© 2015 IJSRSET | Volume 1 | Issue 3 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Converting System of Phonetics Transcriptions to Myanmar Text Using N-Grams Language Models

Kyaw Kyaw Maung
University of Computer Studies, Mandalay, Mandalay Division, Myanmar

ABSTRACT

Converting between Phonetics transcriptions and Myanmar text is a process of converting between the sequence of Phonetics transcriptions and Myanmar text. Phonetics transcription is based on the pronunciation of the language and the Myanmar text is based on the written language. One Phonetics alphabet can be represented many possible forms in written language that leads into word sense ambiguity problem. Another problem is that both of the Phonetics transcriptions and Myanmar text have no space to identify the boundary of syllables and words. This problem can be defined as segmentation problem for matching and mapping between Phonetics transcriptions and Myanmar text. To solve the word-sense ambiguity problem, the research developed n-grams language models from correct training data in Myanmar language. By using these trained n-grams language models, the system can be converted from Phonetics to Myanmar text. Instead of computing the probability on the trained n-grams data, the system matched the input data and the trained n-grams model data. The system has built n-grams models where unigram model, bi-grams model, trigrams model, 4-grams models and 5-grams models to train and convert between Phonetics and Myanmar text. To solve the segmentation problem, the system needed to break the input text into individual tokens. In the system, each token may be represented the consonant, or consonant clusters or vowels. To segment the input text Myanmar text or Phonetics transcriptions correctly, the proposed used the Unicode fonts for both Myanmar text and Phonetics transcriptions.

Keywords: N-grams, Unigram, Bi-grams, Trigrams, 4-grams, 5-grams, Phonetics Transcriptions, Myanmar Text

I. INTRODUCTION

In the linguistic sense, Transcription has been in defined as the process of recording the phonological and/or morphological elements of a language in terms of a specific writing system. Phonetics transcription is the visual representation of speech sounds or phones [1]. The most common type of phonetics transcriptions uses a phonetic alphabet, such as the International Phonetics Alphabets.

Myanmar Language, also known as Burmese, is an official language of Myanmar. All of the researcher who have studies the origin and the development of the Burmese scripts accepted that its source was the Brahmi scripts which flourished in India from 500 B.C to over 300 A.D. Myanmar writing system constructed from consonants, consonants combination symbols, vowels symbols related to the relevant consonants and diacritic marks indicating tone level. Myanmar Language is a type of tonal language, and it using Burmese scripts for writing [2, 3]. Burmese characters are rounded in shape and the script is written from left to right. There is no space between each syllable or each word. The phrases are separated by using space but there are no exact rules to use it. In Myanmar Language, there are 33 consonants, 22 non-nasalized vowels, 21 nasalized vowels and 8 glottal stop vowels. By combining consonant and vowel, a syllable of Myanmar language can be built. Some independent vowels can be constructed a syllable that do not need to combined with any consonant. And combination of consonants called consonant clusters can be combined with vowels to from a Myanmar language

syllable. All of the dependent vowels must be combined with consonant to construct a syllable.

A consonant is a speech sound produced when the either stops or severely constricts the airflows in the vocal track. In articulator phonetics, a consonant is a speech sound that articulated with complete or partial closure of the vocal track. Consonants are all the non-vowel sounds. Consonants correspond to distinct part of a syllable and consonant must be connected with vowels to form a syllable. There are 33 consonants in Myanmar language [4, 5]. The character encoded in Unicode points (U+103B, U+103C, U+103D and U+103E) for Myanmar Language are dependent consonant signs. These signs are combined with another relevant consonant, they become combination of consonants called consonant cluster.

There are 11 basic vowels and 12 extended vowels in Myanmar language. According to the tone level, these vowels can extend to more than 50 vowel sounds. Myanmar vowels can be categorized into three groups: (1) Nasalized Vowels, (2) Non-nasalized Vowels and (3) Glottal Stop Vowels. These vowels are essential part of a syllable [5, 6]. In Nasalized Vowels and Non-nasalized Vowels, tone level may be median, high and low tone level. According to the tone level, the meaning and the pronunciation may be changed in Myanmar language.

Consonants and vowels combine to make a syllable. There is no completely agreed-upon definition of a syllable: in roughly, a syllable is a vowel like sound together with some of the surrounding consonants that are most closely associated with it [9]. In Myanmar language, there are four ways to make a syllable: (1) only Vowel, (2) combination of Consonant/Consonant Clusters and Non-nasalized Vowel, (3) combination of Consonant/Consonant Clusters and Nasalized Vowel and (4) combination of Consonant/Consonant Clusters and Glottal Stop Vowel [3].

Proposed system is a system that converts between the sequence of phonetics symbols and Myanmar text. Phonetics transcription is based on the pronunciations of the language. Myanmar Text is based on the written language. One Phonetic alphabet can be represented many possible forms in written language and this leads to word-sense ambiguity problem. This research proposed a system that develop n-grams language models from correct training data in Myanmar language t solve above problems,. Instead of computing the probability on the trained n-grams data, the system matches the input data and the trained n-grams model. There are three parts of the system: Converting system for Myanmar-Phonetics, training the n-grams models from correct training data, converting system for Phonetics-Myanmar. Accepts the Myanmar text as input to convert into Phonetics and accepts the Phonetics as input, vice versa. The proposed system can be used as speech to text system's language model portion. Unicode fonts for both Phonetics and Myanmar Text are used.

II. RESEARCH OBJECTIVES AND CONTRIBUTIONS

The main objectives of the Phonetics transcriptions to Myanmar text system is to use as part of speech recognition system in Myanmar language. Some Myanmar national races only have spoken language and they have not a written language. The system can help to invent a new language from their spoken language by using Myanmar alphabets. If their language has the same speech sounds of Myanmar language, the system can be used to advice which Myanmar text is most appropriate to represent in written language. The research is intended to develop own segmentation scheme for matching and mapping between Phonetics transcriptions and Myanmar text. And then the system has build the n-grams based language models called unigram, bi-grams, trigrams, 4-grams and 5-grams language models that can be used in speech to text system, text to speech synthesis, spelling correction systems, and so on.

The research contributes the Unicode based segmentation methods for matching and mapping between Phonetics transcriptions and Myanmar text. The system can be classified the types of syllables where it is a Consonant or Consonant cluster or Non-nasalized vowels, Nasalized vowels or Glottal-stop vowels. The proposed segmentation methods can be used in other machine translation researches, NLP researches, Information retrieval systems and text processing system for Phonetics and Myanmar text. The system contributes a computerized mapping between Phonetics symbols and Myanmar alphabets. The proposed system contributes the converting system between Phonetics transcriptions and Myanmar text. The system can be

used as a linguistics tools for linguistics researches for Myanmar language. The proposed system has contributed as n-grams language modelling tool to train Myanmar text. These language models can be used widely in NLP applications, speech recognitions and other text processing systems.

III. LANGUAGE MODELLING AND N-GRAMS LANGUAGE MODELS

A goal of statistical language modelling is to assign a probability to a sentence. In a machine translation system, language modelling is used to distinguish between a correct translation and a bad translation. In a spelling correction system, the language model is used to decide which spelling is more likely to be corrected. In a speech recognition system, the language model can be decided to choose correct speech between the similar speeches of sounds. And language modelling can be used in part-of-speech tagging, parsing, hand writing recognition, summarization, answering the questions, information retrieval systems and so on. The probabilistic relationship among a sequence of words can be directly derived and modelled from the corpora with the so-called stochastic language models, such as n-gram, avoiding the need to create broad coverage formal grammars. Stochastic language models play a critical role in building a working spoken language system.

Most spoken language processing problems can be characterized in a probabilistic framework. One of the most useful concepts in probability theory is that of conditional probability and conditional expectation. The reason is twofold. First, it is often interested in calculating probabilities and expectations when some partial information is available; hence, the desired probabilities and expectations are conditional ones. Secondly, in calculating a desired probability or expectation it is often extremely useful to first "condition" on some appropriate random variable. In conditional Probability, there are two cases for the conditional Probability, discrete case and continuous case.

Probability theory deals with the average of mass phenomena occurring sequentially or simultaneously. Any realistic model of a real-world phenomenon must take into account the possibility of randomness. That is, more often than not, the quantities that interested in will not be predictable in advance but, rather, will exhibit an inherent variation that should be taken into account by the model. This is usually accomplished by allowing the model to be probabilistic in nature. Such a model is, naturally enough, referred to as a probability model.

For a speech recognition system, the input speech signals are continuous in nature. For a spelling correcting system or text to speech synthesis system or this proposed system like phonetics transcriptions to Myanmar Text system, the input tokens are discrete in nature. The leading method for language model is n-gram language modelling. N-gram language models are based on statistical of how likely words are to follow each other. In Language Modelling, the system wants to compute probability of a sentence of a sequence of words. However, most long sequences of words will not occur in the text at all. So, it is need to break down the computation of probability into smaller steps for collecting sufficient statistics and estimate probability distribution. A model that computes either of the probability of words can be defined as a language model.

A leading method of statistical language modelling is n-grams model. The n-grams language model are based on statistical of how likely words are to follow each other. In language modelling, the system wants to compute probability of sentence of a sequence of words. In a spelling correction system, the language model is used to decide which spelling is more likely to be corrected. In a speech recognition system, the language model can be decided to choose a correct speech between the similar speeches of sounds. And language modelling can be used in part-of-speech tagging, parsing, hand writing recognition, summarization, answering the questions, information retrieval systems and so on. The probabilistic relationship among a sequence of words can be directly derived and modelled from the corpora with the so-called stochastic language models, such as n-gram, avoiding the need to create broad coverage formal grammars.

In the research, the n-gram is a contiguous sequence n item from a given sequence of syllable in Myanmar language and Phonetics alphabets. In the research, the n item represents as syllables of Myanmar language and Phonetics transcriptions. The n item of a unigram language model represents only one syllable has used in a unigram model. Bi-gram model is the conditional

probability of the previous one syllable. Trigram model is the conditional probability of the previous two syllables. The 4-grams and 5-grams models are the conditional probability of the previous three syllables and four syllables respectively.

IV. PROPOSED SYSTEM ARCHITECTURE

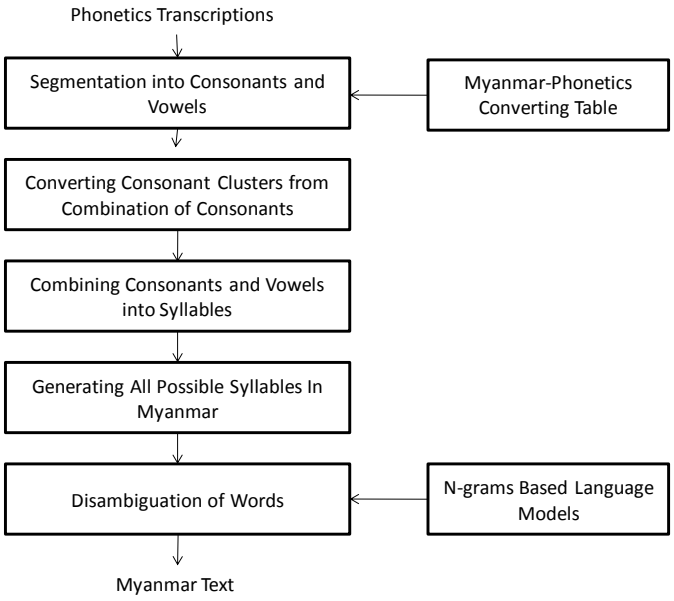

Figure 1: System Flow Diagram for Converting System of Phonetics Transcriptions to Myanmar Text

In the proposed system, there are five steps to convert from phonetics transcriptions to Myanmar text. The first step is to segment of consonants and vowels from input phonetics transcriptions. A phonetics transcription has no space to identify the word boundary. By using Myanmar-Phonetics Converting Table, the input phonetics transcriptions can be identified as consonants and vowels. The detailed segmentation method has been presented in syllable level segmentation between Myanmar text and phonetics transcriptions paper [10]. The second step is to construct consonant clusters by combining of consonants. By combining the relevance consonants can be combined as consonant clusters. In Myanmar language, consonant clusters are important part of the language. The third step is to form a syllable. In Myanmar language, a syllable can be formed by combining the consonant plus vowels or consonants clusters plus vowels. Another possibility is that a syllable may be formed by using only independent vowels. The step four is to generate all possible syllables from input Phonetics transcriptions. One phonetics symbol can be represented many possible Myanmar text. There are 22 speech sounds for 33 Myanmar consonants. One phonetics alphabet for a consonant can be represented one or four Myanmar consonants. So, one speech sounds may be represented many forms of spellings in Myanmar language. It may leads to word sense ambiguity problems. In the fourth step, the proposed system generates many possible forms of syllables in Myanmar language.

Step five is a word sense disambiguation step and it is a core of the proposed system. The proposed system calculates the most possible form of syllable in Myanmar language by using trained n-grams based language model. By using correct data, the proposed system trained uni-gram, bi-gram, trigram, 4-gram and 5-grams in Myanmar language. The data are used from the Web sites such as the President Website and Ministry of Information Website from Myanmar because most of the data from these sites are correct spellings and grammar. These n-grams based language models are syllable level n-grams language models. All possible syllables in Myanmar that generates from step four are matched with trained n-grams language Models and then generate the most appropriate spellings in Myanmar text in final step.

V. EVALUATION RESULTS

The results of converting from phonetics transcriptions to Myanmar text is evaluated by using accuracy, precision and recall. The table presents in the paper is only the sample data from the research. The training data are used from Myanmar president office Web site and used train on the news data. In Table 1, File ID column is represented the test data file name. Other table columns represent as follow:

A. True Positive

True positive is defined as the total numbers of correctly identified and converted from Phonetics transcriptions to Myanmar text. In this case, correctly segmentation is essential for correctly identified from Phonetics to Myanmar text.

TABLE I

EVALUATION RESULTS OF CONVERTING FROM PHONETICS TO MYANMAR TEXT

File ID	Total Syllables	True Positive	True Negative	False Positive	False Negative	Accuracy	Precision	Recall
7677	2914	2899	5	6	4	.9965	.9979	.9986
7702	3864	3832	18	6	5	.9971	.9984	.9986
7721	2322	2293	18	6	5	.9952	.9973	.9978
7720	859	853	4	2	0	.9976	.9976	1
7726	3597	3591	4	2	0	.9994	.9994	1
7734	5092	5077	4	5	6	.9978	.9990	.9988
7737	5239	5233	4	2	0	.9996	.9996	1
7738	1638	1587	40	9	2	.9932	.9943	.9987

B. True Negative

In the research, true negative is defined as the total numbers of correctly identified that the input text is not phonetics. The input test data contains the phonetics text and non phonetics text data.

C. False Positive

False positive is defined as any screening test results that incorrectly detected or classified the input phonetics transcriptions.

D. False Negative

False negative is defined failed to detect phonetics transcriptions and lost of the results.

E. Accuracy

Accuracy is the overall correctness of the system and it can be calculated as the sum of correctly classifications divided the total numbers of classifications. The sum of correctly classifications can be defined as the sum of true positive and true negative. The total number of classification is defined as the sum of true positive, true negative, false positive and false negative.

F. Precision

Precision is a measure of the accuracy provided that the specific class has been predicted. In the research, the precision has computed as true positive is divided by the sum of true positive and false positive.

G. Recall

Recall is a measure of ability of a prediction model to select instances of a certain class from data set. It is also called sensitivity and correspond the true positive rate. In the research, the recall has computed as true positive is divided by the sum of true positive and false negative.

VI. CONCLUSION

The result of the system's accuracy, precision and recall are good. The system used the n-grams language modelling to solve the word sense ambiguity problems that occurs when converting phonetics transcriptions to Myanmar text. The proposed system can be used as part of a speech to text system in Myanmar language. The system cannot be converted some words comes from PALI that is used in Myanmar language.

VII. REFERENCES

[1] A. Kemp, 2006, "Phonetic Transcription: History", University of Edinburgh, Edinburgh, U.K., Elsevier Ltd.

[2] Myanmar Language Commission, 2011, Myanmar-English Dictionary, 11th Edition, University Press, Yangon, Myanmar.

[3] Myanmar Language Commission, 2008, Myanmar-Dictionary, 2nd Edition, University Press, Yangon, Myanmar.

[4] A. Akmajian, R. A. Demers, A. K. Farmer, R. M. Harnish, 2001, "Linguistics, An Introduction to Language and Communication", Fifth Edition, The MIT Press, Cambridge, Massachusetts, London, England.

[5] U. T. Tun, 2007, "Acoustic Phonetics and The Phonology of the Myanmar Language", First Edition, Win Yadanar Press, Yangon, Myanmar.

[6] U. T. Tun, 2012, "The subtleties of the Myanmar Language, Grammar, segments and prosody in the sound system of the language and spelling", First Edition, The Emperor Press, Yangon, Myanmar.

[7] Myanmar Language Commission, Myanmar Grammar, 2005, 30th Year Special Edition, University Press, Yangon, Myanmar.

[8] X. Huang, A. Acero, H. Hon, "Spoken Language Processing, A guide to Theory, Algorithm and System Development", Prentice-Hall, 2001.

[9] A. Kehler, K. V. Linder and N. Ward, "Speech and Language Processing", First Edition, Prentice-Hall, 2000.

[10] K.K.Maung, "Syllable Level Segmentation between Myanmar Text and Phonetics Transcriptions", 2015, Proceeding in International Conference Data Mining, Civil and Mechanical Engineering, ICDMCME'2015, Indonesia.