INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Efficient Thai Grapheme-to-Phoneme Conversion Using CRF-Based Joint Sequence Modeling

Sittipong Saychum, Sarawoot Kongyoung, Anocha Rugchatjaroen, Patcharika Chootrakool, Sawit Kasuriya and Chai Wutiwiwatchai

National Electronics and Computer Technology Center,
National Science and Technology Development Agency,
Pathumthani, Thailand

{sittipong.say, sarawoot.kon, anocha.rug, patcharika.cho,
sawit.kas and chai.wut}@nectec.or.th}

Abstract

This paper presents the successful results of applying joint sequence modeling in Thai grapheme-to-phoneme conversion. The proposed method utilizes Conditional Random Fields (CRFs) in two-stage prediction. The first CRF is used for textual syllable segmentation and syllable type prediction. Graphemes and their corresponding phonemes are then aligned using well-designed many-to-many alignment rules and outputs given by the first CRF. The second CRF, modeling the jointly aligned sequences, efficiently predicts phonemes. The proposed method obviously improves the prediction of *linking syllables*, normally hidden from their textual graphemes. Evaluation results show that the prediction word error rate (WER) of the proposed method reaches 13.66%, which is 11.09% lower than that of the baseline system.

Index Terms: grapheme-to-phoneme conversion, joint sequence modeling, Thai G2P, Conditional Random Fields.

1. Introduction

The complexity of Thai pronunciation and spelling is mostly caused by loan words. The conditional combination of morphemes and syllables are the effect of adoption. Historically speaking, Thai is dominated by Pali, Sanskrit, Chinese, English, Khmer, etc. Therefore, there are many complicated rules applied in writing and reading, even in the same word. Tonglorh [1] extracted and explained the complexity after cultural influences and published the principle of writing and reading Thai in 1982.

Thai writing consists of continuous strings of characters. Separators such as period, comma or punctuation marks are not common. There is only a space, " ", which is used for separating phrases. Text tokenization has always been a significant natural language processing (NLP) problem for Thai [2]. Based on the ambiguous word boundary often found in Thai writing, and the presence of a lot of loan words as described above, Thai grapheme-to-phoneme (G2P) is a challenging task and has been widely researched for many years [4] [5] [6].

Chotimongkol and Black [6] analyzed a pronunciation dictionary and proposed an intelligent Thai orthographic-to-sound converter using a statistical model trained from 22,818 phonemically transcribed words. Ignoring tone errors, one of the most ambiguous problems in Thai G2P, they reached a word error rate of 24.7%. In a year later, Tarsaku et al. [3] proposed using a Probabilistic Generalized LR (PGLR) approach with their Context Free Grammar (CFG) rules for Thai syllable construction, and claimed a word error rate of 27.1%. In 2006, Charoenpornsawat and Schultz [4] introduced their example-based G2P (EBG2P) conversion, aiming at a language-independent approach but using Thai as a case study. They reached a 19.0% word error rate when ignoring tone errors. In the same year, Thangthai et al. [5] also published his idea of syllable induction by finding all possible syllable units in an input string. Syllable sequences containing syllable units not complied with the CFG rules taken from Tarsuku et al. [3] are eliminated. Then probabilistic scores given by a trained syllable n-gram model are assigned to the remaining candidates, and the candidate with the highest score is selected as a result. With their design, a 17.0% word error rate was obtained.

Based on the discussion in Thangthai et al. [5], there is still room to improve Thai G2P by solving three major problems: tone ambiguity, vowel length ambiguity, and prediction of the *linking syllable* which is a syllable hidden in many words and borrowed from Pali-Sanskrit. In this paper, joint sequence modeling proposed for G2P [3] is expected to serve as a more efficient model alleviating the mentioned problems. However, grapheme and phoneme sequence alignment for model training becomes challenging due to character ordering in Thai writing, as well as totally missing the linking syllables. Therefore, this paper aims to introduce a grapheme-phoneme sequence alignment suited for joint sequence modeling. Two-stage Conditional Random Fields (CRFs) are proposed. The first CRF is used for syllable segmentation and syllable type prediction. The assigned syllable type helps in rule-based grapheme-phoneme sequence alignment and the aligned sequences are used to train the second CRF, which finally becomes a phoneme predictor. The next section explains Thai writing and sound systems, and also reviews the problems of Thai G2P found in previous work. Section 3 introduces our proposed method. Section 4 presents experiments and results using our largest Thai pronunciation dictionary [4]. Section 5 concludes and discusses future work.

2. Thai G2P and Its Challenges

This section will briefly describe the basic structure of the Thai language and the difficulties of Thai G2P.

http://dx.doi.org/10.21437/Interspeech.2016-621

2.1. Basics of Thai structure

Thai writing is a spelling system. Basic Thai textual syllables can be represented in the form $\{C_i, V, C_f, T\}$, where C_i, V, C_f and T denote an initial consonant, a vowel, a final consonant, and a tone respectively. The total number of Thai characters and phones is summarized in Table 1. Note that Thai is a tonal language where the meaning of a syllable changes as the syllable tone changes [1].

Table 1: The number of Thai characters and phones.

Type	Character	Phone
Initial consonant (C_i)	44	38
Vowel (V)	16	24
Final consonant (C_f)	37	9
Tone (T)	4	5

2.2. Difficulties of Thai G2P

Since Thai is written without explicit word boundaries, word segmentation becomes the first necessary step. The definition of Thai word boundaries could be ambiguous, which leads to segmentation errors for G2P. Previously proposed Thai G2P systems in [3] and [5] further segment words into smaller units such as syllables, which may result in further errors.

In addition to the basic syllable structure described, more complex and ambiguous syllable forms, caused mainly by loan words, also appear in Thai writing. Some researchers have introduced a Pseudo-syllable (PS) to represent the smallest textual unit [2]. These PS units can make all the contextual units and the syllable itself correctly pronounceable. As there is no ambiguity in PS segmentation, a highly accurate segmentation tool can be made [9]. According to the difficulties of Thai G2P summarized by Thangthai et al. [5], there are three issues that contribute to G2P errors. Table 2 gives examples of these issues, which are described as follows.

2.2.1. Prediction of linking syllable

Generally, the problems of linking syllables are from loan words, such as case 1 and case 2 in Table 2. These cases illustrate words from Pali-Sanskrit. The textual pseudo-syllable "รัฐ" is pronounced as two sound syllables /r á t/tʰ à/, where /tʰ à/ is a sound syllable hidden from its textual form. Case 2 is more complex as the final consonantal character "ก" in the first syllable /c à k/ is shared as an initial consonantal character of the next syllable /kr í:/. One difficulty is that these phenomena do not always occur.

2.2.2. Vowel-length distortion

Thai words differentiate between short and long vowels. Normally, we pronounce words by using textual vowels, but some words do not comply with their textual vowel. For example, in Table 2, even though the textual vowel in case 3 is a long vowel character "า" pronounced as /a:/, it is actually pronounced as a short vowel sound /a/.

2.2.3. Tone ambiguity

This issue often encountered in English loan words as shown in case 4 in Table 2. While the first and the second PS should be pronounced with low and middle tones respectively, their exact pronunciations are replaced with high and falling tones as Thai people attempt to mimic English sounds instead of reading textual forms.

Table 2: Examples of problematic words and their IPA pronunciations.

Case	Word	PS	No. of Syl.	Pronunciation
1	รัฐธรรมนูญ	รัฐ ธรรม นูญ	2 2 1	/r á t/tʰ à/ /tʰ a m/m á/ /n ū: n/
2	จักรี	จักรี	2	/c à k/kr í:/
3	ท่าน	ท่าน	1	/t â n/
4	ด็อกเตอร์	ด็อก เตอร์	1 1	/d ɔ́ k/ /t ɤ̂:/

3. Proposed System

3.1. Fundamental concept

According to the literature review given in the introduction, existing systems rely on different learning machines such as decision trees [6], PGLR [3], and syllable n-gram models [5]. Context Free Grammar (CFG) has often been used at the start of the process to create possible syllable patterns given an input text. Solutions to the difficulties of Thai G2P mentioned in the previous section depend strongly on whether handcrafted CFG rules could cover such complicated phenomena or not. Recently, two major learning machines including Conditional Random Fields (CRFs) [5] [6] and Neural Networks (NN) [7] [8] have successfully been applied to this task. Joint sequence modeling [3] involves tying one or no graphemes together with one or no corresponding phonemes, forming a *graphone* unit. The G2P task then becomes a modeling task for a single sequence of graphones using any type of sequence modeling algorithm.

In addition to the difficulties of Thai G2P, a character-level approach such as the character n-gram model is not feasible to solve the difficulties in Section 2.2. Instead, we applied the concept of graphones to solve Thai G2P problems. Given the definition of a graphone, one-to-one, one-to-many, many-to-one, and many-to-many grapheme-phoneme alignments can be constructed. The linking syllable, which is added between two textual syllables, can then be grouped as graphones as shown in Figure 1. In this figure, all phonemes in the linking syllables /r á t/ and /tʰ à/ are mapped to a null character | $ |. Textual tone marks are also treated as phoneme symbols and are mapped to phonetic tone as shown in the shaded graphones in the figure.

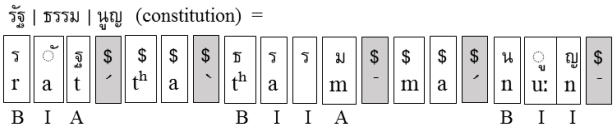

Figure 1: An example of aligned Thai graphones.

3.2. Overall process

Figure 2 illustrates the overall process of the proposed G2P system. The upper part half of the figure is represents system training and the lower part represents testing.

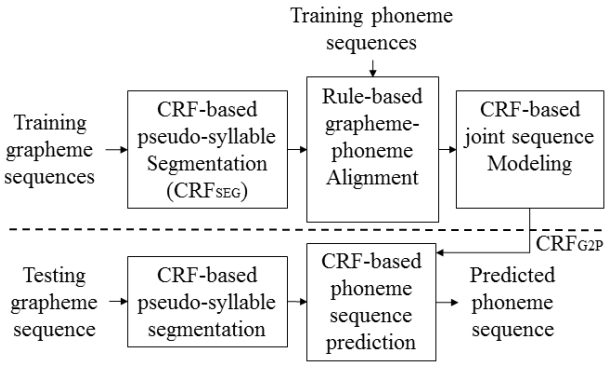

Figure 2: An overall process of the proposed Thai G2P system.

In the training step, training grapheme sequences are first segmented into PS units using a modified version of the previously proposed CRF-based PS segmentation tool [9], hereafter referred to as CRF_{SEG}. CRF_{SEG} takes graphemes and their character classes as input for segmentation. There are four main classes: consonant (C), vowel (V), tone marker (T), and special character (S). Each class can be divided into subclasses. For an example, according to Thai structure in [1], consonants can be subdivided into three subclasses: high-level consonants (C_h), middle-level consonants (C_m), and low-level consonants (C_l). CRF_{SEG} is modified to not only segment text, but also to assign a type to each grapheme character. There are four types of character: 'B' for a PS beginning character, 'I' for a PS internal character, 'A' for a character that could be a linking syllable (e.g. the third character "ฐ" in Figure 1, is a final consonant and also a linking syllable of "รัฐ" pronounced /r á t/t^h à/), and 'S' for a shared character for two connected syllables (e.g. the character "ก" in the second case in Table 2). These assigned character types facilitate the next process to align the training graphemes to their corresponding phonemes as shown in the bottom line of Figure 1.

There are a variety of possible grapheme-phoneme alignment rules. In this paper, the simplest alignment technique, similar to that explained in [3], is first evaluated. The upper half of Figure 3 shows an example of this basic alignment, called joint sequence modeling 1 (JSM1) hereafter. In JSM1, alignment is done character-by-character in the exact order of graphemes, with no grouping of vowels or character re-ordering. Another approach, called joint sequence modeling 2 (JSM2), is shown in the lower half of Figure 3. In this design, multiple graphemes are grouped together as one phone, for example graphemes "เ", " ี", and "ย". They are represented by the vowel "เ ีย" (/i:a/). Each grapheme or group of graphemes is assigned its class, which is similar to that used in the CRF_{SEG}. In JSM2, we introduce some special phonemes in order to improve the accuracy of G2P. For example in Figure 3, /p^hr/ is a phonetic representation of the consonant cluster "พร". We assign both "พ" and "ร" with the same phonetic representation as the consonant cluster /p^hr/ and use /=/ to join them together (/p^hr=/ and /=p^hr/). With this idea, it is easier to predict the phoneme associated with the consonant cluster.

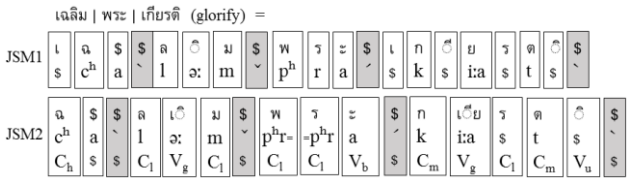

Figure 3: Two variations of grapheme-phoneme alignment.

In the last step of training, the joint sequences given by grapheme-phoneme alignment are modeled by another CRF, called CRF_{G2P} henceforth. An important parameter in training CRF is the window size of element context. The larger the window size, the longer the context dependency. For example, a window of 5 elements, called '5-gram', takes the two previous elements and two following elements into account when modeling.

The lower half of Figure 2 illustrates the model testing stage. Similar to the training stage, a test grapheme sequence is first segmented into PS units and tagged with character type ('B', 'I', 'A', and 'S'). The output tagged PS sequence is then fed into the trained CRF_{G2P} to predict a phoneme sequence.

4. Experiments and Results

Experimental results are presented in two parts. The first part presents the results of experimental data validation and the second part introduces the results of system evaluation.

4.1. Experimental data

The experimental data used for this study is a list of Thai words, labelled with their pronunciations, from a large Thai pronunciation dictionary [4]. The dictionary contains 103,265 unique words tagged with their pronunciations and PS boundary marks. The total number of unique PS segments is 19,601, of which 14,479 are unique, isolatable syllables. This research proposed the use of a training data set that contains all of the PS segments identified, in a real context. Hence, the nineteen thousand segments from the dictionary were used as a criterion for training word selection. Training word selection was done using a scoring equation proposed by Wutiwiwatchai et al. [10]. The resulting word list was the smallest word set that covered all possible PS segments found in the dictionary.

The first experiment was used to validate whether the selected training words could cover all possible Thai PS units. Two large text corpora were used to validate the PS coverage. The first corpus was a large set of text taken from an online newspaper, and the second corpus, named BEST [11], is a large text corpus collected from various sources for general Thai text processing research. In total, the data contained 26 million PS segments. Detailed statistics are presented in Table 3. From observation, most of the unknown PS units come from typing errors and incorrect segmentation.

Table 3: PS statistics of overall data used in the experiment.

Characteristics	Newspaper	BEST
Total no. of words	13,525,944	7,616,968
Total no. of PS units	16,203,551	10,174,761
No. of unique words	26,824	90,697
No. of unique PS units	14,693	16,828
No. of unknown unique PS units given the training set and %	3,532 (24%)	5,897 (35%)

4.2. System evaluation

The proposed system was evaluated using a set of 87,329 words from the dictionary that were not used for training. This system evaluation experiment compared three systems: baseline, JSM1, and JSM2. The baseline system was the latest G2P tool developed by Thangthai et al. [5]. It is based on syllable trigrams, in which candidate syllables are from handcrafted CFG rules. The JSM1 and JSM2 systems are the proposed CRF-based joint sequence modeling approaches with different grapheme-phoneme alignment design as described in Section 3.2. Two sizes of window, 3-gram and 5-gram, were tested in CRF_{G2P} modeling. Word error rates (WER) were counted in two cases: 'exact match' where all phonemes in a testing word have to be correctly predicted, and 'ignoring tones' where predicted tone mismatch is acceptable.

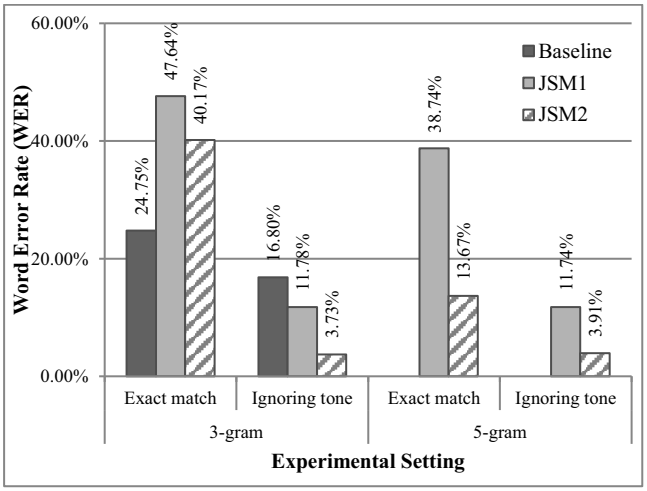

Figure 4: Word error rates comparison of Baseline, JSM1, and JSM2 systems.

Figure 4 shows the comparative results. When using the 3-gram window size, the baseline system achieves 24.75% WER for the exact match case and degrades to 40.17% WER by the JSM2 system. However, after expanding the window size to 5-gram, the proposed JSM2 system reduces the WER to 13.67% for the exact match case, and down to only 3.91% when ignoring tone errors. These results are reasonable as the baseline system uses the CFG rules for syllable structuring but the proposed JSM2 system considers the continuous input sequence without any syllable boundaries. Although the JSM2 system could function better for contextual analysis, a longer analysis window is needed to compensate the absence of syllable boundaries. Results from the JSM1 system are the worst in all cases. This shows that using a simple alignment like that used in English [3] is not applicable.

Considering errors at phone level, the baseline system achieves 6.68% Phone Error Rate (PER), while the JSM2 system achieves a PER value of 4.72% for 3-gram and 1.54% for 5-gram models. Moreover, when ignoring tone errors, linking syllable errors, and vowel length errors, the PER of the JSM2 system can reach 0.54%. This means that the three major errors presented in Section 2.2 cause only 1% phone error in the JSM2 system.

Table 4 shows further interesting analysis of the errors. In this table, the three G2P difficulties – tone ambiguity, vowel-length distortion, and linking syllables – are taken into account. The first row in the table gives the percentages of PS errors counted only on vowel-length distorted PS units. Similarly, the second and third rows are percentages of PS errors counted only on tone-distorted PS units and PS units containing linking syllables, respectively. The baseline system works without tone prediction, so the tone ambiguity statistic is not available. This analysis shows that the proposed JSM2 system can successfully handle the linking syllable case with only 1.80% PS errors. It can also improve the errors caused by vowel-length distortion, whereas tone ambiguity is still a problem to solve.

Table 4: Percentages of PS errors counted on each of the three major G2P problematic cases: vowel-length distortion, tone ambiguity, and linking syllable (5-gram used for JSM systems)

Major Thai G2P problem	Baseline	JSM1	JSM2
Vowel-length distortion	69.57%	78.22%	60.03%
Tone ambiguity	N/A	43.53%	43.79%
Linking syllable	26.28%	28.42%	1.80%

5. Conclusion and Future Work

This paper presented our recent attempt to improve the performance of Thai grapheme-to-phoneme conversion (G2P). The tone ambiguity problem, in particular, occurs more often in modern Thai language because many English loan words are introduced daily. The CRF-based joint sequence modeling successfully used in G2P was investigated. However, for Thai, it was not straightforward to align the training grapheme sequences to their corresponding phoneme sequences. A CRF-based pseudo-syllable (PS) segmentation module was introduced to indicate PS boundaries and at the same time to identify PS units potentially containing linking syllables. Grapheme-phoneme alignment rules were developed to solve the character-ordering problem by introducing vowel grouping, and to solve the consonant cluster ambiguity by introducing special phoneme symbols indicating the consonant cluster. The proposed system clearly outperformed the baseline system using CFG syllabification and syllable n-gram scoring. The highest merit of the proposed system was its ability to handle linking syllables, while the other difficulties such as vowel-length distortion could be partly resolved by the capability of CRF in long-context sequence modeling. In conclusion, the proposed system achieved a promising result of 13.67% word error rate, which is 11.09% lower than that of the baseline system.

Future work on this topic will focus on improvement of tone prediction and vowel-length specification. The tone ambiguity problem is of particular concern as it occurs with increasing regularity in Thai language as English loan words are introduced. Well-known machine learning methods, such as Deep Neural Networks (DNN) and Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN) will be considered as the means for such improvements.

6. References

[1] K. Tonglohr, Principle of Thai Language (In Thai), Bangkok: Ruamsarn (1977), 1994.

[2] W. Aroonmanakun, "Collocation and Thai Word Segmentation," in *SNLP-Oriental COCOSDA*, Huahin, 2002.

[3] P. Tarsaku, V. Sornlertlamvanich and R. Thongprasirt, "Grapheme-to-Phoneme for Thai," in *NLPRS*, Tokyo, 2001.

[4] P. Charoenpornsawat and T. Schultz, "Example-Based Grapheme-to-Phoneme Conversion for Thai," in *Interspeech*, Pittsburgh, 2006.

[5] A. Thangthai, C. Hansakunbuntheung, R. Siricharoenchai and C. Wutiwiwatchai, "Automatic Syllable-Pattern Induction in Statistic Thai Text-to-Phone Transcription," in *Interspeech*, Pittsburgh, 2006.

[6] A. Chotimongkol and A. W. Black, "Statistically trained orthographic to sound models for Thai," in *ICSLP*, Beijing, 2000.

[7] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication,* vol. 50, no. 5, pp. 434-451, 2008.

[8] P. Chootrakool, C. Wutiwiwatchai and K. Kosawat, "A Large Pronunciation Dictionary for Thai Speech Processing," in *ASIALEX*, Bangkok, 2009.

[9] S. Kongyoung and A. Rugchatjaroen, "Thai Pseudo Syllable Segmentation using Conditional Random Fields," in *Oriental-COCOSDA*, Shanghai, 2015.

[10] J. Lafferty, A. McCallum and F. C. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *ICML 2001*, Williamstown, 2001.

[11] A. Guta, Grapheme to Phoneme Conversion Using CRFs with Integrated Alignments, RWTH Aachen University of Technology, 2011.

[12] K. Yao and G. Zweig, "Sequence-to-Sequence Neural Net Models for Grapheme-to-Phoneme Conversion," in *Interspeech*, Dresden, 2015.

[13] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *NIPS*, Montreal, 2014.

[14] C. Wutiwiwatchai, P. Chootrakool, S. Saychum, N. Thatphithakkul, A. Rugchatjaroen and A. Thangthai, "TSynC-2: Thai Speech Synthesis Corpus Version 2," NECTEC, NSTDA, Bangkok, 2008.

[15] K. Kosawat, M. Boriboon, P. Chootrakool, A. Chotimongkol, S. Klaithin, S. Kongyoung, K. Kriengket, S. Phaholphinyo, S. Purodakananda, T. Thanakulwarapas, And C. Wutiwiwatchai, "BEST 2009: Thai Word Segmentation Software Contest," SNLP, 2009.