Syllable Pronunciation Features for Myanmar Grapheme to Phoneme Conversion

[†]Ye Kyaw Thu, [‡§]Win Pa Pa, [†]Andrew Finch, [‡§]Aye Mya Hlaing, [‡§]Hay Mar Soe Naing,
[†]Eiichiro Sumita, [§]Chiori Hori

[†]Multilingual Translation Lab., NICT, Kyoto, Japan
[§]Spoken Language Communication Lab., NICT, Kyoto, Japan
[‡]Natural Language Processing Lab., UCSY, Myanmar
yekyawthu, andrew.finch, eiichiro.sumita, chiori.hori@nict.go.jp,
winpapa, ayemyahlaing, haymarsoenaing@ucsy.edu.mm

Abstract

Grapheme-to-Phoneme (G2P) conversion is a necessary step for speech synthesis and speech recognition. This paper addresses the problem of grapheme to phoneme conversion for the Myanmar language. In our method, we propose four simple Myanmar syllable pronunciation patterns as features that can be used to augment the models in a Conditional Random Field (CRF) approach to G2P conversion. Our results show that our additional features are able to improve a strong baseline model that does not include them. We found that combination of all four features gave rise to the highest performance for Myanmar language G2P conversion.

1. Introduction

G2P conversion is the task of predicting the pronunciation of words given only the spelling. A grapheme is the smallest semantically distinguishing unit in a written language analogous to the phonemes of spoken languages. The correspondence between graphemes and phonemes of Myanmar language is not as simple as one to one since the relationship between syllables and pronunciations is context dependent, and there are many exceptional cases. In particular, we believe the most important information for G2P conversion is to be found in dependencies between adjacent syllables.

In this paper, Myanmar pronunciation patterns are discussed with examples. The proposed method exploits these patterns directly, incorporating them as features within CRF models of pronunciation.

This paper also investigates how each feature and combination of features work with CRF models and shows the improvement of using these features over baseline CRF models with a typical n-gram feature set.

The remainder of the paper is structured as follows. Section 3 describes grapheme to phoneme mapping, Section 4 explains the building of the phonetic dictionary, Sections 5 and 6 explains how syllables interact with each other during pronunciation, Section 7 describes the CRF model used for G2P conversion, Sections 8 and 9 describe the experiments and the results, Section 10 provide the discussion and we conclude in Section 11.

2. Related Work

As far as the authors are aware there have been only one published methodology for

Myanmar language G2P conversion. It was dictionary based approach and analyzed only on pure Myanmar syllables and not considered Pali or subscript consonants [1]. Since it was a dictionary-based approach, out of vocabulary word (OOV) problem is also the drawback. For other languages such as English, several techniques already proposed and these techniques can be roughly classified into rule-based, data-driven and statistical methods [2][3][4][5].

3. Grapheme to Phoneme Mapping

This work uses the Myanmar Language Commission (MLC) Pronunciation Dictionary as a basis for pronunciation mapping [6]. However, we found it necessary to extend the dictionary, with foreign pronunciations. We also needed to modify some of the mappings to ensure consistency of syllable order, and to facilitate mapping the syllables to the International Phonetic Alphabet (IPA). In the proposed mapping table there are 23 phonetic symbols for 33 consonants (some consonants share the same pronunciation, for example “ဒ”, “ဓ”, “ဍ” and “ဎ” in Table 1), 87 vowel combinations and 20 special symbols for foreign pronunciations.

Table 1. Groups of Myanmar Consonants

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

Table 1 shows the groups of characters according to their pronunciation, the groups are: unaspirated, aspirated, voiced and nasal. Many Myanmar syllables containing unaspirated and aspirated consonants are pronounced as voiced consonants depending on the neighboring context. The proposed set of Myanmar pronunciation features were designed to allow statistical g2p conversion models to take these dependencies into account.

Table 2. Examples of phonetic mapping

Myanmar	MLC	Our Mapping
ကျ	kj	ky
ချ	ch	ch
ဂျ	gj	gy
ည	nj	nj
ရှ	sh	sh
၍	jwei.	jwei.
၌	hnai'	nhai'
၎င်း	la̲gaun:	la-gaun:
(စ်)	not defined	S
(ခ်)	not defined	KH
(ပ်)	not defined	P

The main difference between the mapping used in the MLC dictionary and our phoneme mapping is that our mapping produces sequences of phonemes in the same order as they are spoken. For example, the Myanmar single syllable word “လှ” is written as the sequence {လ, ှ}. And thus, our phoneme mapping for this word is “lha.” (IPA: “/l̥/”, where, l corresponds to “လ”, h to “ှ” and “a.” represents the diphthong sound). However, the MLC mapping is “hla.” which is inconsistent with the order in which the characters are pronounced (“l” is a consonant, and “h” is a vowel, and within Myanmar syllables, vowels are always pronounced after consonants). Furthermore, this mapping causes an ambiguity in speech to text processing, since “h” can also express a consonant that is pronounced like a vowel. We also extend the

MLC dictionary to include phoneme mappings for foreign words. For example, the Myanmar phonetic representation of the foreign name Alex (အဲလက်(စ်)) is "e:le'S" (here, S is for "(စ်)"). Some mapping examples from the MLC dictionary and the corresponding mappings resulting from the proposed method are shown in Table 2.

4. Building the Phonetic Dictionary

We built the phonetic dictionary used for training the statistical g2p conversion model by modifying the MLC phonetic dictionary to make it conform to the principles set out in the previous sections. The following steps were applied to the dictionary in order:

1. Words from MLC dictionary were broken into syllables using a heuristic approach, which is 100% accurate [7].

2. Syllables were aligned to their phonemes using a combination of rules and human annotation. Initially, single syllable words were used to align by exact match on the phoneme sequences. This was sufficient to unambiguously align about 80% of the words in the dictionary. The alignment of the remainder was completed manually. For example, က/ka. က/ga- are both single syllable words. The word ကစား/ga-za: can be aligned as: က/ga- စား/za:. The alignment of "စား" being unambiguous given the alignment of "က".

3. Map MLC phonemes to the proposed phoneme set using a manually prepared conversion table.

5. Contextually Independent Pronunciation of Syllables

In this section we will briefly explain how the pronunciation of Myanmar syllables is normally derived from orthographic structure. The pronunciations of consonants when they are combined with vowels are shown in Table 3. Myanmar syllables are generally composed of sequences of consonants and (zero or more) vowel combinations starting with a consonant. Here, vowel combinations can be single vowel, sequences of vowels and sequences of vowels starting with a consonant that modifies the pronunciation of the first vowel. Some examples of the 87 pronunciations of the (>87) vowel combinations are shown in Table 3.

Table 3. Examples of vowel combinations

အီ	i	အိ	i.	အီး	i:	အစ်	i'	အင်	in	အင့်	in.	အင်း	in:
အေ	ei	အေ့	ei.	အေး	ei:	အိတ်	ei'	အိန်	ein	အိန့်	ein.	အိန်း	ein:
အယ်	e	အယ့်	e.	အဲ	e:	အိုက်	ai'	အိုင်	ain	အိုင့်	ain.	အိုင်း	ain:
အာ	a	အာ့	a.	အား	a:	အတ်	a'	အန်	an	အန့်	an.	အန်း	an:
အော်	o	အော့	o.	အော	o:	အောက်	au'	အောင်	aun	အောင့်	aun.	အောင်း	aun:
အူ	u	အု	u.	အူး	u:	အွတ်	u'	အွန်	un	အွန့်	un.	အွန်း	un:
အို	ou	အို့	ou.	အိုး	ou:	အုပ်	ou'	အုန်	oun	အုန့်	oun:	အုန်း	oun:

Syllables are composed of consonant and vowel combinations. In general, the pronunciation of syllables can be obtained directly from the pronunciation of these components. All of the pronunciations of consonants are shown in Table 1, and some example pronunciations of vowel combinations are shown in Table 3. The pronunciation of the full syllable is a concatenation of the pronunciations of each component. These pronunciations do not modify each other. In this paper, we will refer to this type of pronunciation as the "standard pronunciation" of the syllable. Following are some examples of the standard pronunciation:

စာ => s+a

စက် => s + e'

ချိုင့် => ch+ain.

6. Contextual Dependency Among Syllables

Some Myanmar syllables do not conform to these standard rules of pronunciation. The pronunciation of the syllables can depend on the context of syllables. In this section we will briefly present a set of patterns that capture the dependencies among syllables in Myanmar language. In [8] 10 patterns are proposed to capture these dependencies. In this work we consider only a subset of 4 patterns, which cover most cases. These are the first four patterns described briefly below. The remaining patterns either require information (such as POS tags) that is scarce for Myanmar, or occur rarely in the corpus.

6.1. Pattern 1

If the vowel combination of first syllable is ‘င်/in/aun’ or ‘ဉ်/in’ or ‘န်/an/un’ or ‘မ်/an/ein’ or ‘ယ်/e’ or ‘ဲ/e:’ or ‘ံ/an’, and the consonant of second syllable is an unaspirated or an aspirated consonant that is unvoiced, then that second syllable’s pronunciation is voiced. Example pronunciations of some words are as follows:

တောင် ပြုံး (taun pjoun:) => taun bjoun:

ပိုင်း ခြေ (pain: chei) => pain: gyei

There are many exceptions (as an approximate indication, more than half of the cases in our training data were exceptions) and the following are examples in which the next unvoiced syllable does not change to voiced. The following two words pronounced as their standard pronunciations.

ရင်ခုန် jin khoun

စည်းကမ်း si: kan:

6.2. Pattern 2

If the first syllable is a stopped syllable (glottal stop) terminated with any of four vowel combinations (က်၊, စ်၊, တ်၊, ပ်), then the second syllable is pronounced independently from the first in the standard manner.

For example:

စက် သီး => se' thi:

The first syllable စက် terminated with က် vowel combination, the second syllable သီး is pronounced, as it’s standard pronunciation thi:

လွှတ် တော် => lhu' to

The original တ-t does not change to voiced group.

Approximately 50% of all cases are exceptions, for example the following case where the first syllable of a word ends with တ်/a', but the second syllable’s pronunciation is changed to voiced.

နတ် က တော် (na' ka. to) => na' ga- do

6.3. Pattern 3

The pronunciation of certain vowel combinations is occasionally non-standard, for example: the vowel "အု/u.", when preceded by "k/က", "kh/ခ", "p/ပ", "b/ဘ", "m/မ", "l/လ" and "th/သ" is usually pronounced "a-"/"အ".

For example:

ကု ဋေ ku. dei => "ga- dei"
ပု လဲ pu. le: => "pa- le:"
ဘု ရင် bu. jin => "ba- jin"

Although there are 32 consonants in Myanmar (see Table 1), only above 7 consonants usually follow this pattern.

6.4. Pattern 4

A syllable's pronunciation can be changed if the syllable before it was changed. Aspirated and unaspirated syllables typically change to their voiced forms. This phenomenon can cause a cascade of changes that can affect several syllables.

For example:
ကွင်း ဆက် (kwin: hse') => gwin: ze'
စ ကား (sa. ka:) => za- ga:
တ ပို့ တွဲ (ta. pou. twe:) => da- bou. dwe:

There are some exceptions in which the second syllable remains unchanged, and the following are examples:

ချိတ် ဆက် (chei' hse') => gyei' hse'
ခေါင်း ခေါက်(khaun: khau') => gaun: khau'

6.5. Pattern 5

Pattern 5 relates to compound words. If a noun and verb combine to form a noun phrase, the final syllable (unaspirated or aspirated) is changed to the voiced form.

For example:
ထ မင်း ချက် (hta min: che') => hta min: gye'
စာ ရင်း စစ် (sa jin: si') => sa- jin: zi'

6.6. Pattern 6

Some of the syllable pronunciations are changed to voiced sound depending on context or on their part of speech (POS).

The pronunciation of ချိုင့် သော ချိုင့် is "chain. dho: gjain.". This word contains three syllables (ချိုင့်, သော and ချိုင့်). Here, the first syllable and the last syllable are the same spelling but pronounced differently. This is because they are differ semantically and also in their part of speech, as follows:

ချိုင့် (concave, Adjective), သော (a particle suffixed to a word to form an adjective), ချိုင့် (depression in the ground, Noun)

6.7. Pattern 7

Sometimes nasal vowels are omitted in pronunciation and sometimes added.

For example:
မင် အိုး (min ou:) => mhin ou:
Note: a nasal vowel (h) is added here.

ကိုယ် စား လှယ် (kou sa: lhe) => kou za- le
Note: a nasal vowel (h) is omitted in the last syllable.

ပ ညာ (pa. nja) => pjin nja
Note: a nasal vowel (jin) is added to the first syllable.

6.8. Pattern 8

Some syllable pronunciation durations are halved in length, for example: အာ/a, အား/a:, အိ/i., အီ/i, အူ/u, အင်း/in:, အက်/e' may be changed to "အ/a-".

For example:

ဒ ကာ မ (da. ka ma.) => da- ga- ma.

ဓား မ (da: ma.) => da- ma.

ပိ တောက် (pi. tau') => ba- dau'

သိ ချင်း (thi. chin:) => tha- chin:

သူ ဌေး (thu htei:) => tha- htei:

ဆံ ပင် (hsan pin) => za- bin

ထန်း ရည် (htan: jei) => hta- jei

လင်း ကွင်း (lin: kwin:) => la- gwin:

လက် ဖက် (le' hpe') => la- hpe'

7. Modeling with CRFs

MLC dictionary words were mapped to our phoneme symbols to create the training data for building a baseline CRF model. We used the CRFsuite tool [9] for training and testing CRF models. Our proposed method augments the baseline model with features designed to capture the dependencies between syllables manifested in patterns 1-4. The features are numbered the same as their respective patterns.

For each feature, we labeled syllables with tags from the following set {0, c, w}. Here, 0 indicates that the pattern is not applicable to this syllable; c indicates that the syllable was changed from standard pronunciation by the pattern; and w indicates that the pronunciation of this syllable an exception to the pattern.

7.1. Baseline Feature Set

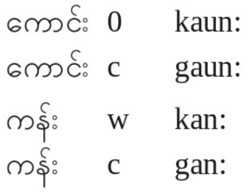
The baseline feature set consisted of unigrams and bigrams of syllables, and unigrams, bigrams and trigrams of pronunciation change labels for each feature, and is shown below:

⇨ s[t-2], s[t-1], s[t], s[t+1], s[t+2]
⇨ s[t-1]|s[t], s[t]|s[t+1]
⇨ l[t-2], l[t-1], l[t], l[t+1], l[t+2]
⇨ l[t-2]|l[t-1], l[t-1]|l[t], l[t]|l[t+1], l[t+1]|l[t+2]
⇨ l[t-2]|l[t-1]|l[t], l[t-1]|l[t]|l[t+1], l[t]|l[t+1]|l[t+2]

Where s[t] is the syllable at position t (t being the position of the syllable being labeled), and l[t] is the label at position t; and s[t-1]s[t] is a bigram of syllables, and so on. In addition the model includes transition features for up to bigrams of phonemes.

7.2. Labeling the Proposed Feature Set

Figure 1 is an example of how we labeled the syllables of a word for feature 1. The first syllable "ကောင်း" is labeled 0 because there is no previous syllable and therefore pattern 1 does not apply. The second syllable is an instance of pattern 1 because the first syllable ends in the required vowel combination and the second syllable has changed from unaspirated to voiced sound. The third syllable also fits pattern 1 but has the standard pronunciation, and is therefore labeled an exception. The final syllable is another correct example of pattern 1.



ကောင်း	0	kaun:
ကောင်း	c	gaun:
ကန်း	w	kan:
ကန်း	c	gan:

Figure 1. A 4-syllable word with feature 1 labeled for each syllable

7.3. Combination of Features 1 to 4

A word can be composed of one or more syllables and it can belong to more than one pattern. Figure 2 shows a word that contains 1 to

4 columns for features 1 to 4. The first syllable has 0 0 0 c values for the pronunciation change labels of the four features; this means that feature 1, feature 2, feature 3 are not applicable to this syllable and feature 4 takes the value 'c' indicating a pattern 4 pronunciation change. The first and fourth features of second syllable follow patterns 1 and 4, but the third syllable's first feature 'w' is covered by pattern 4, and leads to no pronunciation change.

ခံ	0 0 0 c	ga-
တွင်း	c 0 0 c	dwin:
ကောင်း	w 0 0 0	kaun:

Figure 2. Example word with 4 features

8. Experiments

8.1. Data

For this experiment the phonetic dictionary (built from the words in the MLC dictionary) consisted of 27,747 words (2,482 syllables, 1,901 phonemes), and it was used to train the CRF models used in all experiments. We used two evaluation data sets. One was used in a closed evaluation (i.e. the evaluation used only words from the MLC dictionary that were in the training data) consisting of a random sample of 2,000 words (1,254 syllables, 1,158 phonemes). The other test set consisted of 414 words (460 syllables, 476 phonemes) retrieved from 500 sentences of the multilingual Basic Travel Expressions Corpus (BTEC3) [10].

8.2. Methodology

We built seven CRF models in total based on the following feature sets: baseline, feature-1, feature-2, feature-3, feaure-4, feature-123 and feature-1234. Here, the baseline model was the model that we built only with words and their tagged phonemes (described in Section 7.1). The Feature-1 model consisted of the baseline features together with pronunciation change labels of pattern 1 (see Figure 1), Similarly, Feature-2 extended the baseline feature set with pronunciation change labels for pattern2, and so on. In addition, in order to test the effectiveness of combining these feature sets, we also built two feature combination models: feature sets 1, 2, and 3 and feature sets 1, 2, 3, and 4.

For the evaluation, we calculated accuracy as a percentage at both the word-level and the phoneme-level as follows:

Accuracy (%) = (*#correct tags/#of total tags)*100*

We tested differences in accuracy between all of the proposed models and the baseline model using the McNemar test at a difference significance level of 0.95.

9. Results

Table 4 shows the accuracy at the word level for all of the models with the closed MLC test data. The results shown in bold face are statistically significantly different to the baseline model.

Table 4. Word accuracy (%)

Models	MLC	BTEC-3
baseline	84.20	75.12
feature-1	**86.05**	76.57
feature-2	**85.40**	75.36
feature-3	84.80	75.36
feature-4	84.85	75.85
feature-123	**87.50**	**76.81**
feature-1234	**87.90**	**77.78**

Table 5 shows the accuracy at the phoneme level for all of the models with the BTEC-3 test data. Again, the results shown in bold face are significantly different to the baseline model.

Table 5. Phoneme accuracy (%)

Models	MLC	BTEC-3
baseline	93.30	87.92
feature-1	**94.15**	**89.04**
feature-2	**93.81**	88.43
feature-3	93.61	88.12
feature-4	93.64	**88.73**
feature-123	**94.83**	**89.34**
feature-1234	**95.04**	**89.85**

We see an improvement in accuracy of both words and phonemes when the proposed features are added to the baseline CRF model for both test sets. As expected, adding combinations of features to the baseline model (feature-123 and feature-1234) gave the largest improvements in accuracy.

10. Discussion

In this discussion, we want to show the phoneme tagging improvement over baseline by adding features in detail. Generally, we can see precision, recall and F1 score improvements on phonemes by applying features as follows:

Table 6. Precision, Recall and F1 comparison between baseline and feature1

Phoneme	Baseline (precision, recall, F1)	Feature 1 (precision, recall, F1)
ga-	0.62, 1.00, 0.76	0.67, 1.00, 0.80
dan:	0.70, 0.88, 0.78	1.00, 0.88, 0.93
zan:	1.00, 0.67, 0.80	1.00, 1.00, 1.00

Table 7. Precision, Recall, F1 comparison between baseline and feature-2

Phoneme	Baseline (precision, recall, F1)	Feature 2 (precision, recall, F1)
che’	1.00, 0.67, 0.80	1.00, 1.00, 1.00
sa	1.00, 0.83, 0.91	0.92, 0.92, 0.92
tin	1.00, 0.50, 0.67	1.00, 1.00, 1.00

The phoneme “sa” in Table 7, is commonly found in Myanmar words and is usually modified by pattern 2. The result in table 7 show that feature 2 is able to improve the labeling of this phoneme. An example from the test set is the word lei’ sa (လိပ်စာ), in the baseline model the phoneme sa is erroneously labeled as hsa. In the proposed model, using feature 2, the phoneme has been correctly changed to “sa” according to pattern 2.

We found the combination of features feature-1234 can predict some phonemes correctly that other single features can’t do. An example from the test set is the word da- ge lou. (တကယ်လို့) is erroneously labeled as “ta- ke lou.” in feature-1, feature-2 and feature-4 models. Feature-3 also erroneously labeled it as “da- ke lou.” but feature-1234 labeled correctly. Furthermore, by using contextual dependency patterns among syllables as features can predict unknown tags. For example, (null) tag of ja dhi (null) du. (ရာသီဥတု) in the baseline can be predicted by features.

11. Conclusions and Future Work

In this paper we have addressed the problem of grapheme to phoneme conversion for the Myanmar language. Sets of commonly occurring pronunciation patterns were identified, and corresponding sets of features were created to capture the dependencies between phonemes in these patterns. We augmented a respectable n-gram CRF model with the proposed sets of features and evaluated its performance on open and closed test sets relative to the baseline model. The results show that the new features can substantially improve the accuracy of grapheme to phoneme conversion. We analyzed the output of the baseline and proposed systems at the syllable level and show that the proposed approach improves the conversion of those syllables specifically targeted by the new feature sets.

In future work we hope to extend the data sets available. Adding POS-tagged data will allow us create features for more pronunciation patterns. We also plan to study the effects of longer-range context on pronunciation, and intend to extend the approach to compound words, and work towards building a sentence-level model of grapheme-to-phoneme conversion.

References

[1] Ei Phyu Phyu Soe, "Grapheme-to-Phoneme Conversion for Myanmar Language", the 11th International Conference on Computer Applications (ICCA 2013), pp. 195-200

[2] R. I. Damper, Y. Marchand, M.J. Adamson and K. Gustafson, "A comparison of letter-to-sound conversion techniques for English text-to-speech synthesis", Proceeding of the Institute of Acoustics 20(6), 1999

[3] Black, A., Lenzo, K. and Pagel, V., "Issues in building general letter to sound rules", 3rd ESCA on Speech Synthesis, 1998, pp. 77-80

[4] F. Jelinek, "Statistical Methods for Speech Recognition", MIT Press, 1998

[5] Stanley F. Chen, "Conditional and joint models for grapheme-to-phoneme conversion", Eurospeech 2003, pp. 2033-2036

[6] Department of the Myanmar Language Commission (1993), Myanmar-English Dictionary, Yangon, Ministry of Education

[7] Ye Kyaw Thu, Andrew M. Finch, Yoshinori Sagisaka, and Eiichiro Sumita (2013), "A study of Myanmar and segmentation schemes for statistical machine translation", Processing of the 11th International Conference on Computer Applications, pp. 167-179

[8] Myanmar Thadda (2005), Myanmar Language Commission, Ministry of Education, Myanmar, pp. 362-367

[9] Naoki Okazaki, "CRFsuite: a fast implementation of Conditional Random Fields (CRFs)", http://www.chokkan.org/software/crfsuite/, 2007

[10] Genichiro Kikui, Seiichi Yamamoto, Toshiyuki Takezawa, and Eiichiro Sumita (2006), "Comparative study on corpora for speech translation", In IEEE Transactions on Audio, Speech and Language, 14(5), pp. 1674-1682