F-measure than NS measurement (i.e., using CLSG ver.2). Consequently, our proposed method has potential application for CLIR, MT, and various linguistic studies.

7.2 CONTRIBUTIONS

The main research has contributed to CLIR and MT in two ways: first by building cross-language sound grouping (i.e., extension of Soundex phonetic system) for Asian languages, and second by developing a transliteration mining system based on stochastic edit distance in learning approach for low resource languages.

1. The first contribution of this work is the development of cross-language sound grouping for Asian languages. Based on the study of language borrowing and phonological concept of those languages, the Soundex phonetic coding system is extended to use in our methodology (i.e., referred to Section 3.1).
2. The second contribution of this work is a development of cross-language phonetic similarity metrics (CLPSM) in learning approach to retrieve phonetic borrowing word pairs from parallel corpora, especially for the low resource Asian languages (i.e., referred to Section 3.3).

A brief background and usefulness of methodology

Usually, phonetic borrowing words appear among the languages based on the etymological relationship they have with each other or the borrowing language used. Today, many spoken languages appear in a variety of texts/graphemes on the web and digital documents, CLIR (e.g. cross-language search engine), data mining, and MT systems are necessary to raise the coverage and accessibility for those languages. Most of the lexicons/corpuses do not provide a good coverage of proper nouns (e.g. names and technical terms) and it turns out these appear often in queries and translation tasks to constitute the largest class of out-of-vocabulary terms in CLIR and MT systems. Many examples of literature have been developed with various sophisticated approaches to explore in these areas, there still remains a lot of improvement in the complex multilingual retrieval labyrinth. This development work is an extension to existing systems in terms of added support for retrieving loanwords/phonetic borrowing words/transliterated words/OOV terms from various sources and recipient Asian languages.

Improving performance of methodology

A noble attempt of this work, the knowledge of CLSG helps to improve the performance of both classical edit-distance and stochastic edit-distance measurements. Moreover, this work is expected to be most useful when working with low resource/limited resource languages as human knowledge of CLSG can be used to mitigate issues of data sparseness resulting from lack of data and learning problems. The use of CLSG like human knowledge is illustrated in cross-language edit-distance measure to distinguish between phonetic borrowing and semantic adopting words in applications: mining training data for transliterations, improving lexical coverage for MT, and CLIR via translation resource expansion. In addition, their use in searching out-of-vocabulary terms indicates the applications of CLIR systems and various linguistic areas. Furthermore, providing knowledge of CLSG to the weights for the machine learning process reduced the number of parameters to be learned, simplifying and speeding up the learning task. Moreover, the experiments show that in a Myanmar-English transliteration mining task, this approach substantially improves mining performance when mining realistic data [referred to in Section 5.2.4].

7.3 LIMITATIONS

In the remainder of this chapter, some limitations on the work in this dissertation is presented that occurred during the development and evaluation of the systems presented. Some of these limitations result from experiments, but that does not represent fundamental deficiencies in the approach. Section 7.4 presents further work to be done in this area that could extend the applicability and performance of the approach.

- Experimentation with more languages

  Current methodology attempted to test eight Asian languages in both manual designing and learning approaches. The first Myanmar-English transliteration mining experiment is attempted in the learning approach. We need to test in additional languages from other types of language family (e.g., Indic and Sanskrit).

- Failed to differentiate between phonetic borrowing words and semantic adopting words in Chinese

  Most of the loanwords in Chinese are semantic adopting words, generated by creating new loanwords (i.e., a type of loan creation in semantic adoption). Our phonetic similarity metrics in both manual designing and EMn learning approaches could not differentiate between phonetic and semantic borrowing word pairs properly.

- One-to-one alignment in EM and EMn

  Though the nonparametric BAYESn is able to learn many-to-many alignment with different script parallel corpora/bilingual data, EM and EMn models failed to work in it.

7.4 FUTURE DIRECTION

This section explores other areas to be explored within the cross-language sound grouping and stochastic edit distance in learning approach framework that have not yet been fully addressed by this dissertation, as well as some new research lines:.

- Expansion of CLSG

The current version of CLSG (CLSG ver. 2) treats the phonetic coding for source language: English and eight recipient Asian languages: Japanese, Korean, Malay, Myanmar, Thai, Malay, Indonesian, and Vietnamese within five types of language family. The knowledge of CLSG allows room for modifications and can be easily adopted as per the required additional languages from different language family. With a practicable approach of phonological knowledge, this methodology confirms that a standard framework can be part of any CLIR and MT applications.

- Increasing performance of aligning without romanization in EMn and BAYESn

In order to study the utility of performing alignment without romanization, an explicit noise model is added to a non-parametric Bayesian alignment model. The results show that this model works as well as the model based on EM training, but when applied to un-romanized data the model was not able to perform at the same level as the systems based on romanized data. As observed this approach failed for the same reason the CLSG phonetic coding approach succeeded: the direct alignment of Myanmar to English introduces too many parameters to be learned from the small amount of available data. This technique still may be viable for languages with more data, and/or smaller input grapheme/syllable set sizes where the data sparseness issues are less severe, but this remains for future work.

- Learning realistic data within a semi supervised framework

The experimental results clearly show that the choice of edit cost is a strong factor in determining the performance of the edit-distance-based techniques used in these experiments (i.e., referred to Section 5.2, 5.3 and 6.2). Often edit costs are selected to have plausible values by human experts (i.e., knowledge of cross-language sound grouping-CLSG), but better results can be obtained through the application of machine learning techniques to learn appropriate edit costs. Furthermore, the mining performance using stochastic models depends heavily on the romanization system used (i.e., referred to results in section 5.2), and this motivates further research in the area of string representation. All the experiments in this work were performed using an unsupervised mining approach, and in future research it would be interesting to study realistic/noisy data mining within a semi supervised framework.

Appendix

List of Chemical Element Names in Asian Languages Table is attached here.

- Japanese, Korean, Chinese, Myanmar, and Thai are romanized names.

No.	Name	Sym.	Japanese	Myanmar	Malay	Chinese	Korean	Thai	Vietnamese	Indonesian
1	hydrogen	H	suiso	huiakdruigrang	hidrogen	qing	suso	haidrohjaehn	hidro	hidrogen
2	helium	He	heriumu	hiliyam	helium	hai	hellyum	heeliiam	heli	helium
3	lithium	Li	richiumu	lisiyam	litium	li	rityum	lithiiam	liti	litium
4	beryllium	Be	beririumu	bhairileyam	berilium	pi	berillyum	boerinliam	berili	berilium
5	boron	B	houso	buirwan	boron	peng	bungso	bohrawn	bo	boron
6	carbon	C	tanso	kabwan	karbon	tan	tanso	khaarbaawn	cacbon	karbon
7	nitrogen	N	chisso	nuiakdruigrang	nitrogen	dan	jilso	naitrochen	nito	nitrogen
8	oxygen	O	sanso	aauakhcigrang	oksigen	yang	sanso	awksijen	oxy	oksigen
9	fluorine	F	fusso	hpluirang	fluorin	fu	peulluorin	fluorin	flo	fluor
10	neon	Ne	neon	niywan	neon	nai	neon	nion	neon	neon
11	sodium	Na	natoriumu	hcuidiyam	sodium	na	sodyum	sohdiiam	natri	natrium
12	magnesium	Mg	maguneshiumu	mggniciyam	magnesium	meng	mageunesyum	maekneesiiam	magie	magnesium
13	aluminium	Al	aruminiumu	aluminiyam	aluminium	lu	alluminyum	aluuminiiam	nhom	alumunium
14	silicon	Si	shirikon	cilikwan	silikon	gui	gyuso	silikhawn	silic	silikon
15	phosphorus	P	rin	hpecprp	fosforus	lin	in	faawtfaawrat	photpho	fosfor
16	sulfur	S	iou	hcalhpa	sulfuroksida	liu	hwang	gammathan	luuhuynh	belerang
17	chlorine	Cl	enso	kluirang	klorin	lu	yeomso	khlaawreen	clo	Klor
18	argon	Ar	arugon	aagwan	argon	ya	areugon	aargawn	agon	argon
19	potassium	K	kariumu	puitakciyam	kalium	jia	potasyum	bpohdtaaetsiiam	kali	kalium
20	calcium	Ca	karushiumu	htuandhat	kalsium	gai	kalsyum	bpuun	canxi	kalsium
21	scandium	Sc	sukanjiumu	hckandiyam	skandium	kang	seukandyum	sagohndiayohm	scandi	skandium
22	titanium	Ti	chitan	tuikteniyam	titanium	tai	taitanyum	thaithaehniiam	titan	titanium
23	vanadium	V	banajiumu	bndiyam	vanadium	fan	banadyum	waanaehdiiam	vanadi	vanadium
24	chromium	Cr	kuromu	hkruimiyam	kromium	ming	keuromyum	khromiam	crom	Krom
25	manganese	Mn	mangan	mangggni	mangan	mei	mangganijeu	maaenggaaneet	mangan	mangan
26	iron	Fe	tetsu	sam	besi	tie	cheol	lek	sat	besi
27	cobalt	Co	kobaruto	kuibhe	kobalt	gu	kobalteu	khohbaawnd	coban	kobal
28	nickel	Ni	nikkeru	nikai	nikel	nie	nikel	nikgeern	niken	nikel
29	copper	Cu	dou	kre	kuprum	tong	guri	hang	dong	tembaga
30	zinc	Zn	aen	swp	zink	xin	ayeon	sanggasee	kem	seng
31	gallium	Ga	gariumu	gailiyam	galium	jia	gallyum	gaaenliiam	gali	galium
32	germanium	Ge	gerumaniumu	gamanniyam	germanium	zhe	jeomanyum	choemeniam	gecmani	germanium
33	arsenic	As	hiso	ciandrpcang	arsenik	shen	biso	sannuk	asen	arsen
34	selenium	Se	seren	cailiniyam	selenium	xi	sellenyum	seleniam	selen	selenium
35	bromine	Br	shuuso	bhruiman	bromin	xiu	beuromin	baromin	brom	brom
36	krypton	Kr	kuriputon	hkracptwan	kripton	ke	keuripton	khripthon	krypton	kripton
37	rubidium	Rb	rubijiumu	rubhidiyam	rubidium	ru	rubidyum	rubitdiam	rubidi	rubidium
38	strontium	Sr	sutoronchiumu	ctuirwantiyam	strontium	si	seuteurontyum	sotronchiam	stronti	stronsium
39	yttrium	Y	ittoriumu	rihtriyam	ytrium	yi	iteuryum	itriam	yttri	itrium
40	zirconium	Zr	jirukoniumu	jakuiniyam	zirkonium	gao	jireukonyum	soekhoniam	zirconi	zirkonium
41	niobium	Nb	niobu	niauibhiyam	niobium	ni	naiobyum	naiohbiiayohm	niobi	niobium
42	molybdenum	Mo	moribuden	muilbhdiyam	molibdenum	mu	mollibeudeneom	moleepdinam	molypden	molibdenum
43	technetium	Tc	tekunechiumu	takhkyniyam	technetium	de	tekeunetyum	theksatneechiiap	tecneti	teknesium
44	ruthenium	Ru	ruteniumu	ruSiyam	rutenium	liao	rutenyum	ruthiniam	rutheni	rutenium
45	rhodium	Rh	rojiumu	ruidiyam	rodium	lao	rodyum	rodiam	rhodi	rodium
46	palladium	Pd	parajiumu	plakdiyam	paladium	ba	palladyum	phaenlediam	paladi	paladium
47	silver	Ag	gin	ngw	argentum	yin	eun	ngern	bac	perak
48	cadmium	Cd	kadomiumu	katmiyam	kadmium	ge	kadeumyum	khaetmiam	cadmi	kadmium
49	indium	In	injiumu	aandiyam	indium	yin	indyum	indiam	indi	indium
50	tin	Sn	suzu	samhpru	timah	xi	juseok	deebook	thiec	timah
51	antimony	Sb	anchimonii	aantimuini	antimoni	ti	antimoni	phluaang	antimon	antimon
52	tellurium	Te	teruru	teluriyam	telurium	di	telluryum	thenyuriam	telua	telurium
53	iodine	I	youso	auiangauidang	iodin	dian	aiodin	aiodin	iot	yodium
54	xenon	Xe	kisenon	jinwan	xenon	xian	jenon	sinon	xenon	xenon
55	caesium	Cs	seshiumu	cihciyam	sesium	se	sesyum	seesiiam	xezi	sesium
56	barium	Ba	bariumu	bhariyam	barium	bei	baryum	bariam	bari	barium
57	lanthanum	La	rantan	laksnam	lantanum	lan	rantaneom	laaenthaanan	lantan	lantanum
58	cerium	Ce	seriumu	ciriyam	serium	shi	seryum	chaiphasida	xeri	serium
59	praseodymium	Pr	puraseojimu	praciauidimiyam	praseodimium	pu	peuraseodimyum	phrseeohdimiyohm	praseodymi	praseodinium
60	neodymium	Nd	neojimu	niauidimiyam	neodimium	nu	neodimyum	neeohdimiian	neodymi	neodinium
61	promethium	Pm	puromechiumu	pruimaksiyam	prometium	ju	peurometyum	phrmeethiiang	promethi	prometium
62	samarium	Sm	samariumu	hcmariyam	samarium	shan	samaryum	saamariiam	samari	samarium
63	europium	Eu	yuuropiumu	airupiyam	europium	you	yuropyum	yuurophiyohm	europi	europium
64	gadolinium	Gd	gadoriniumu	gaduiliniyam	gadolinium	ga	gadollinyum	gaaedooyliniian	gadolini	gadolinium
65	terbium	Tb	terubiumu	tabhiyam	terbium	te	teobyum	thebiian	terbi	terbium
66	dysprosium	Dy	jisupuroshiumu	duiangcpruiciyam	disprosium	di	diseupeurosyum	dithosiian	dysprosi	disprosium
67	holmium	Ho	horumiumu	huiliyam	holmium	huo	holmyum	hohtmiiam	holmi	holmium
68	erbium	Er	erubiumu	aabiyam	erbium	er	eobyum	uuhrbiian	ecbi	erbiam
69	thulium	Tm	horumiumu	suliyam	tulium	diu	tullyum	thuuliiam	thuli	tulium
70	ytterbium	Yb	itterubiumu	yaktabhiyam	yterbium	yi	iteobyum	ipthebiian	yttecbi	iteribium
71	lutetium	Lu	rutechiumu	lutetiyam	lutetium	lu	rutetyum	lutheediam	luteti	lutesium
72	hafnium	Hf	hafuniumu	hahpniyam	hafnium	jia	hapeunyum	rhafaniiayohm	hafni	hafnium
73	tantalum	Ta	tantaru	tantailiyam	tantalum	dan	tantalleom	thaaenthalan	tantali	tantalum
74	tungsten	W	tangusuten	taungctan	tungsten	wu	bolpeulam	wulferm	volfram	tungsten
75	rhenium	Re	reniumu	rhiniyam	renium	lai	renyum	reeniiama	rheni	renium
76	osmium	Os	osumiumu	auicmiyam	osmium	e	oseumyum	aawtmiamichu	osmi	osmium
77	iridium	Ir	irijiumu	airidiyam	iridium	yi	iridyum	aireetdiiam	iridi	iridium
78	platinum	Pt	purachina	rwhhpru	platnium	bo	baekgeum	phathinaa	platin	platinum
79	gold	Au	kin	rwh	emas	jin	geum	thaawngkham	vang	emas
80	mercury	Hg	suigin	prda	raksa	gong	sueun	phoot	thuyngan	raksa
81	thallium	Tl	tariumu	sailiyam	talium	she	tallyum	thalaliiam	tali	talium
82	lead	Pb	namari	hkai	plumbum	qian	nap	dtaguaa	chi	timbal
83	bismuth	Bi	bisumusu	bhccmat	bismut	bi	biseumuteu	bitmat	bitmut	bismut
84	polonium	Po	poroniumu	puiluiniyam	polonium	pu	pollonyum	phaawlohniian	poloni	polonium
85	astatine	At	asutachin	aakcttang	astatina	ai	aseutatin	raawthaathini	astatin	astatin
86	radon	Rn	radon	rdwan	radon	dong	radon	raehdaawn	radon	radon
87	francium	Fr	furanshiumu	pranciyam	fransium	fang	peurangsyum	phaaesiian	franxi	fransium
88	radium	Ra	rajiumu	rdiyam	radium	lei	radyum	raehdiiam	radi	radium
89	actinium	Ac	akuchiniumu	aaktangniyam	aktinium	ei	aktinyum	aawthitniian	actini	aktinium
90	thorium	Th	toriumu	suiriyam	torium	tu	toryum	thaawriiam	thori	torium
91	protactinium	Pa	purotoakuchiniu	pruitaktangyam	protaktinium	pu	peurotaktinyum	phrohthohtniian	protactini	protaktinium
92	uranium	U	uran	yureniyam	uranium	you	uranyum	yuuraehniiam	urani	uranium

Glossary

ACL	The Association for Computational Linguistics
BAYESn	Bayesian Noise Model
CLIR	Cross-Language Information Retrieval
CLPSM	Cross-Language Phonetic Similarity Metrics
CLSG	Cross-Language Sound Grouping
CRP	Chinese Restaurant Process
EM	Expectation Maximization Model
EMn	Expectation Maximization Noise Model
IPA	The International Phonetic Alphabet
LD	Levenshtein distance
MLCTS	Myanmar Language Commission Transcription System
MT	Machine Translation
NS	Normalized Similarity
OOV	Out-Of-Vocabulary
UFL	University of Foreign Language

References

[1] 社団法人日本ローマ字会.「99 式」日本語のローマ字表記方式. 1999-7. Roomazi Sekai. No. 675, pp.3-19.

[2] Naing, Aung Win. (2006). An Introductory Course in Myanmar Language: University of Foreign Languages, Yangon, pp.1-26.

[3] Barker Milton E. (1969). The phonological adaptation of French loanwords in Vietnamese: Mon-Khmer Studies, 3, pp. 138–47.

[4] Bhagat, Rahul. and Eduard Hovy. (2007). Phonetic Models for Generating Spelling Variants: In proceedings of the 20th international joint conference on Artifical intelligence, pp.1570-1575.

[5] Brill, Eric. Gary Kacmarcik. Chris Brockett. (2001). Automatically Harvesting Katakana-English Term Pairs from Search: Asia Federation of Natural Language Processing.

[6] Charles, B. Chang. (2008). Phonetic vs. Phonology in Loanword Adaption: Revisiting the Role of the Bilingual: UC Berkeley Phonology Lab Annual Report 2008.

[7] Colin de la Higuera. Luisa Mico. (2008). A contextual normalised edit distance: In proceeding of Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference, pp.354–361.

[8] CONGRESS, LIBRARY OF. Cataloging Policy and Support Office. ALA-LC Romanization Tables: Transliteration Schemes for Non-Roman Scripts, 1997 edition [online]. 17 June 2004 [cited 21 May 2013]. Available from: http://www.loc.gov/catdir/cpso/roman.html

[9] Delahunty, Gerald P. and Garvey James J. (2010). The English Language: From Sound to Sense: Perspectives on Writing. Fort Collins, Colorado: The WAC Clearinghouse and Parlor Press, http://wac.colostate.edu/books/sound/

[10] Denshi Jisho Online Japanese dictionary: Electronic Dictionary Research Group. Monash University. (accessed 2013/03/18). http://jisho.org/kanji/details/%E7%94%9F%E7%89%A9%E5%AD%A6,

[11] Dictionary- Myanmar Language Commission. (2008). Ministry of Education, Myanmar. Myanmar-English Dictionary, 9th Edition. 2008.

[12] Emsley, John. (2011). Nature's Building Blocks: An A-Z Guide to the Elements (New ed.). New York, NY: Oxford University Press. ISBN 978-0-19-960563-7.

[13] Encyclopedia Britannica. (accessed 2013/03/14). Linguistic-Borrowing: http://global.britannica.com/EBchecked/topic/342418/linguistics/35113/Borrowing

[14] Ennis, Daniel M. and F.Gregory (2007). Similarity Measures : Scholarpedia. (accessed 2013/01/27). http://www.scholarpedia.org/article/Similarity_measures

[15] Eugene, E. Loos, Susan Anderson, Dwight H., Day, Jr. , Paul C. Jordan, J. Douglas Wingate. (2003). Glossary of linguistic terms: International Linguistics Department, SIL International.

[16] Finch, Andrew. Keiji Yasuda. Hideo Okuma. Eiichiro Sumita. and Satoshi Nakamura. (2011). A Bayesian Model of Transliteration and Its Human Evaluation When Integrated into a Machine Translation System: IEICE Transactions on Information and Systems E94-D, 10, pp. 1889-1900.

[17] Finch, A. Dixon, P. and Sumita. (2011a). Integrating models derived from non-parametric bayesian co-segmentation into a statistical machine transliteration system: In proceedings of the Named Entities Workshop. Asian Federation of Natural Language Processing, Chiang Mai, Thailand, pp. 23-27.

[18] Finch, Andrew. and Eiichiro Sumita. (2010). A Bayesian Model of Bilingual Segmentation for Transliteration: proceedings of the 7th International Workshop on Spoken Language Translation, pp. 259-266.

[19] Francis, John De. (1984). The Chinese Language - Fact and Fantasy. (ISBN: 0-8248-1068-6 / 0-8284-0866-5)

[20] Freeman, Andrew T. Sherri L. Condon. Christopher M.Ackerman. (2006). Cross Linguistic Name Matching in English and Arabic: A “One to Many Mapping” Extension of the Levenshtein Edit Distance Algorithm: In proceeding of the Human Language Technology Conference of the North American Chapter of the ACL, New York, pp.471-478.

[21] Fukunishi, Takaaki. Andrew Finch. Seiichi Yamamoto. Eiichiro Sumita. (2012). A Bayesian Alignment Approach to Transliteration Mining: ACM Transactions on Asian Language Information Processing, vol. 9, no. 4, article. 39.

[22] Gillian, Kay. (1995). English loanwords in Japanese: Journal of World Englishes, vol.14, no.1, pp.67-76

[23] Halvor Eifring & Rolf Theil. (2005). Linguistics for Students of Asian and African Languages: University of Oslo. link. http://www.uio.no/studier/emner/hf/ikos/EXFAC03-AAS/h05/larestoff/linguistics/Chapter%206.%28H05%29.pdf (accessed 2013/03/21)

[24] Hassan, Sajjad. Alexander Fraser. and Helmut Schmid. (2012). A Statistical Model for Unsupervised and Semi-Supervised Transliteration Mining:

proceedings of Association for Computational Linguistics (ACL-2012) conference.

[25] He, X. (2007). Using word dependent transition models in HMM based word alignment for statistical machine translation: proceeding of 2nd ACL Workshop on Statistical Machine Translation.

[26] Heeringa, Wilbert. Peter Kleiweg. Charlotte Gooskens & John Nerbonne. (2006). Evaluation of String Distance Algorithms for Dialectology: Proceeding of the workshop on Linguistic Distance, Association for Computational Linguistics, Sydney, July 2006, pp. 51-62

[27] Hiramoto, Mie. (2007). Lexical strata of Indonesian vocabulary: In IWASAKI, S., SIMPSON, A., ADAMS, K., SIDWELL , P. (eds.), SEALSXIII: Papers from the 13th meeting of the South Asian Linguistic Society 2003.

[28] Holmes D. & C.M.McCabe. (2002). Improving precision and recall for soundex retrieval: In proceedings of the IEEE International Conference on Information Technology – Coding and Computing (ITCC), Las Vegas.

[29] Ho-Min Sohn. 1999. The Korean Language: Cambridge University Press, pp. 12-13.

[30] Htun, Ohnmar, Shigeaki Kodama, Yoshiki Mikami. (2011). Cross-language Phonetic Similarity Measure on Terms Appeared in Asian Languages: International Journal of Intelligent Information Processing 2(2):9-21

[31] Htun, Ohnmar, Andrew Finch, Eiichiro Sumita, and Yoshiki Mikami. (2012). Improving Transliteration Mining by Integrating Expert Knowledge with Statistical Approaches: International Journal of Computer Applications 58(17):12-22

[30] Ibrahiem M.M EI Emary. Jafar Atwan. (2005). Designing and building an automatic Information retrieval system for handing the Arabic data: Journal of American joural of applied science, vol. 2(11): no. 1520-1525.

[31] IPA. (accessed 2013/03/25). International Phonetic Alphabet: http://en.wikipedia.org/wiki/International_Phonetic_Alphabet

[32] Iverson Gregory K. and Ahrong Lee. (2006). Perception of Contrast in Korean Loanword Adoption: International Circla of Korean Linguistic 2006. Korean Linguistic, Volume 13, pp. 49-87

[33] Jone, Russell. (2007). Loan-Words in Indonesian and Malay: Compiled by the Indonesian etymological project, p.xi.

[34] K Saravanan. Raghavendra Udupa. and A Kumaran. (2011). Improving Cross-Language Information Retrieval by Transliteration Mining and Generation: proceedings of Tamil Internet Conference, in Philadelphia.

[35] Kahki, Ali EI. Kareem Darwish. Ahmed Saad EI Din. Mohamed Abd EI-Wahab. Ahmed Hefny. and Waleed Ammar. (2011). Improved Transliteration Mining Using Graph Reinforcement: proceedings of EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1384-1393.

[36] Kahki, Ali EI. Kareem Darwish. Ahmed Saad EI Din. and Mohamed Abd EI-Wahab. (2012). Transliteration Mining Using Large Training Test Sets, proceedings of 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 243-252.

[37] Kareem Darwish. (2010). Transliteration Mining with Phonetic Conflation and Iterative Training: proceedings of the 2010 Named Entities Workshop, ACL 2010, pp. 53-56.

[38] Kenstowicz, Michael & Atiwong. (2006). Issue in Loanword Adoption: a Case Study from Thai: Lingua 116: pp. 921-949.

[39] Kodama, Shigeaki. (2010). String Edit Distance for Computing Phonological Similarity between Words: proceedings of International Symposium on Global Multidisciplinary Engineering, 2010.

[40] Kondark Grzegorz and Sherif Tarek. (2006). Evaluation of Several Phonetic Similarity Algorithms on the Task of cognate Identification: Proceeding of the Workshop on Linqguestic Distance (Sydney), pp. 43-50

[41] Korean language, The national institute of the. (2000). The Revised Romanization of Korean", Ministry of Culture & Tourism.

[42] Kumaran A. Mitesh Khapra. Haizhou Li. (2010). Whitepaper on NEWS 2010 Shared Task on Transliteration Mining: Whitepaper of NEWS 2010 Shared Task on Transliteration Generation.

[43] Ladefoged, Peter. (2001). Vowels and Consonants: An Introduction to the Sounds of Languages (Second Edition): Blackwell Publishing, USA, UK, Australia.

[43] Levenshtein, V. I.. (1966). Binary codes capable of correcting deletions, insertions, and reversals: Journal of Soviet Physics Doklady, vol. 10, no. 8, pp. 707-709.

[44] Marina Sokolova. Guy Lapalme. (2009). A systematic analysis of performance measures for classification tasks: Journal of Information Processing and Management, Elsevier, vol.45, no.4, pp.427-437.

[45] Martin Haspelmath and Uri Tadmor. (2009). Loanwords in the World's Languages, A Comparative Handbook: Walter De Gruyter Mouton & Co., KG, 10785 Berlin, Germany. pp. 573-598.

[46] Mikami, Yoshiki. (accessed 2013/03/18). World of Scripts in Asia: http://gii2.nagaokaut.ac.jp/ws/ideographs.html

[47] MLCTS. (1980). Myanmar Language Commission Transcription_System: Myanmar Language Commission. http://en.wikipedia.org/wiki/MLC_Transcription_System (accessed 2012/10/12)

[48] MTD. (2009). (accessed 2013/03/18). Multilingual Terminology Dictionary: Nagaoka University of Technology. http://gii2.nagaokaut.ac.jp/mtd/

[49] Myanmar spelling book. (2003). မြန်မာစာအဖွဲ့ဦးစီးဌာန၊ ၂၀၀၃၊ မြန်မာစာလုံးပေါင်းသတ်ပုံကျမ်း၊ ပညာရေး ဝန်ကြီးဌာန၊ ပြည်ထောင်စုမြန်မာနိုင်ငံတော်အစိုးရ

[50] Naing, Aung Win. (2006). An Introductory Course in Myanmar Language: University of Foreign Languages, Yangon, pp.1-26.

[51] Nerbonne, John. Erhard Hinrichs. (2006). Linguistic Distance: Proceeding of the workshop on Linguistic Distance, Association for Computational Linguistics, Sydney, July 2006, pp. 1-6.

[52] Odden, David. (2005). Introducing Phonology: Cambridge University Press, pp.34-39.

[53] Okell, John. (1971). A Guide to the Romanization of Burmese: The Royal Asiatic Society of Great Britain and Ireland.

[54] Olshausen, Bruno A. (2004). Bayesian probability theory: NPB 163/PSC 128 - Information processing models in neuroscience and psychology (Winter 2004)

[55] Raghavendra Udupa, K. Saravanan, Anton Bakalov, and Abhijit Bhole. (2009). "They Are Out There, If You Know Where to Look": Mining Transliterations of OOV Query Terms for Cross-Language Information Retrieval: Proceeding in 1th European Conference on IR Research, ECIR 2009. pp. 437-448

[56] Royal Thai Institute of Thailand. (2002). Royal Thai General System of Transcription for Thai, http://en.wikipedia.org/wiki/Royal_Thai_General_System_of_Transcription

[57] Russell R.C. and Odell K.M. (1918 and 1922). Soundex phonetic comparison system: [cf. U.S. Patents 1261167(1918), 1435663(1922)], USA.

[58] Ristad, Eric Sven and Yianilos, Peter N. (1998). Learning String-Edit Distance: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 5.

[59] Ricardo Baeza-Yates. Berthier Ribeiro-Neto. (1999). Modern Information Theory: ACM Press, pp.73-97.

[60] Romanization. (accessed 2013/03/25). http://en.wikipedia.org/wiki/Romanization

[61] Romanization of Japanese. (accessed 2013/03/26). http://en.wikipedia.org/wiki/Romanization_of_Japanese

[62] Royal Thai Institute of Thailand. (2002). Royal Thai General System of Transcription for Thai 2002.

[63] Sasaki Yutaka. (2007). The Truth of F-measure: Teaching: Tutorial materials Version. 26th October 2007.

[64] Seman, Noraini. (2008). Acoustic Pronunciation Variations Modeling for Standard Malay Speech Recognition: Journal of Computer and Information Science, vol. 1(4):pp.112-120.

[64] Shibles, Warren A. (1994). IPA Transliteration of Chinese Romanization Systems: SINO-PLATONIC PAPERS Number 52 November, 1994

[65] Soundex. (accessed 2013/03/26). http://en.wikipedia.org/wiki/Soundex

[66] Soundex - Ex. (accessed 2013/03/26). www.sound-ex.com/soundex_method.htm

[67] Steuart, Bradley W. (1994). The Soundex Daitch-Mokotoff Reference Guide. 2 v. Bountiful, Utah: Precision Indexing, 1994.

[68] Suwanvisat, Prayut. Somchai Prasitijutrakul. (1998). Thai-English Cross-Language Transliterated Word Retrieval using Soundex Technique: In proceedings of National Computer Science and Engineering Conference, Bangkok, Thailand, pp.210-213.

[69] The Korean language, the national institute of. (2000). The Revised Romanization of Korean: Ministry of Culture & Tourism, 2000.

[70] UFL. (2005). An introductory course in Myanmar language CD-Rom: University of Foreign Language. Yangon, Myanmar.

[71] U.S. Board on Geographic Names (1994). Romanization Systems and Roman-Script Spelling Conventions (PDF). Defense Mapping Agency. OCLC 31881487. Retrieved 8 January 2013. [cited 21 May 2013]. Available from: http://libraries.ucsd.edu/bib/fed/USBGN_romanization.pdf

[72] Volker Tresp. (2006). (accessed 2013/04/12). Dirichlet Processes and Nonparametric Bayesian Modelling: http://videolectures.net/mlss06au_tresp_dpnbm/

[73] Wells, J. C. (2006). Phonetic transcription and analysis. Encyclopedia of Language and Linguistics. 2nd edn. London: Elsevier, 9, 396-410.

[74] Wii-vun Taiffalo Chiung. (2001). Missionary scripts from Vietnam and Taiwan: In proceeding of the sixth meeting of Southeast Asian Linguistics Society, Mahidol University, Bangkok, Thailand.

[75] Wiki-1. 2013. (accessed 2013/05/14) http://en.wikipedia.org/wiki/Velar_nasal

[76] Ying Zhang. (2006). Improved Cross-language Information Retrieval via Disambiguation and Vocabulary Discovery: Schoold of Computer Science and Information Technology, RMIT University, Melbourne, Victoria, Australia. pp. 3.

[77] Yuzana. Khin Marlar Tun. (2008). Sound Alike Name Matching for Myanmar Language: Journal of World Academy of Science, Engineering and Technology, vol. 40, no. 56, pp.339-343.

[78] Zhai, Cheng Xiang. (2009). Statistical Language Models for Information Retrieval: Synthesis Lectures on Human Language Technologies. The Morgan & Claypool Publishers series. pp. 73.

[79] Zhiwei, Feng. (2004). Standardization of Chinese Scientific Loanwords: In proceedings of the 11th International Symposium of the National Institute for Japanese Language, pp. 71-77.

[80] Zobel, Justin. and Philip Dart. (1996). Phonetic String Matching: Lessons from Information Retrieval: In proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 166 – 172.

Vita

Name: Ohnmar Htun @ Ohnmar Htun Pe

Permanent address: No. D-6/A, Thazin-2 Street, Hlaing Yadanar
Duplex Housing Estate, Hlaing Township, Yangon
11051, Myanmar.

Email: ohnmar@iuj.ac.jp;
ohnmar@cytroncomputing.com

PUBLICATIONS/REFERENCES

Journal

[1] Ohnmar Htun, Andrew Finch, Eiichiro Sumita and Yoshiki Mikami, "Improving Transliteration Mining by Integrating Expert Knowledge with Statistical Approaches", International Journal of Computer Applications 58(17):12-22, November 2012.

[2] Ohnmar Htun, Shigeaki Kodama, Yoshiki Mikami, "Cross-language Phonetic Similarity Measure on Terms Appeared in Asian Languages", International Journal of Intelligent Information Processing 2(2):9-21, June 2011.

International Conference

[1] Andrew Finch, Ohnmar Htun, Eiichiro Sumita, "The NICT Translation System for IWSLT 2012", In proceeding of the International Workshop on Spoken Language Translation (IWSLT), pp.121-125, 2012.

[2] Ohnmar Htun, Shigeaki Kodama, Yoshiki MIKAMI, “Measuring Phonetic Similarity in Myanmar IDNs”, In proceeding of the 8th International Conference on Computer Applications, pp. 129-135, 2010.

[3] Ohnmar Htun, Shigeaki Kodama, Yoshiki Mikami, "Analysis of Terminology Terms in Multilingual Terminology Dictionary", In proceeding of the 8th International Conference on Computer Applications , pp. 122-128, 2010.

[4] Ohnmar Htun, Maung Maung Thant, Pann Yu Mon, Chew Yew Choong, Yoshiki MIKAMI, "Development of Community based Multilingual Terminology Dictionary on KUI Platform", In proceeding of the 7th International Conference on Computer Applications, pp.11-16, 2009.

[5] Pann Yu Mon, Maung Maung Thant, Ohnmar Htun, San Ko Oo, Yoshiki MIKAMI, "Statistical Analysis of Myanmar Word on the World Wide Web for Search Engine Development", In proceeding of the 7th International Conference on Computer Applications, pp.219-224, 2009.

[6] Jay Rajasekera, Maung Maung Thant, Ohnmar Htun, "Effective Use of Environmental Management Information Systems with Data Crawling Techniques", In proceeding of 11th International Conference on Humans and Computers, pp. 263-269, November 20th-23rd, 2008.

[7] Ohnmar Htun, Christopher Tun, "The Future Project Managers in Myanmar: From Tribal PM to Competency-Based PM", International Project and Program Management Symposium Tokyo 2008, Project Management Association of Japan, Tokyo, Japan, Mar. 2008.

Others

[1] Ohnmar Htun, Yoshiki Mikami, "Enhancing Distance-Based Similarity Measure Method", In proceeding of GII-GITS Workshop on Natural Language Processing for Asian Languages, Jointly held between Urano Lab, Waseda University and Mikami Lab, Nagaoka University of Technology, Japan, Oct. 2011.

[2] Ohnmar Htun, Shigeaki Kodama, Yoshiki Mikami, "Cross Language Phonetic Similarity Metrics ",電子情報通信学会信越支部大会IEEE 信越支部セッション, Nagaoka University of Technology, Nagaoka, Japan, p.198, IEEE Shin-etsu session, Oral section, 2010. (*Young Researcher Paper Award*)

[3] Ohnmar Htun, Shigeaki Kodama, Yoshiki Mikami, "Measuring phonetic similarities of words across languages", In proceeding of GII-GITS Workshop on Natural Language Processing for Asian Languages, Jointly held between Urano Lab, Waseda University and Mikami Lab, Nagaoka University of Technology, Waseda University, Tokyo, Japan, Oct. 2009.

[4] Jay Rajasekera, Maung Maung Thant, Ohnmar Htun, "Improving the Timeliness of Environmental Management Information Systems with Data Crawling Techniques", GSIM Working Papers, Series No. IM-2009-06, International University of Japan, Jun. 2009.