Thai-Lao Machine Translation based on Phoneme Transfer
音素トランスファータイ・ラス機械翻訳

Virach Sornlertlamvanich, Chumpol Mokarat
Thai Computational Linguistics Lab., NICT Asia Research Center,
Pathumthani, Thailand
{virach, chumpol}@tcllab.org

Hitoshi Isahara
NICT, Japan
isahara@nict.go.jp

タイ語とラス語では文法、単語、文字コード体系化のスキームも非常に似ている。そのため、国境に住んでいる人たちはお互いの言葉が分る。新聞、ラジオ、テレビおよびインターネットの普及とともにタイの番組、情報がラス側で流通している。タイ語のままで良さそうだがやはり正確なラス語で流通することが望ましい。一方、ラス語処理の研究そして応用ソフトの開発するためにはラス語のコーパスは欠かせないもの。ラス語コーパスの開発、辞書も同様、ラス語より多く存在しているタイ語コーパスから変換してできればラス語の資源として使えるであろう。本研究では、基礎として音節単位での変換ルールを適応する。まずは、Grapheme to phoneme の段階でタイ語の入力を PGLR によって音素に変換する。PGLR の特徴を生かして読みのあいまい性が状態遷移の確立の値で解消される。その後、音素ごとの変換は一対一でラス語の音節を合成することができる。いくつかの注意点があるのだが、大抵の場合では、音節の合成ルールによって各音素に使用する文字が確定できる。以上のルール・ベースで変換できない単語か表現には、予め辞書に登録しておけばタイ・ラス機械翻訳の実現ができる。その結果、タイ語とラス語の対だけではなく、タイ語と他の言語対の資源を基にしてラス語の言語資源を増設することができる。

Thai and Lao languages are very similar in terms of grammar, lexicon, and character encoding scheme. The people who live in the border of the two countries can commonly understand each other's language very well. Under the circumstance of information push from Thailand through Internet, TV programs, radio, newspapers for instance, many Lao people having chances in consuming the information can understand the Thai language very well. However, reading in a foreign text is not so natural. Moreover, to conduct a research in NLP in the recent years, we need a large amount of corpus due to the effectiveness of corpus-based approaches. If we can convert from the language having richer resources, it is possible to make use of the converted resources for creating the model of the target language. This paper introduced a phoneme-based transfer method for Thai to Lao machine translation. We first analyze the input text to generate the most probable sequence of phoneme by probabilistic GLR (PGLR) approach. On the output of sequence of phoneme, we apply Thai-Lao phoneme conversion rule to obtain the proper Lao pronunciation. Morphological generation is then applied to get the Lao translation. In case of language specific expression, we only need to prepare a dictionary for word to word translation. As a result, we can generate a Lao text from an input Thai text by the phoneme-based machine translation approach. A preliminary evaluation on 35,125 Thai words conversion was conducted. Without using a dictionary the accuracy of translation is 76%.

1 Introduction

Machine translation [1] has been a major topic in natural language processing for years. Until the present day, many types of translation, for instance, direct, transfer, interlingual, and many approaches, for instance, rule-

base, statistical base, probabilistic base, have been proposed to solve the sophisticate tasks of machine translation for several pairs of languages. An arbitrary approach may suit a language pair of translation rather than another in terms of its performance of translation.

This paper introduces a phoneme-based approach for machine translation between the languages that considerably share their language phenomena in terms of grammar, vocabulary or character encoding scheme. We selected a language pair of Thai and Lao for the evaluation of the proposed machine translation based on phoneme transfer.

We adopted PGLR approach for Thai grapheme to phoneme conversion [2] to generate an output sequence of phoneme. This sequence of phoneme is then mapped into a Lao phoneme sequence by a phonetic conversion rule. This rule can convert the words that share the same root words in both languages. Exceptional rules are also prepared for the words that do not follow the regular phonetic conversion. There are some words, especially the words which are derived from different root, cannot be converted by the phoneme transfer approach. In this case, a word mapping dictionary is prepared to increase the translation coverage. Finally, the Lao phonetic sequence is mapping into its script to produce the Lao text output.

A preliminary evaluation on 35,125 Thai words conversion was conducted. It yielded an accuracy of .76% without using a dictionary.

The rest of the paper is organized as follows. Section 2 describes the preparation of Thai-Lao phonetic conversion rule. Section 3 discusses the effectiveness of the implementation of Thai-Lao machine translation based on phoneme transfer and a preliminary evaluation result. Section 4 concludes the paper.

2 Thai-Lao Phonetic Conversion Rule

A Thai phoneme can be correctly converted into a Lao phoneme according to the type of syllable. There are two types of syllable and each type can have four different patterns. These patterns are defined by the type of character set and its position in the structure. Table 1 shows the two types of syllable together its four possible patterns. The live syllable with only 'a' vowel has a different structure. Therefore, it is defined in a separated pattern.

Table 1 Pattern of Syllable Type

Dead Syllable	Live Syllable (excluding 'a' vowel)	Live Syllable (with 'a' vowel)
1.1 C^{M} V ?	2.1 C^{M} V [N,SV]	3.1 C^{M} V N : N = n, ng
1.2 C^{M} VV S	2.2 C^{M} VV [N,SV]	3.2 C^{M} V [N,SV] : N = m, SV = j, w
1.3 $C^{L,H}$ V ?	2.3 $C^{L,H}$ V [N,SV]	3.3 $C^{L,H}$ V N : N = n, ng
1.4 $C^{L,H}$ VV S	2.4 $C^{L,H}$ VV [N,SV]	3.4 $C^{L,H}$ V [N,SV] : N = m, SV = j, w

Note:

C^{M} : Mid consonant C^{L} : Low consonant C^{H} : High consonant
SV : Semi Vowel (-j, -w) S : Stop (-p, -t, -k) N : Nasal (-n, -m, -ng) ? : Glottal (-z)
V : Short Vowel VV: Long Vowel T : Tone (0, 1, 2, 3)

Once the syllable pattern is recognized, the phonetic conversion rule can map for the initial consonant, select an appropriate vowel sign and tonal mark according to the syllable pattern.

Some exceptional rules are provided for the irregular phoneme mapping. Those are:

1. If it is a live syllable having $C^{L,H}$ as the initial consonant and a tone '2' then map to tonal mark '◌่'.
2. If the initial consonant is a double consonant such as 'อย', 'หม', or 'คร' then map to a correspondent consonant those are 'ย', 'ฆ', or 'ถ', respectively.
3. If the vowel sign is 'ไ' or 'ใ' then map to 'ไ' or 'ใ', respectively.

These exceptional rules occur due to the preference of each language in the process of derivation. It is a predictable pattern. Therefore, the pattern for each rule is prepared.

Most of the Lao words can be converted by the above direct phoneme mapping and the exceptional rules. For the words that the phonetic conversion rule cannot cover, it needs to prepare a pre-defined word mapping dictionary.

3 Thai-Lao Machine Translation based on Phoneme Transfer

In a Thai input string, there is no an explicit word boundary marker. To generate a proper string of phoneme, it is necessary to know the appropriate word boundary. Previously, there was a work on grapheme to phoneme (G2P) for the Thai language using the PGLR approach. It was reported a high accuracy of 90.44% when ignoring the vowel length [2]. The advantage of the approach is that the output of PGLR is the most probable string of phoneme with a syllable boundary. To apply a phonetic conversion rule, it is necessary to generate a sequence of phoneme for an input string. Therefore, the G2P module based on PGLR approach is adopted to prepare an input phoneme sequence for phonetic conversion rule.

The similarity between the Thai and Lao languages is significant in terms of their vocabulary, grammar or character set [3]. In general, they share their vocabulary but differentiated by their scripts i.e. /kin/ means 'to eat' in Thai it is '**กิน**' and in Lao it is 'ກິນ'. In this case, it works by simply applying a character mapping process. However, it is not true in all cases. But if we can capture a rule for phoneme mapping for applying to the case of pronunciation derivation, we can generate a Lao word of 'ເຄື່ອງລ້ອນ' /kh-vv-ng^-2|l-@-n^-2/ from a Thai word of '**เครื่องร่อน**' /khr-vv-ng^-2|r-@-n^-2|/. In this case, we need to prepare a rule set to map /khr/ to /kh/ and /r/ to /l/. This phoneme mapping rule is not always true, therefore, a dictionary for word conversion is also needed for different word origin to express the same word meaning.

Figure 1 shows the diagram of Thai-Lao machine translation. The implementation includes a G2P module to convert a Thai input string into a sequence of phoneme with syllable break markers. A phonetic conversion rule set is prepared to map the regular phoneme derivation of consonant, vowel or tone. The rule set is applied to generate a string of Lao phoneme from Thai phoneme.

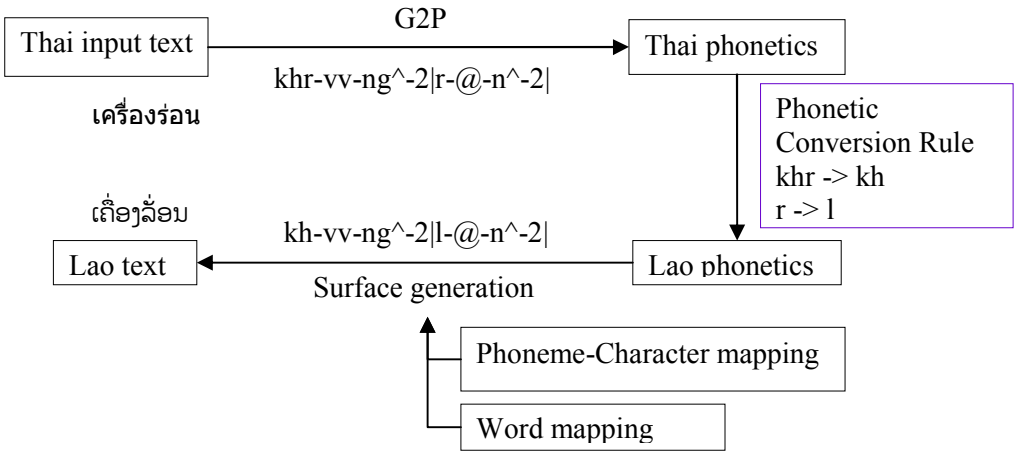

Figure 1 Thai-Lao Phoneme-based Machine Translation

When the string of Lao phoneme is determined, we apply a phoneme-character mapping table as shown in Table 2 and 3 to generate the final Lao output string. Multiple consonants are introduced for some sounds in the Thai language, but not in the Lao language. Therefore, there is no ambiguity in selecting a Lao character, once its sound is determined. Though both languages have their own script for the tonal mark, they both contain four tonal marks.

4 Conclusion

A phoneme transfer approach for Thai to Lao machine translation has been proposed. It yielded an acceptable evaluation result on a list of 35,125 word entries from a Thai dictionary. The effort has yielded a promising result to implement a new approach of machine translation for the similar languages in terms of their grammar, vocabulary and character encoding scheme. The coverage evaluation is still needed to confirm the reliability of the translation. The approach can also support the development of language resources from a language having richer resources. We would like express our sincere thanks to Mr. Valaxay Dalaloy and his

colleagues at Science Technology and Environment Agency (STEA) of Lao PDR for their contributions in the evaluation of translation.

Table 2 Phoneme-Character Mapping for Thai and Lao Consonant

Thai	Phoneme	Lao	Thai		Phoneme	Lao	
Mid		Mid	Low	High		Low	High
ก	K	ภ	ค, ฆ	ข	kh	ถ	ฃ
จ	C	จ	ช, ฌ	ฉ	ch	ซ	ส
ด, ฎ	D	ถ	ซ	ส, ศ, ษ	ch	ซ	ส
ต, ฏ	T	ต	ง	หง	ng	ง	หง
บ	B	บ	ญ, ย	หญ, หย	j	ย	หย
ป	P	ป	ฑ, ฒ, ธ	ฐ ถ	th	ท	ฤ
อ	Z	ฮ	ณ, น	หน	n	น	หน
			พ, ภ	ผ	ph	พ	ผ
			ฟ	ฝ	f	ฟ	ฝ
			ม	หม	m	ม	หม
			ร	หร	r	ธ ล	หล
			ล, ฬ	หล	l	ล	หล
			ว	หว	w	อ	หอ
			ฮ	ห	h	ธ	ห

Table 3 Phoneme-Character Mapping for Thai and Lao Monophthong, Diphthong, and Consonant Vowel

Thai	Phoneme	Lao	Thai	Phoneme	Lao	Thai	Phoneme	Lao
อะ	a	ะ	อา	aa	า	เอียะ	ia	◌ັย
อิ	i	◌ิ	อี	ii	◌ี	เอีย	iia	เ - ย
อึ	v	◌ึ	อือ	vv	◌ื	เอือะ	va	เ◌ือ
อุ	u	◌ุ	อู	uu	◌ู	เอือ	vva	เ◌ือ
เอะ	e	เ - ะ	เอ	ee	เ	อัวะ	ua	◌ົวะ
แอะ	x	แ - ะ	แอ	xx	แ	อัว	uua	◌ົว
เออะ	q	เ◌ิ	เออ	qq	เ◌ิ	อำ	am	◌ำ า
โอะ	o	โ - ะ	โอ	oo	โ	ไอ	aj	ไ
เอาะ	@	เ - าะ	ออ	@@	◌ॅ	ใอ	aj	ใ
						เอา	aw	เ - า

Reference

[1] John Hutchins, 2004, Research Methods and Systems Designs in Machine Translation a Ten-year Review, 1984-1994, Cranfield University, England.

[2] Tarsaku, P., Sornlertlamvanich, V., and Thongpresirt, R., 2001, Thai Grapheme-to-Phoneme Using Probabilistic GLR Parser, Eurospeech 2001, vol. 2, pp. 1057-1060.

[3] Meesathan, W., 2000, Lao-Thai Dictionary, Institute of Language and Culture for Rural Development, Mahidol University, Sahathammik Press.