Comparison of Grapheme–to–Phoneme Conversion Methods on a Myanmar Pronunciation Dictionary

Ye Kyaw Thu[λ,†], **Win Pa Pa**[‡], **Yoshinori Sagisaka**[†] **and Naoto Iwahashi**[λ]
[λ]Artificial Intelligence Lab., Okayama Prefectural University, Japan
[‡]Natural Language Processing Lab., University of Computer Studies Yangon, Myanmar
[†]Language and Speech Science Research Lab., Waseda University, Japan
{ye, iwahashi}@c.oka-pu.ac.jp, winpapa@ucsy.edu.mm, ysagisaka@gmail.com

Abstract

Grapheme-to-Phoneme (G2P) conversion is the task of predicting the pronunciation of a word given its graphemic or written form. It is a highly important part of both automatic speech recognition (ASR) and text-to-speech (TTS) systems. In this paper, we evaluate seven G2P conversion approaches: Adaptive Regularization of Weight Vectors (AROW) based structured learning (S-AROW), Conditional Random Field (CRF), Joint-sequence models (JSM), phrase-based statistical machine translation (PBSMT), Recurrent Neural Network (RNN), Support Vector Machine (SVM) based point-wise classification, Weighted Finite-state Transducers (WFST) on a manually tagged Myanmar phoneme dictionary. The G2P bootstrapping experimental results were measured with both automatic phoneme error rate (PER) calculation and also manual checking in terms of voiced/unvoiced, tones, consonant and vowel errors. The result shows that CRF, PBSMT and WFST approaches are the best performing methods for G2P conversion on Myanmar language.

1 Introduction

Grapheme-to-Phoneme (G2P) conversion models are important for natural language processing (NLP), automatic speech recognition (ASR) and text-to-speech (TTS) developments. Although many machine learning approaches are applicable for G2P conversion, most of them are supervised learning approaches and as a prerequisite we have to prepare clean annotated training data and this is costly. As a consequence, G2P models are rarely available for under-resourced languages such as South and Southeast Asian languages. In practice, we need to perform bootstrapping or active learning with a small manually annotated G2P dictionary for efficient development of G2P converters. In this paper, we examine seven G2P conversion methodologies for incremental training with a small Myanmar language G2P lexicon. We used automatic evaluation in the form of phoneme error rate (PER) and also manually evaluated Myanmar language specific errors such as inappropriate voiced to unvoiced conversion and tones, on syllable units.

2 G2P Conversion for Myanmar Language

Myanmar language (Burmese) is one of the under-resourced Southeast Asian languages for NLP. It has SOV (Subject–Object–Verb) typology and syntactically is quite similar to Japanese and Korean in that functional morphemes succeed content morphemes, and verb phrases succeed noun phrases. In Myanmar text, words composed of single or multiple syllables are usually not separated by white space. Although spaces are used for separating phrases for easier reading, it is not strictly necessary, and these spaces are rarely used in short sentences. In this paper, we only consider phonetic conversion of syllables within words for G2P bootstrapping with a dictionary. 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 vowels, sequences of vowels and sequences of vowels starting with a consonant that modifies the pronunciation of the first vowel. Some examples of Myanmar vowel combinations are အင်း(in:), အိန်း(ein:), အိုင်း(ain:), အန်း(an:) and ဧအောင်း(aun:). The relationship between words and the pronunciation of Myanmar language is not completely consistent, ambiguous, and context dependent, depending on adjacent syllables. Moreover, there are many exceptional cases and rules that present difficulties for G2P conversion (Ye Kyaw Thu et al., 2015a).

Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing,
pages 11–22, Osaka, Japan, December 11-17 2016.

Some Myanmar syllables can be pronounced in more than 4 ways depending on the context and Part-of-Speech (POS) of the syllable. As an example, consider the pronunciation of the two-syllable word ရောင်းဝယ် (meaning trade) with corresponding standard pronunciation of its syllables "ရောင်း" (pronunciation: jaun:) and "ဝယ်" (pronunciation: we). This is a simple pronunciation pattern of a Myanmar word and it has no pronunciation change (i.e. jaun: + we => jaun:). However, many pronunciations of syllables are changed depending on their combination such as in the Myanmar word မေတ္တာ (မေတ် syllable + တာ syllable), love in English; the pronunciation changes from "mi' + ta" to "mji' + ta", နားရွက် (နား syllable + ရွက် syllable) , ear in English; the pronunciation changes from "na: + jwe'" to "na- + jwe'" .

POS is also a factor for pronunciation. The Myanmar word ထမင်းချက် can be pronounced in two ways; "hta- min: che'" when used as a **verb** "cook rice" and "hta- min: gye'" when used as a **noun** "a cook". In another example, the three syllable Myanmar word စာရင်းစစ် can be pronounced "sa jin: si'" when used to mean **verb** "audit" or "sa- jin: zi'" when used to mean a **noun** "auditor"; the single-syllable Myanmar word ချိုင့် can be pronounced "chein". for usage as an adjective "dented" or can be pronounced "gyein.". when used as a noun meaning "food carrier"; one syllable Myanmar word ချေ can be pronounced "gyi" when used as a noun meaning "barking deer" or can be pronounced "chei" when used as a verb.

The most common pronunciation change of Myanmar syllables is unvoiced to voiced and it is contextually dependent, for example the change from: "pi. tau'" to "ba- dau'" for the word ပိတောက် (Pterocarpus macrocarpus flower) , "pja. tin: pau'" to "ba- din: bau'" for ပြတင်းပေါက် (window) word. Some same syllables within a word can be pronounced differently, for example, the Myanmar consonant က pronounced "ka." and "ga-" for three syllables Myanmar word ကကတစ် "ka. ga- di'" (giant sea perch in English). In some Myanmar words, the pronunciation of a syllable is totally different from its grapheme or spelling such as one old Myanmar name လုလင်ကျော် "lu. lin kyo" pronounced as "na- lin gyo".

3 Related Work

(Davel and Martirosian, 2009) designed a process for the development of pronunciation dictionaries in resource-scarce environments, and applied it to the development of pronunciation dictionaries for ten of the official languages of South Africa. The authors mentioned that it is a means of developing practically usable pronunciation dictionaries with minimal resources. (Schlippe, 2014) proposed efficient methods which contribute to rapid and economic semi-automatic pronunciation dictionary development and evaluated them on English, German, Spanish, Vietnamese, Swahili, and Haitian Creole. A novel modified Expectation-Maximization (EM)-driven G2P sequence alignment algorithm that supports joint-sequence language models, and several decoding solutions using weighted finite-state transducers (WFSTs) was presented in (Novak et al., 2012). G2P conversion using statistical machine translation (SMT) was proposed in (Laurent et al., 2009), (Karanasou and Lamel, 2011). In (Laurent et al., 2009), it is shown that applying SMT gives better results than a joint sequence model-based G2P converter for French. The automatic generation of a pronunciation dictionary is proposed in (Karanasou and Lamel, 2011), and their technique used Moses phrase-based SMT toolkit (Koehn et al., 2007) G2P conversion. (Damper et al., 1999) compared different G2P methods and found that data-driven methods outperform rule-based methods.

As far as the authors are aware, there have been only three published methodologies for Myanmar language G2P conversion. (Ei Phyu Phyu Soe, 2013) proposed a dictionary based approach and analyzed it only on pure Myanmar syllables without considering subscript consonants or Pali words. It is a simple approach with a dictionary that is not able to handle out-of-vocabulary (OOV) words. (Ye Kyaw Thu et al., 2015a) proposed four simple Myanmar syllable pronunciation patterns as features that can be used to augment the models in a CRF approach to G2P conversion. The results show that the new features can substantially improve the accuracy of G2P conversion especially on conversion of syllables specifically targeted by the new feature sets. (Ye Kyaw Thu et al., 2015b) applied a phrase-based SMT (PBSMT) approach to Myanmar

G2P conversion and found that G2P conversion using SMT outperformed a CRF approach, with a considerably faster training time. Their comparison between the CRF and PBSMT models shows that the PBSMT approach can handle pronunciation prediction on new compound words (a common form of OOV) well, and can also handle the influence of neighbouring words on the pronunciation of a word.

4 G2P Conversion Methodologies

In this section, we describe the G2P conversion methodologies used in the experiments in this paper.

4.1 Structured Adaptive Regularization of Weight Vectors (S-AROW)

(Kubo et al., 2014) proposed Structured AROW extending AROW (Crammer et al., 2013) to structured learning for G2P conversion. AROW is an online learning algorithm for binary classification that that has several useful properties: large margin training, confidence weighting, and the capacity to handle non-separable data. To overcome the overfitting problems encountered by competitive methods such as Margin Infused Relaxed Algorithm (MIRA) (Crammer and Singer, 2003) and the Confidence Weighted Algorithm (CW) (Dredze et al., 2008) AROW recasts the terms for the constraint of CW as regularizers. S-AROW is applicable for G2P conversion tasks and has a shorter learning time than MIRA. It also has been shown to have a lower phoneme and word error rate compared to MIRA (Kubo et al., 2014).

4.2 Conditional Random Fields

Linear-chain conditional random Fields (CRFs) (Lafferty et al., 2001) are models that consider dependencies among the predicted segmentation labels that are inherent in the state transitions of finite state sequence models and can incorporate domain knowledge effectively into segmentation. Unlike heuristic methods, they are principled probabilistic finite state models on which exact inference over sequences can be efficiently performed. The model computes the following probability of a label sequence $\mathbf{Y} = \{y_1, ..., y_T\}$ of a particular character string $\mathbf{W} = \{w_1, ..., w_T\}$.

$$P_{\boldsymbol{\lambda}}(\mathbf{Y}|\mathbf{W}) = \frac{1}{Z(\mathbf{W})} exp(\sum_{t=1}^{T} \sum_{k=1}^{|\boldsymbol{\lambda}|} \lambda_k f_k(y_{t-1}, \mathbf{W}, t)) \tag{1}$$

where $Z(\mathbf{W})$ is a normalization term, f_k is a feature function, and $\boldsymbol{\lambda}$ is a feature weight vector.

4.3 Joint-sequence models (JSM)

The joint-sequence models (JSM) approach for G2P was proposed by (Bisani and Ney, 2008) and it is also one of the most popular approaches for G2P conversion. The fundamental idea of JSM is that both the grapheme and phoneme sequences can be generated jointly by means of a sequence of joint units (graphones) which carry both grapheme and phoneme symbols. The goal of the JSM is to find a sequence of Y phonemes, $Q = Q_1^Y = \{q_1, q_2, ..., q_Y\}$, that given by a sequence of X graphemes defined by $G = G_1^X = \{g_1, g_2, ..., g_X\}$. This problem can be describe as the determination of the optimal sequence of phonemes, $\acute{Q}$, that maximizes their conditional probability, Q, given a sequence of graphemes, G:

$$\acute{Q} = \arg\max_{Q} P(Q|G). \tag{2}$$

The calculation for all possible sequences of Q directly from $P(Q|G)$ is difficult and we can express it using Bayes' Rule as follows:

$$\acute{Q} = \arg\max_{Q} P(Q|G) = \arg\max_{Q} \left\{P(G|Q) \cdot P(Q)/P(G)\right\} \tag{3}$$

Here, $P(G)$ is common to all sequences Q. The above equation can be simplified as follows:

$$\acute{Q} = \arg\max_{Q} P(G|Q) \cdot P(Q) \tag{4}$$

4.4 Phrase-based Statistical Machine Translation (PBSMT)

A PBSMT translation model is based on joint phrasal units analogous to graphones (Koehn et al., 2003b), (Och and Marcu, 2003). A phrase-based translation system also includes length models, a language model on the target side, and a re-ordering model (which is typically not used for monotonic transduction such as G2P conversion. The models are integrated within a log-linear framework.

4.5 Recurrent Neural Network (RNN) Encoder-Decoder

The RNN Encoder-Decoder technique for machine translation (Cho et al., 2014), (Bahdanau et al., 2014) is a neural network model that links blocks of Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) in an RNN that encodes the source language and decoder units that generate the target language. The basic architecture of the Encoder-Decoder model includes two networks: one encodes the source sentence into a real-valued vector, and the other decodes the vector into a target sentence. In the case of G2P, input is a sequence of graphemes of a Myanmar word, and the output is a phoneme sequence. For example, G2P conversion for Myanmar word ရွှက်ပုန်းသီး (hidden talent in English), the model takes the graphemes of the source word as input: ရွှက်, ပုန်း, သီး and outputs the target phoneme sequence `jwe'`, `poun:` and `dhi:`, which is terminated by an end-of-sequence token (see Figure 1).

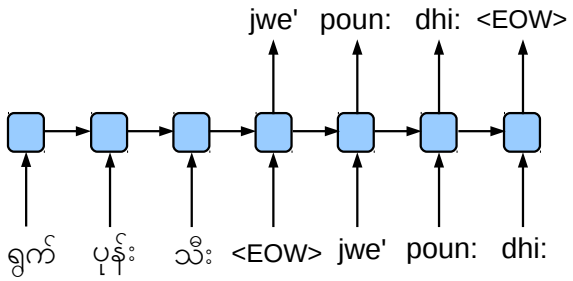


Figure 1: An Architecture of Encoder-Decoder Machine Translation for G2P conversion of Myanmar word ရွှက်ပုန်းသီး (hidden talent in English)

4.6 Support Vector Machine (SVM) based Point-wise classification

Generally, sequence-based pronunciation prediction methods such as (Nagano et al., 2005) require a fully annotated training corpus. To reduce the cost of preparing a fully annotated corpus and also considering possible future work on domain adaptation from the general to the target domain, techniques involving only partial annotation have been developed (Ringger et al., 2007), (Tsuboi et al., 2008). (Neubig and Mori, 2010) proposed the combination of two separate techniques to achieve more efficient corpus annotation: point-wise estimation and word-based annotation. Point-wise estimation assumes that every decision about a segmentation point or word pronunciation is independent from the other decisions (Neubig and Mori, 2010). From this concept, a single annotation model can be trained on single annotated words, even if the surrounding words are not annotated such as ငါ/{ } ကျေးဇူး/{kyei: zu:} တင်ပါတယ်/{tin ba de} (Thank you in English). In this paper, we applied this approach for phonemes of syllables within a word and thus the previous example will change to ငါ/{ } ကျေး/{`kyei:`} ဇူး/{`zu:`} တင်/{`tin`} ပါ/{`ba`} တယ်/{`de`}.

4.7 Weighted Finite-state Transducers (WFST)

(Novak et al.,) introduced a modified WFST-based many-to-many Expectation Maximization (EM) driven alignment algorithm for G2P conversion, and presented preliminary experimental results applying a RNN language model (RNNLM) as an N-best rescoring mechanism for G2P conversion. Their many-to-many approach contained three main modifications to G2P alignment, (1) only many-to-one and one-to-many arcs are trained, (2) a joint WFSA alignment lattice is built from each sequence pair using a log semiring (3) all remaining arcs (including deletion and substitution) are initialized to and constrained to maintain a non-zero weight. This

approach provides EM training to produce better estimation for all possible transitions. The authors applied an RNNLM-based N-best rescoring method to G2P conversion.

5 Experimental Setup

5.1 Data Preparation

In the experiments, we used 25,300 words of Myanmar Language Commission (MLC) Dictionary data (Lwin, 1993). We randomized the original MLC dictionary and prepared 25,000 words for training, 300 words for three open test sets (100 words for each test set) for evaluation. In order to study how the seven G2P approaches behave with varying amounts of training data, we ran a sequence of experiments that trained G2P models from 2,500 words to 25,000 (2393 unique graphemes, 1864 unique pronunciations and 113 unique phonemes) words in increments of 2,500 words. 100 words from the training data also used for closed testing. The G2P mapping is used same mapping proposed by (Ye Kyaw Thu et al., 2015b) and some examples are given in Table 1.

Consonant	**Vowel**	**Independent Vowel**	**Foreign Pronunciation**
က => k	◌ွ: => wa:	ဥ => au.	(ကိ်) => K
ခ => kh	◌ွ့ => wa.	ဦး => u	(ခိ်) => KH
ဂ => g	◌ွေ: => wei:	ဦး => u:	(လိ်) => L
ဃ => gh	◌ွေ့ => wei.	အြ => i.	(စိ်) => S
င => ng	◌ွန် => un	ဤ => i	(ထိ်) => HT

Table 1: An example of grapheme to phoneme mapping for Myanmar language

5.2 Software

We used following open source G2P converters, software frameworks and systems for our G2P experiments:

- Chainer: A framework for neural network development that provides an easy and straightforward way to implement complex deep learning architectures. (Tokui et al., 2015). A deep learning framework developed by Preferred Infrastructure, Inc. (PFI) (`https://preferred.jp/en/`) and Preferred Networks, Inc. (PFN) (`https://www.preferred-networks.jp/en/`). It was released as open source software in June, 2015 (`https://github.com/pfnet/chainer`). Some key features of Chainer are that it is supported as a Python library (PyPI: Chainer) and is able to run on both CUDA with multi-GPU computers. We used the Chainer Python module (version 1.15.0.1) for the G2P conversion experiments based on RNN and RNNA approaches. For both the RNN and the RNNA models, we trained for 100 epochs.

- CRFSuite: We used the CRFsuite tool (version 0.12) (Okazaki, 2007), (`https://github.com/chokkan/crfsuite`) for training and testing CRF models. The main reason was its speed relative to other CRF toolkits.

- KyTea: is a general toolkit (version 0.47) (Neubig and Mori, 2010), (`https://github.com/neubig/kytea`) and it is able to handle word segmentation and tagging. It uses a point-wise classifier-based (SVM or logistic regression) approach and the classifiers are trained with LIBLINEAR (`http://www.csie.ntu.edu.tw/~cjlin/liblinear/`). We used the KyTea toolkit for studying G2P bootstrapping with SVM based point-wise classification for Myanmar language.

- Moses: We used the PBSMT system provided by the Moses toolkit (`http://www.statmt.org/moses/`) for training the PBSMT model for G2P conversion. The word segmented source language was aligned with the word segmented target language using GIZA++ (Och

and Ney, 2000). The alignment was symmetrized by grow-diag-final-and heuristic (Koehn et al., 2003a). The lexicalized reordering model was trained with the msd-bidirectional-fe option (Tillmann, 2004). We used SRILM for training the 5-gram language model with interpolated modified Kneser-Ney discounting (Stolcke, 2002), (Chen and Goodman, 1996). Minimum error rate training (MERT) (Och, 2003) was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1.1). We used default settings of Moses for all experiments.

- Phonetisaurus: A WFST-driven G2P converter (Novak et al., 2012), (`https://github.com/AdolfVonKleist/Phonetisaurus`). Version 0.8a was used. An EM-based many-to-many aligner was applied to grapheme and phoneme sequences (training data) prior to building a G2P model. In the updated version of Phonetisaurus, dictionary alignment is performed with OpenFst (`http://www.openfst.org/twiki/bin/view/FST/WebHome`). In order to estimate an n-gram language model, any language model toolkit such as MITLM (`https://github.com/mitlm/mitlm`)or SRILM (`http://www.speech.sri.com/projects/srilm/`) can be used. We used MITLM toolkit and conversion from ARPA format to a binary FST representation was done with OpenFST.

- Sequitur: A data-driven G2P converter developed at RWTH Aachen University - Department of Computer Science by Maximilian Bisani (Bisani and Ney, 2008). The 2016-04-25 release version (`https://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html`) was used for the JSM G2P conversion experiment.

- Slearp: Structured LEarning And Prediction (Kubo et al., 2014). We used Slearp (version 0.96) (`https://osdn.jp/projects/slearp/`) for S-AROW G2P model building.

We ran all above software with default parameters for building the G2P models. Although feature engineering is usually an important component of machine-learning approaches, the G2P models were built with features from only the grapheme and phoneme parallel data, to allow for a fair comparison between the seven approaches.

5.3 Evaluation

To evaluate the quality of the G2P approaches, we used two evaluation criteria. One is automatic evaluation of phoneme error rate (PER) with SCLITE (score speech recognition system output) program from the NIST scoring toolkit SCTK version 2.4.10 (`http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm`). The other evaluation was done manually by counting voiced/unvoiced, tones, consonant and vowel errors on G2P outputs.

The SCLITE scoring method for calculating the erroneous words in Word Error Rate (WER), is as follows: first make an alignment of the G2P hypothesis (the output from the trained model) and the reference (human transcribed) word strings and then perform a global minimization of the Levenshtein distance function which weights the cost of correct words, insertions (I), deletions (D) and substitutions (S). The formula for WER is as follows:

$$WER = (I + D + S) * 100/N \tag{5}$$

In our case, we trained G2P models with syllable segmented words and thus alignment was done on syllable units and the PER was derived from the Levenshtein distance at the phoneme level rather than the word level. For example, phoneme level of syllable alignment, counting I, D and S for Myanmar word "ချင်းချက်" (exception in English), left column and "စိတ်ပျက်လက်ပျက်" (disappointed in English), right column is as follows:

```
Scores: (#C #S #D #I) 0 2 0 1
REF: *** CHWIN: GYE'
HYP: CHI NWIN: CHE'
Eval:    I     S     S
```

```
Scores: (#C #S #D #I) 2 1 1 0
REF: sei' PJEI le' PJAU'
HYP: sei' PJAUN: le' *****
Eval:       S           D
```

6 Results

6.1 Automatic Evaluation with Phoneme Error Rate (PER)

We used PER to evaluate the performance of G2P conversion. We computed the PER scores using `sclite` (`http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm`) on the hypotheses of G2P models and references. The results are presented in Figure 2 and lower PER is better in performance as we mentioned in Section 5.3. The experimental results also show the learning curve variations of seven G2P conversion approaches on the training data. We can clearly see that there is no significant learning improvement for the SVM based point-wise classification from the evaluation results on both the closed and the three open test sets (see Figure 2, (g)). Also, the PER results of S-AROW, JSM, PBSMT and RNNA on the closed test data are unstable. Each of the graphs show the performance of G2P conversion and the best PER scores (i.e. 0) was achieved on the closed test data by the RNN, S-AROW and WFST. The best PER scores of the CRF and PBSMT on closed test data were 6.4 and 7.5 respectively. On the other hand, the final models of the CRF and WFST achieved the lowest PER scores for all three open test data sets (open1, open2 and open3). A PER score 14.7 for open1 was achieved by WFST, 11.4 for open2, and 15.7 for open3 by both CRF and WFST. An interesting point is that the PBSMT approach achieved close to the lowest PERs for the three open test sets (16.1 for open1, 13.1 for open2 and 22.0 for open3). Figure 2, (e) shows the RNN approach is able to learn to reach zero PER score on the closed test data from epoch two (i.e. with 5,000 words). The PER of RNN is lower than RNNA approach for both the closed and the open test data (see Figure 2, (e) and (f)).

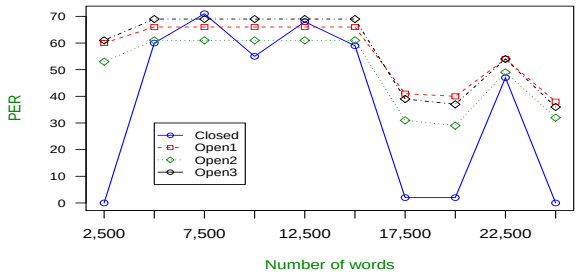
6.2 Manual Evaluation

Manual evaluation was mainly done on the results from the models trained with 25,000 words in terms of errors on voiced/unvoiced pronunciation change, vowel, consonant and tone. The results show that voiced/unvoiced error is the highest among them. (Ye Kyaw Thu et al., 2015a) discussed the importance of the pronunciation change patterns, and our experimental results also show how these patterns affect the G2P performance. Pronunciation error rates for PBSMT and WFST are comparable and the PBSMT approach gives the best performance overall. The SVM based point-wise classification approach produced the highest phoneme errors on unknown words (i.e. UNK tagging for OOV case by KyTea) among the seven G2P approaches. Generally, all methods can handle tone well and we assume that almost all the tonal information of Myanmar graphemes is covered in the training dictionary. The lowest error rate on tone was achieved by PBSMT. From the overall manual evaluation results from train1 (training number 1: trained with 2,500 words) to train10 (training number 2: trained with 25,000 words), we can see clearly that RNN, PBSMT and WFST approaches gradually improve with increasing training data set size. Some difficult pronunciation changes at the consonant level (such as pronunciation prediction from `ljin` to `jin` for the Myanmar word "`kau'jin`", "ကော်ကျင်") can be predicted correctly by the PBSMT approach and the RNN but not by the other approaches. Although the training accuracy of RNN is higher than the other techniques, in the automatic evaluation, some OOV predictions are the worst (refer Table 2).

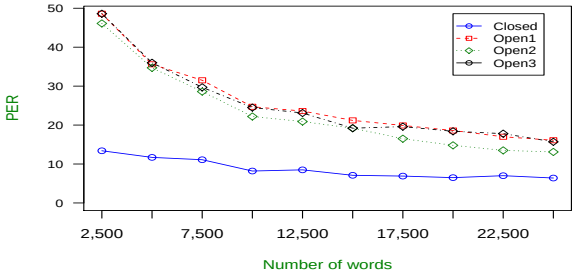
6.3 Discussion

As we presented in the previous section, some evaluation results of the G2P conversion approaches on closed test data are inconsistent especially for S-AROW and JSM (refer Figure 3, (a) and (c)). However all models are generally improve on the three open test evaluation sets. Here we investigate the OOV rates over test data. Figure 3 shows the OOV rate for graphemes of the three open test data sets over the incremental training process from train1 to train10. As expected, the OOV rate gradually decreases as the the training data size increases.

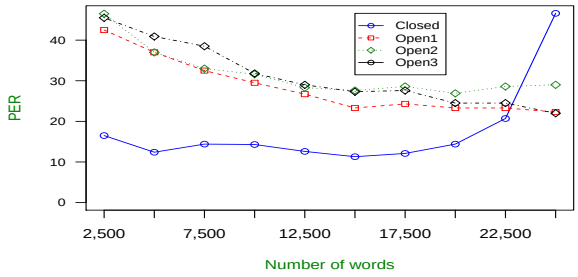
We performed a detailed analysis of each error type by manual evaluation, and the results are shown in Figure 4. From the results, we can clearly see that SVM based point-wise classification produced highest number of voiced/unvoiced errors, and we have already discussed UNK tags or KyTea pronunciation estimation errors in Section 6.2. We now turn to RNN specific errors. RNNs are capable sequence models with high potential for building G2P conversion models and thus we present a detailed error analysis. The RNN produced some reordering errors and the automatic evaluation counts one reordering error as one deletion and one insertion. For example,



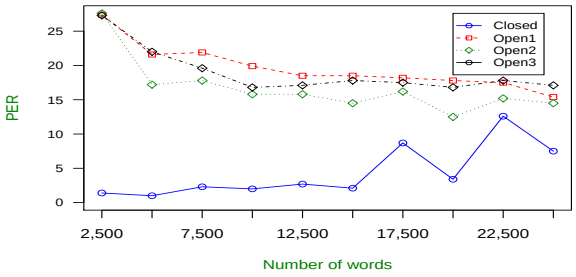
(a) Method: S-AROW, Program: Slearp



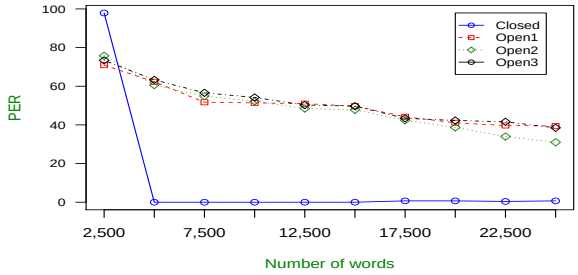
(b) Method: CRF, Program: CRFSuite



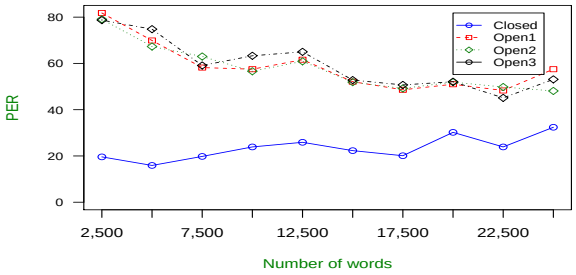
(c) Method: JSM, Program: Sequitur



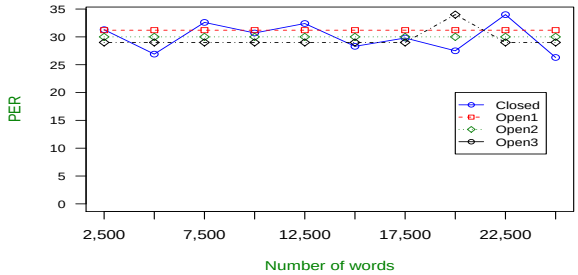
(d) Method: PBSMT, Program: Moses



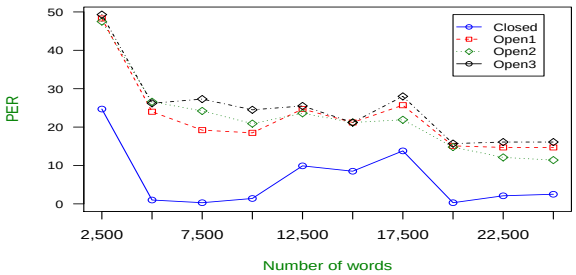
(e) Method: RNN, Program: Chainer



(f) Method: RNNA, Program: Chainer



(g) Method: SVM based point-wise classification, Program: KyTea



(h) Method: WFST, Program: Phonetisaurus

Figure 2: Phoneme Error Rate (PER) of G2P conversion methodologies

Method	Hypothesis	Note on Error
S-AROW	tha' ba. ja. nan baun:	tone error in "ba." and consonant error in "ja."
CRF	tha' ba- ja. nan baun:	consonant error in ja.
JSM	tha' ba. ra- baun:	tone error in "ba." and "ra-" one phoneme deletion
PBSMT	tha' ba. ja- nan baun:	tone error in "ba."
RNN	tha' ba- WA. SA MI:	3 syllables "WA. SA MI:" are predicted and they are far from the correct pronunciation
SVM based point-wise	UNK ba- ja- nan baun:	OOV error
WFST	tha' ba- ra. nan baun:	0 Error

Table 2: An example of phoneme prediction errors of G2P conversion methods.

the RNN model output for the Myanmar word "ထုံဇေဝ", htoun pei BEI (recalcitrantly in English). Its SCLITE alignment and scoring is shown in the left column below:

Scores: (#C #S #D #I) 2 0 1 1	Scores: (#C #S #D #I) 3 1 0 0
REF: *** htoun pei BEI	REF: mwei: tha- MI. gin
HYP: PEI htoun pei ***	HYP: mwei: tha- HPA. gin
Eval: I D	Eval: S

Some RNN pronunciation prediction errors were semantic in nature, and we were surprised to discover them. For example, the RNN model output for the Myanmar word "မွေးသမိခင်", mwei: tha- MI. gin (mother in English) is similar word "မွေးသဖခင်", mwei: tha- HPA. gin (father in English). Similar semantic errors were also produced by the PBSMT approach. Another interesting point is that the RNN and WFST approaches can predict correctly for some rare patterns (i.e. where all syllable pronunciations of a word are changed) even when all other models made errors. For example, the errors for the Myanmar word "စားပွဲခင်း", za- bwe: gin: (tablecloth in English) made by the other approaches were: S-AROW: za- bwe: khin:, JSM: za- bwe: khin:, RNN: za- bwe: gin:, WFST: za- bwe: gin: and SVM based point-wise classification: za- bwe: khin:.

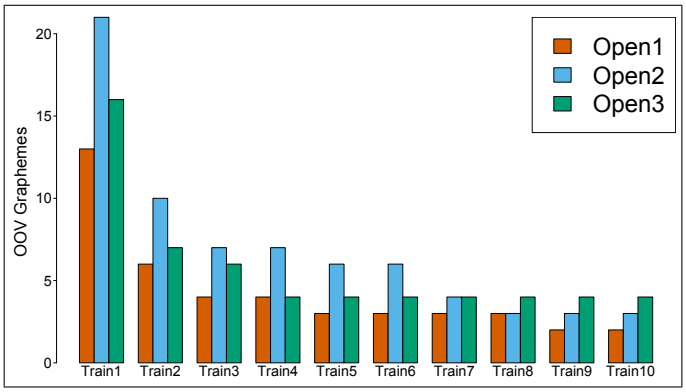


Figure 3: OOV graphemes over incremental training process

7 Conclusion and Future Work

The aim of this work is to show the relative performance of different machine learning techniques on Myanmar G2P conversion. Both automatic evaluation and manual evaluation showed that CRF, Phonetisaurus, SMT and RNN have their own unique advantages when applied to

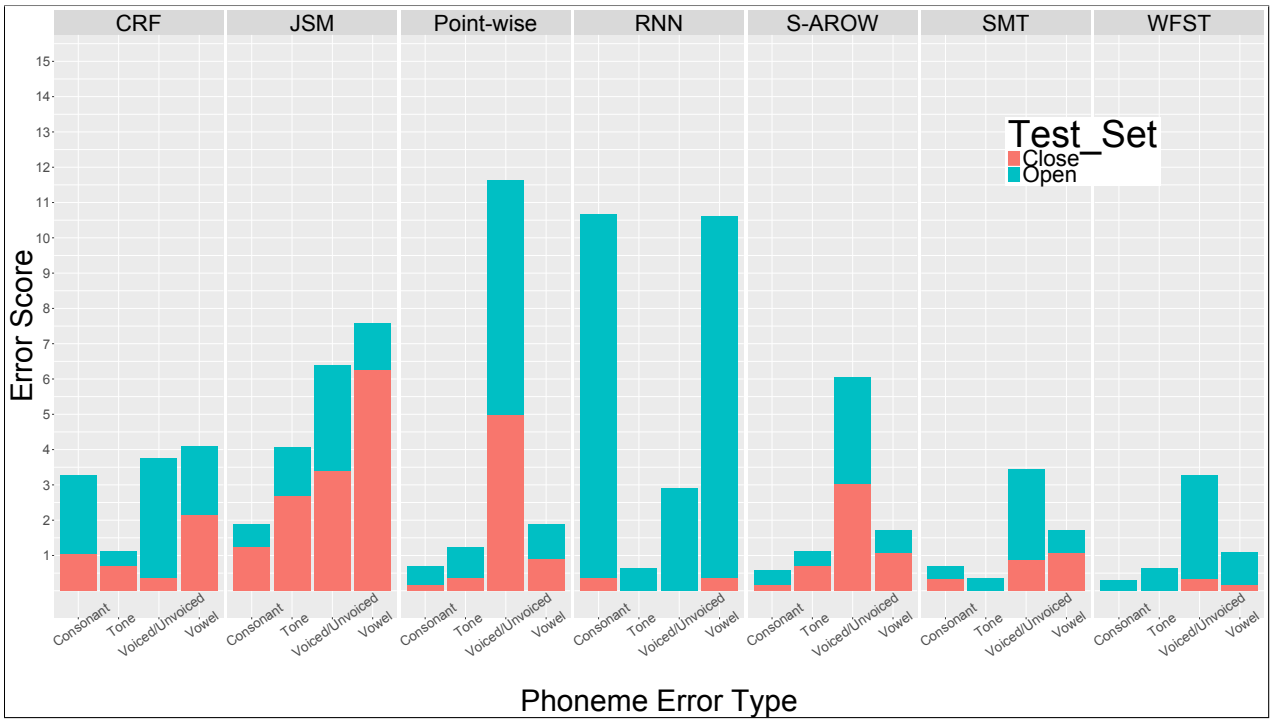


Figure 4: Average error scores of manual checking for G2P conversion methods

Myanmar pronunciation prediction. Although the manual evaluation was expensive, we believe it was necessary in order to analyse these approaches in depth. In summary, our main findings are that the CRF, Phonetisaurus, SMT approaches gave rise to the the lowest error rates on the most important features of Myanmar G2P conversion: voiced/unvoiced, vowel patterns and tone. We plan to find out the performance of these approaches on sentence level since Myanmar pronunciation highly depends on the context.

Acknowledgements

The authors would like to thank Dr. Andrew Finch, Multilingual Translation Lab., Advanced Speech Translation Research and Development Promotion Center, National Institute of Information and Communications Technology (NICT), Japan for valuable comments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Commun.*, 50(5):434–451, May.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the ACL*, pages 310–318. Santa Cruz, California, June.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *J. Mach. Learn. Res.*, 3:951–991, March.

Koby Crammer, Alex Kulesza, and Mark Dredze. 2013. Adaptive regularization of weight vectors. *Machine Learning*, 91(2):155–187.

R.I. Damper, Y. Marchand, M.J. Adamson, and K. Gustafson. 1999. A comparison of letter-to-sound conversion techniques for english text-to-speech synthesis.

Marelie Davel and Olga Martirosian. 2009. Pronunciation dictionary development in resource-scarce environments. In *in Proc. Interspeech*, pages 2851–2854.

Mark Dredze, Koby Crammer, and Fernando Pereira. 2008. Confidence-weighted linear classification. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 264–271, New York, NY, USA. ACM.

Ei Phyu Phyu Soe. 2013. Grapheme–to–phoneme conversion for myanmar language. In *The 11th International Conference on Computer Applications (ICCA2013)*, pages 195–200, Yangon, Myanmar, Feb.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Panagiota Karanasou and Lori Lamel. 2011. Automatic generation of a pronunciation dictionary with rich variation coverage using smt methods. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'11, pages 506–517, Berlin, Heidelberg. Springer-Verlag.

Philipp Koehn, Franz Josef Och, , and Daniel Marcu. 2003a. Statistical phrase-based translation. In *In Proceedings of the Human Language Technology Conference*, Edmonton, Canada.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003b. Statistical phrase-based translation. In *HLT-NAACL*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Keigo Kubo, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2014. Structured adaptive regularization of weight vectors for a robust grapheme-to-phoneme conversion model. *IEICE Transactions on Information and Systems*, E97-D(6):1468–1476, June.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Antoine Laurent, Paul Deléglise, and Sylvain Meignier. 2009. Grapheme to phoneme conversion using an smt system. In *INTERSPEECH*, pages 708–711. ISCA.

San Lwin. 1993. *Myanmar - English Dictionary.* Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar.

Tohru Nagano, Shinsuke Mori, and Masafumi Nishimura. 2005. A stochastic approach to phoneme and accent estimation. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 3293–3296. ISCA.

Graham Neubig and Shinsuke Mori. 2010. Word-based partial annotation for efficient corpus construction. In *The seventh international conference on Language Resources and Evaluation (LREC 2010)*, pages 2723–2727, Malta, May.

Josef R. Novak, Paul R. Dixon, Nobuaki Minematsu, Keikichi Hirose, Chiori Hori, and Hideki Kashioka. Improving wfst-based g2p conversion with alignment constraints and rnnlm n-best rescoring.

Josef R. Novak, Nobuaki Minematsu, and Keikichi Hirose. 2012. Wfst-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing, FSMNLP 2012, Donostia-San Sebastiían, Spain, July 23-25, 2012*, pages 45–49.

Franz Josef Och and Daniel Marcu. 2003. Statistical phrase-based translation. pages 127–133.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *ACL00*, pages 440–447, Hong Kong, China.

Franz J. Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Eric Ringger, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, Kevin Seppi, and Deryle Lonsdale. 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 101–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tim Schlippe. 2014. *Rapid Generation of Pronunciation Dictionaries for new Domains and Languages*. Ph.D. thesis, Uni Karlsruhe.

Lucia Specia. 2011. Tutorial, fundamental and new approaches to statistical machine translation. In *International Conference Recent Advances in Natural Language Processing*.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 101–104, Stroudsburg, PA, USA. Association for Computational Linguistics.

Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*.

Yuta Tsuboi, Hisashi Kashima, Hiroki Oda, Shinsuke Mori, and Yuji Matsumoto. 2008. Training conditional random fields using incomplete annotations. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 897–904, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ye Kyaw Thu, Win Pa Pa, Finch Andrew, Aye Mya Hlaing, Hay Mar Soe Naing, Sumita Eiichiro, and Hori Chiori. 2015a. Syllable pronunciation features for myanmar grapheme to phoneme conversion. In *The 13th International Conference on Computer Applications (ICCA2015)*, pages 161–167, Yangon, Myanmar, Feb.

Ye Kyaw Thu, Win Pa Pa, Finch Andrew, Ni Jinfu, Sumita Eiichiro, and Hori Chiori. 2015b. The application of phrase based statistical machine translation techniques to myanmar grapheme to phoneme conversion. In *The Pacific Association for Computational Linguistics Conference (PACLING2016)*, pages 170–176, Legian, Bali, Indonesia, May.