THAI GRAPHEME-TO-PHONEME USING PROBABILISTIC GLR PARSER

Pongthai Tarsaku, Virach Sornlertlamvanich and Rachod Thongprasirt

Information R&D Division, National Electronics and Computer Technology Center (NECTEC)
Gypsum Metropolitan Bldg., 22nd Fl., 539/2 Sri Ayudhaya Rd., Rajthevi, Bangkok 10400, Thailand.
ptarsaku@notes.nectec.or.th

Abstract

Many difficulties in the Thai language such as the absence of boundary word, linking syllables in pronunciation, and homographs are challenging us in developing a Thai Grapheme-to-Phoneme (G2P) converter. Presently there are a couple Thai G2P systems which are proposed in ruled-based and decision-tree approach. The rule-based approach has a drawback in the limitation of employing the context. The decision-tree approach is somehow able to capture the local context for making the decision. On the contrary, the Probabilistic Generalized LR (PGLR) approach is reported that both the global and local context are efficiently captured in the probabilistic model. In this paper, we implement a Thai G2P system based on the PGLR approach. The result of experiment shows 90.44% of word accuracy in case of ignoring vowels length and 72.87% of word accuracy in case of exact match evaluation. These results are superior to those of rule-based and decision-tree approaches.

1. Introduction

A Grapheme-to-Phoneme (G2P) module is a routine that converts an input word sequence into their corresponding phonetic transcription. It is one of the essential routines in developing a text-to-speech (TTS) module.

In order to implement a Thai G2P system, many language specific problems i.e. homograph, non-word booundary, tonal language, etc., need to be solved.

Various approaches, i.e. dictionary-based, rule-based, and statistical-based approaches have already been proposed for Thai G2P system [1], [2], [3], [5], [6], [9].

Luksaneeyanawin [3] and Mittapiyanurak [5] proposed a dictionary-based Thai G2P system for using in the text-to-speech (TTS) system. This approach requires a large dictionary, and it can not deal with unregistered word problem. To overcome the problem, Narupiyakul [6] and Khamya [2] proposed a rule-based approach for Thai G2P system. In their system, they prepared a set of rules of syllable construction and implemented with a finite state machine for detecting syllable structures. Wiboon [9] proposed a rule-based approach using regular expression to deal with the unregistered word problem. However, the rule-based approach can not deal with the ambiguity in syllabification of the Thai language. Chotimongkol [1] proposed a decision tree approach, but the problem of inversion of phonemes across syllable boundaries is left unsolved.

PGLR parsing is a technique invented for incorporating probabilistic value with the parsed trees of a sentence [7]. The PGLR has the advantage of context-sensitivity based on the GLR parsing framework [8]. The PGLR can efficiently capture the information from both the global and local context during the parse. This paper is organized in 5 sections. Section 2 describes Thai syllable structure and the difficulties in determining word pronunciation. Section 3 explains the architecture of G2P implemented on PGLR framework. Section 4 shows the experimental results.

2. Thai Pronunciation System

2.1. Thai Syllabic Representation

A basic Thai-pronunciation unit is a syllable that can be represented in the form of 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 tonal marker, respectively. The subunits inside the parenthesis, i.e. C_f and T, are optional. Details of each subunit are described in the followings:

1. Initial consonant (C_i): There are 44 consonants in Thai and all of them can be an initial consonant of a syllable (two of them, "ฃ" and "ฅ", are not often used in modern Thai). However, only 21 phonemes are needed to represent all these consonants as shown in Table 1.

Table 1: Phonetic symbol of Thai consonants.

Consonant	Phoneme		Consonant	Phoneme	
	Initial	Final		Initial	Final
ก	/k/	/k/	บ	/b/	/p/
ข,ค,ฆ	/kh/	/k/	ป	/p/	/p/
ง	/ng/	/ng/	ผ,พ,ภ	/ph/	/p/
จ	/c/	/t/	ฝ,ฟ	/f/	/p/
ฉ,ฌ,ช	/ch/	/t/	ม	/m/	/m/
ซ,ศ,ษ,ส	/s/	/t/	ร	/r/	/n/
ญ,ย	/j/	/j/	ล,ฬ	/l/	/n/
ฎ,ด	/d/	/t/	ว	/w/	/w/
ฏ,ต	/t/	/t/	ห,ฮ	/h/	-
ฐ,ฑ,ฒ,ถ,ท,	/th/	/t/	อ	/?/	-
ณ,น	/n/	/n/			

Some initial consonants are written pairly and act as an initial consonant. This type of initial consonants is called 'double consonant'. There are four catagories of double consonant:

1. True clusters: (“ปร”, /pr/), (“ตร”, /tr/), (“กร”, /kr/), (“กล”, /kr/), (“กว”, /kw/), (“พร”, /phr/), (“ทร”, /thr/), (“คร”, /khr/), (“ขร”, /khr/), (“พล”, /phl/), (“ผล”, /phl/), (“คล”, /khl/), (“ขล”, /khl/), (“คว”, /khw/), (“ขว”, /khw/).
2. Pseudo clusters: (“ทร”, /s/), (“จร”, /c/), (“ซร”, /s/).
3. Parallel consonant characters: For example (“กล”, /k-a-1/l-/), (“ปร”, /p-a-1/r-/), etc.
4. Leading consonant characters.

More information can be found in [4].

2. Vowel (V): There are 28 vowels in Thai composing of 18 monophthongs, 6 diphthongs, and 4 vowel letters as shown in Table 2.

Table 2: Phonetic Symbols of Thai vowels.

Type	Short vowel		Long vowel	
	grapheme	**phoneme**	**grapheme**	**phoneme**
Monophthong	–ะ	/a/	–า	/a:/
	–ิ	/i/	–ี	/i:/
	–ึ	/v/	–ื –ื	/v:/
	–ุ	/u/	–ู	/u:/
	เ-ะ	/e/	เ-	/e:/
	แ-ะ	/x/	แ-	/x:/
	โ-ะ	/o/	โ-	/o:/
	เ-าะ	/@/	เ-า	/@:/
	เ-อะ	/#/	เ-อ	/#:/
Diphthong	เ–ียะ	/ia/	เ–ีย	/i:a/
	เ–ือะ	/va/	เ–ือ	/v:a/
	–ัวะ	/ua/	–ัว	/u:a/
Vowel Letter	-ำ	/am/	-	-
	ไ-, ใ-	/aj/	-	-
	เ-า	/aw/	-	-

3. Final consonant (C_f): Not all Thai consonants can be a final consonant. For example, “ห”, and “ฮ” can not be used as a final consonant. Only 9 phonemes are needed to represent the final consonant. Like the initial consonant, the final consonant can be a double consonant. For example, /กร/ is mapped to /k/, /คร/ is mapped to /k/, /ตร/ is mapped to /t/, /ทร/ is mapped to /t/, and /ปร/ is mapped to /p/.

4. Tone (T): Like Chinese, Thai is a tonal language. There are five tones in Thai, i.e. mid, low, falling, high and rising. Four tonal markers and one non-mark are used to indicate the tone. A tone is determined by the combination of syllable structure, initial consonant and the tonal marker.

2.2 Difficulties in Thai pronunciation system

Difficulties in transcribing a Thai letter can be classified as follows.

1. Ambiguity in grapheme-phoneme mapping: Some graphemes can be mapped to multiple phonemes depending on the context. For example, the grapheme “ฑ” in the word “มณฑา” is mapped to /t/, while it is mapped to /d/ in the word “มณฑป”.
2. Homograph: The words that are pronounced differently according to the semantic context. In order to pronounce a homograph word, the sentential context is required. For example, “เพลา” is pronounced as /phl-a-w-0/ to mean “axle”, but “เพลา” is pronounced as /ph-e:-0/l-a:-0/ to mean “time”.
3. Vowel’s length: The problem occurs when a vowel is pronounced as a short (long) vowel according to its grapheme, but by any reasons it is pronounced as a long (short) vowel instead. For example, the word “น้ำ” which is pronounced as /n-a-m-3/ (short vowel) by its grapheme, but it is usually pronounced as /n-a:-m-3/ (long vowel). The word “ท่าน” should be pronounced as /t-a:-n-2/ (long vowel), but it is usually pronounced as /t-a-n-2/ (short vowel).
4. Linking syllable pronunciation: The problem occurs in a polysyllabic word where the final consonant of the forthcoming syllable is explicitly pronounced with /a/ vowel as an additional syllable. For example, the polysyllabic word “วิทยา” is pronounced as /w-i-t-3/th-a-2/j-a:-0/. In this case, the grapheme “ท”, the final consonant of the first syllable, “วิท”, is added with /a/ vowel to pronounce /th-a-2/ as the second syllable. As a result, the overall pronunciation is /w-i-t-3/th-a-2/j-a:-0/.
5. Ambiguity in consonantal functionality: In some Thai words, a consonant can be both the final consonant of forthcoming syllable and the initial consonant for the next syllable. For example, “ฐ” in “อัฐิ” (/?-a-t-1/th-i-1/) is both the final consonant /t/ of the first syllable “อัฐ” and the initial consonant /th/ of the following syllable “ฐิ”.
6. Word boundary: Unlike some other languages such as English, Thai has no word boundary. Therefore, different segmentation can yield different words and pronouce in different syllable units. For example, the word “ตากลม” can be segmented into “ตา|กลม” (round eye) and “ตาก

ลม" (to expose wind) which are pronouced /t-a:-0/kl-o-m-0/ and /t-a:-k-1/l-o-m-0/ .

3. Grapheme-to-Phoneme System

3.1 PGLR Approach

The entire Thai G2P system is implemented on the PGLR [7] approach as shown in Fig. 1. The input text, a sequence of Thai graphemes, are parsed by the GLR parser [8] with the CFG rules for syllable construction. The outputs of the GLR parser are the parsed trees with the corresponding probabilities. The most likely parsed tree is selected and then transformed to the Grapheme-Phoneme mapping module to generate the phonetic transcription by table look-up.

example, the final consonant "ร" can not follow some vowels, e.g. ิ, ี, ุ, ู, ึ, and ื. This kind of restriction is introduced to restrict the invalid syllable sequences.

In order to handle the linking-syllable pronunciation problem, the possible final consonants for pronunciation as a linking-syllable are defined with a new non-terminal type <*FinConwithLinking*> composed in a syllable structure such as

<*Syl_TypewithLinking*> → <*InitCon*> า <*FinConwithLinking*>

corresponding to the original rule:

<*Syl_Type*> → <*InitCon*> า <*FinCon*>.

Since some final consonants, e.g. "ง", "ด", and "บ", can not be pronouced as a linking-syllable, the new non-terminal <*FinConwithLinking*> is defined as a different non-terminal from the ordinary final consonant <*FinCon*>.

Furthermore, consonant type specific rules are also prepared for solving the ambiguity in consonant functionality problem.

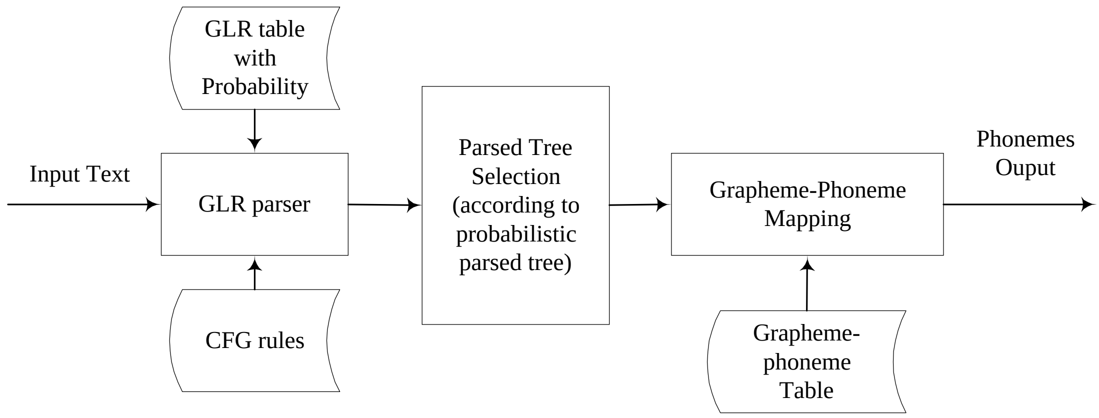

Figure 1: Thai G2P System: the PGLR approach.

3.2 Context-Free Grammar Rules

Syllable structures are described in CFG rules for preparing the LR parsing table. The CFG rules are designed according to the vowel structure. For example, the CFG rules for the syllable structure. For example,

[*initial consonant*] [*vowel*] [*Final consonant*] to parse the words "การ", "นาก", and "ยาว" is represented as:

<*Syl_Type*> → <*ini-cons*> *Vx* <*fin-cons*>,

where <*Syl_Type*>, <*ini-cons*> and <*fin-cons*> are non-terminal nodes representing syllable type, initial consonant, and final consonant, respectively and *Vx* is a terminal node representing a vowel.

Since a Thai word can be either monosyllable or polysyllable, followings are the sample of CFG rules for word formation from syllable units:

<*Word*> → <*Syllable*>
<*Syllable*> → <*Syllable*> <*Syl_Type*>
<*Syllable*> → <*Syl_Type*>.

In case that some combinations of charactors are not allowed due to the spelling regulation, additional restricted rules are prepared to reduce the acceptable parsed candidates. For

These rules depend on the syllable structure. For example, the rule:

<*SylTypeSpecial*> → <*InitCon*> *Vowel1* <*FinCons*> *Vowel2*

is prepared for parsing the word such as "อัฐิ" (ash). In this case, <*InitCon*>, *Vowel1*, <*FinCons*>, and *Vowel2* correspond to the characters "อ", ั, "ฐ", and ิ, respectively.

3.3 Training of the PGLR parsing table

The probabilistic GLR parsing table is a LR table in which each action is assigned a probability gained from parsing a bracketed training set [7]. LEXiTRON is a Thai electronic dictionary containing 23,000 entries with pronunciation provided [10]. A four-fifth of the LEXiTRON is randomly selected for training. All possible parsed trees for each word are generated by the GLR parser. Only those with the correct parsed trees (phoneme sequence) comparing with the word pronunciation in LEXiTRON, are selected. As a result, the correct parsed tree for each word is listed up for training.

4. Experimental results

The remained one-fifth of the LEXiTRON, is used to test the PGLR model. Table 3 shows the results of the PGLR model comparing with the rule-based [9] and the decision-tree models [1]. The evaluation shows the grapheme-to-phoneme conversion accuracy in both case of exact match and ignorance of vowel length match. The exact match accuracy indicates that the output of the model has exactly the same phoneme sequence as the one given in LEXiTRON. The ignorance of vowel length match accuracy shows the model accuracy in yielding the phoneme sequence by ignoring the length of vowel which is less important than the phoneme sequence. It is expected that the vowel length may be determined by other means.

The PGLR approach yields the best result in comparing with those of the rule-based and decision-tree approaches. However, about half of the errors are the words that include linking syllables. This problem is also not trival for native speakers either when the word is unseen.

Table 3: Grapheme-to-phoneme conversion accuracy

Model	Conversion (word) accuracy(%)	
	Exact match	Ignorance of Vow. Length
PGLR	72.87	90.44
Rule-based	67.14	83.81
Decision Tree	68.76	86.94

5. Conclusion

Theoretically, PGLR can capture both local and global context which is reported the high performance in syntactic parsing for sentences. Our research has shown the best result in applying PGLR in G2P problem. This concludes that the context for probabilistic decision in PGLR is good enough for a relatively small structure such as words (phoneme sequence) as well.

6. Reference

[1] Chotimongkol, A. "Statistically Trained Orthographic to Sound Models for Thai," Proceeding of the 6th International Conference on Spoken Language Processing," pp. 533-554, Oct. 2000.

[2] Khamya, A. "SATTS: Syllable Analysis for Text-to-Speech System," Proceeding of the Fourth Symposium on Natural Language Processing 2000 (SLNP'2000), pp. 336-340, May 2000.

[3] Luksaneeyanawin, S. "A Thai Text to Speech System," Proceeding of the Conference of the Regional Workshops on Computer Processing of Asian Languages. Asian Institute of Technology, pp. 305-315, 1990.

[4] Luksaneeyanawin, S. "Speech Computing and Technology in Thailand," NLP in Thailand, pp. 276-321, 1993.

[5] Mittapiyanurak, P., Hansakunbuntheung, C. and Tepprasit, V. and Sornlertlamvanich, V. "Issues in Thai Text-to-Speech Synthesis: The NECTEC Approach," Proceedings of NECTEC Annual Conference, Bangkok Thailand, pp. 483-495, 2000.

[6] Narupiyakul, L. "The Phonetic Transcription of Thai Word," Proceeding of the International Symposium on Intelligent Signal Processing and Communication Systems, pp. 789-792,1999.

[7] Sornlertlamvanich, V. " Probabilistic Language Modeling for Generalized LR Parsing," Technical Report, Department of Computer Science, Tokyo Institute of Technology, September, 1998.

[8] Tomita, M., "Generalized LR Parsing," Kluwer Academic Publishers, 1991.

[9] Wiboon, T. "The Engineering Training Report at NECTEC," Engineering Faculty, Khasetsart University, 1999.

[10] http://www.links.nectec.or.th/lexit/.