Regulating Orthography-Phonology Relationship for English to Thai Transliteration

Binh Minh Nguyen
National University of Singapore, Singapore
nguyen.binh.minh92@u.nus.edu

Gia H. Ngo
National University of Singapore, Singapore
ngohgia@u.nus.edu

Nancy F. Chen
Institute for Infocomm Research, Singapore
nfychen@i2r.a-star.edu.sg

Abstract

In this paper, we discuss our endeavors for the Named Entities Workshop (NEWS) 2016 transliteration shared task, where we focus on English to Thai transliteration. The alignment between Thai orthography and phonology is not always monotonous, but few transliteration systems take this into account. In our proposed system, we exploit phonological knowledge to resolve problematic instances where the monotonous alignment assumption breaks down. We achieve a 29% relative improvement over the baseline system for the NEWS 2016 transliteration shared task.

1 Introduction

Transliteration is the process of transforming a word from one writing system (source word) to a word in another writing system (target word) (Knight and Graehl, 1998). Transliteration is often used to borrow names and technical terms from the source language into the target language when translation is difficult or awkward. For example, *British* is transliterated into Thai as [*bri-tit*] using Royal Thai General System of Transcription (RTGS) notation (Royal Institute, 1999).

Transliteration can be formulated as a special case of translation. Instead of converting words from one language to the semantically equivalent words in another language, transliteration converts the source word to a phonetically equivalent target word (Knight and Graehl, 1998).

In transliteration, character-reordering is important to ensure the transliterated words follow the phonotactic rules of the target language. Character-reordering in transliteration is similar to word-reordering in translation, where the translated sentence needs to satisfy the grammatical rules of the target language. However, most transliteration systems do not take into account character-reordering.

If the phonetically equivalent characters are reordered in the target word, the source word and the transliterated word are said to be non-monotonously aligned (Toms and Casacuberta, 2006). A classical approach to transliteration is using phrase-based Statistical Machine Translation (pbSMT). While pbSMT approach can model non-monotonous alignment of characters in the transliteration task, the pbSMT systems introduced in Kunchukuttan and Bhattacharyya (2015), Nicolai et al. (2015) and Finch et al. (2015) did not model such character-reordering. In addition to pbSMT, Nicolai et al. (2015) also used grapheme-to-phoneme (G2P) conversion tools, namely DirecTL+ (Jiampojamarn et al., 2010) and Sequitur G2P (Bisani and Ney, 2008), for the transliteration task. Such G2P conversion tools also make a similar assumption of monotonous alignment between source and target word. While this assumption is reasonable for transliteration between most language pairs, there are cases in English to Thai transliteration whereby this assumption is invalid.

In this paper, we regulate the relationship between Thai orthography and phonology in the English to Thai transliteration task. We show that the transliteration accuracy can be improved by addressing the mismatch between Thai orthography and phonology that causes the monotonous alignment assumption to break down.

2 Thai Phonology

2.1 Syllable

A syllable is considered the basic unit of a word in both written and spoken language (Ladefoged and Johnson, 2014). Most languages have the following syllable structure, including English and Thai (Kessler and Treiman, 1997; Luksaneeyanawin, 1992):

Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL, pages 83–87,
Berlin, Germany, August 12, 2016.

$$[O] \quad N \quad [Cd] \quad [T] \qquad (1)$$

where O, N, Cd, T denotes onset, nucleus, coda, tone respectively.

In Thai, an onset (O) has at most two consonants and a coda (Cd) has at most one consonant, while a nucleus (N) can be a vowel or a diphthong (Luksaneeyanawin, 1992). Tone (T) is a feature of many tonal languages (Yip, 2002). Tone is a variation in pitch that is used to distinguish different words (Yip, 2002). For example, both the Thai words *value* and *to trade* have the same syllable [*khaa*] (RTGS), but *value* is pronounced with a falling tone while *to trade* is pronounced with a high tone.

The aim of transliteration is to generate a word in the target language that best matches the pronunciation of the word in the source language (Knight and Graehl, 1998). Although the syllable structure in English and in Thai are similar, the idiosyncratic relationship between Thai pronunciation and Thai orthography makes English to Thai transliteration complex.

1. In Thai orthography, the position of the onset and nucleus can be inverted for a syllable with a leading vowel (Chotimongkol and Black, 2000). A leading vowel is part of the syllable's nucleus. While the vowels of the nucleus are pronounced after the consonants of the onset as specified by the syllable structure in (1), leading vowels precede the consonants of the onset in written form. The order of consonants and vowels in Thai written form therefore does not always match the pronunciation order. This is unlike English whereby the order of consonants and vowels in the written form matches the order in pronunciation. For example, *Reagan* is transliterated into Thai as [*er-aekn*] (RTGS) but is pronounced in Thai as [reːkɛːn] (IPA). [*e*] and [*ea*] are leading vowels and are written before their corresponding onset which are [*r*] and [*k*] respectively.
2. Lexical tones in Thai are not uniquely defined by tone marks, but are determined by the type of the syllable, the class of the onset consonant, and the length of the nucleus (Smyth, 2002). For example, the Thai words for both *to pass* and *glancing* are pronounced with a low tone. However, only *to pass* has a tone mark in Thai script, while *glancing* does not.
3. Some vowels are implicit in certain phonetic contexts (Chotimongkol and Black, 2000). Such vowels are present when the syllable is pronounced but it is omitted when the syllable is transcribed. For example, *Steve* is transliterated as [*s-tip*] but is pronounced as [sa-tip] (RTGS). In this case, the vowel [*a*] is implicit.

In this work, we focus on the onset-nucleus inversion case (case 1), which occurs much more often.

2.2 Alignment between Thai Orthography and Phonology

Even though the syllable structure of English and Thai are similar, the mismatch between Thai orthography and Thai syllable structure can lead to challenges in aligning English-Thai transliteration word pairs. In this section, we describe how onset-nucleus inversion may make monotonous alignment difficult.

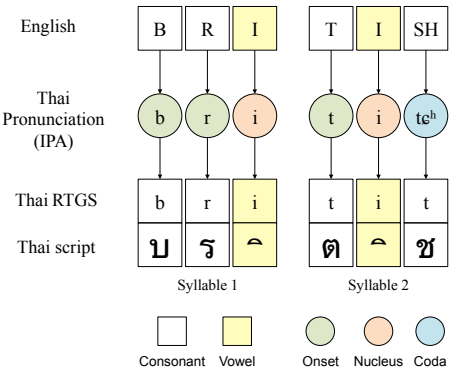

Figure 1: Monotonous Alignment, the source word is *British*

When there is no onset-nucleus inversion such as the case in Figure 1, the alignment between the English word and the Thai word is monotonous.

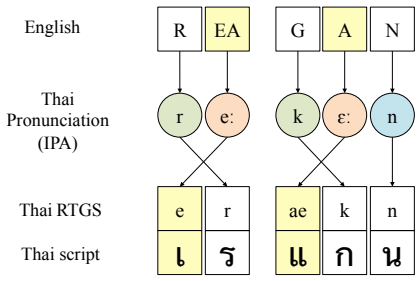

Figure 2: Non-Monotonous alignment of English-Thai transliteration pairs due to onset-nucleus inversion of monophthongs.

However, when onset-nucleus inversion occurs such as the case in Figure 2, the alignment is

non-monotonous. Under the monotonous alignment assumption, the English onset [R] may be wrongly aligned to the Thai nucleus [*e*], and the English nucleus [EA] may be wrongly aligned to the Thai onset [*r*]. The English syllable [REA] may still be aligned correctly to the Thai syllable [*er*] if the syllables appear together frequently in the training data. Nevertheless, the presence of onset-nucleus inversion increases the number of possible alignments between English-Thai transliteration pairs.

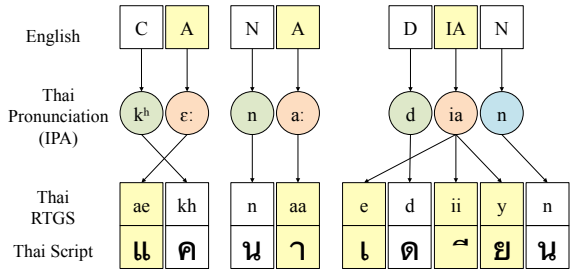

Figure 3: Non-Monotonous alignment of English-Thai transliteration pairs due to onset-nucleus inversion of diphthongs.

Onset-nucleus inversion also occurs in syllables with diphthongs. In Figure 3, the nucleus of the third syllable, namely [e] [ii] [y], is a diphthong in Thai, which comprises of three phonemes [e], [ii] and [y]. Although [e], [ii] and [y] are components of the same nucleus, they are not adjacent. Under the monotonous alignment assumption, it is unclear how to align the English onset, nucleus and coda with Thai onset, nucleus and coda respectively. We attempt to address these issues in the proposed transliteration system.

3 Baseline Systems

We considered two classic approaches as our baseline systems. Transliteration seeks to convert an English string $\mathbf{f} = (f_1, f_2, \ldots, f_n)$ to a Thai string $\mathbf{e} = (e_1, e_2, \ldots, e_m)$.

3.1 Phrase-based Statistical Machine Translation System

Under the pbSMT system, the objective function is given by:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} p(\mathbf{e})p(\mathbf{f} \mid \mathbf{e}), \qquad (2)$$

where $p(\mathbf{e})$ is estimated from an n-gram language model of Thai, and $p(\mathbf{f} \mid \mathbf{e})$ is estimated from the alignment of segments (phrases) of $\mathbf{f}$ with segments (phrases) of $\mathbf{e}$.

We implement the pbSMT system with the Moses toolkit (Koehn et al., 2007). We use GIZA++ (Och and Ney, 2003) to perform alignment, and SRILM (Stolcke, 2002) to train a 5-gram language model of Thai transliteration units with Witten-Bell smoothing (Witten and Bell, 1991). While the pbSMT systems implemented in (Kunchukuttan and Bhattacharyya, 2015) and (Nicolai et al., 2015) did not model word reordering, we tried various reordering models offered by the Moses toolkit.

3.2 Joint Source-channel System

The baseline system is based on the joint source-channel model, formulated for transliteration in (Li et al., 2004). A similar model (joint sequence model) was proposed for grapheme-to-phoneme conversion in (Bisani and Ney, 2002). We use the Sequitur G2P tool from (Bisani and Ney, 2008) to train the joint source-channel model by assuming a direct correspondence between phoneme and grapheme in the target language.

Given an English string $\mathbf{f}$ and a Thai string $\mathbf{e}$, the joint source-channel model estimates the co-segmentation $\mathbf{q}$, defined as $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ where, $q_i = (f_i, e_i)$, $\mathbf{f} = (f_1, f_2, \ldots, f_n)$ and $\mathbf{e} = (e_1, e_2, \ldots, e_n)$. During decoding, the output Thai string corresponds to the co-segmentation that matches the input English string and yields the maximum likelihood.

The monotonous alignment assumption is built into the joint source-channel model (Bisani and Ney, 2008). Under this assumption, $\forall i < j$, f_i appears before f_j and e_i appears before e_j. In a Thai syllable with onset-nucleus inversion, e_i corresponds to a leading vowel and e_{i+1} corresponds to the onset that comes after the leading vowel. However, f_i and f_{i+1} may still be matched to the onset and nucleus in the English syllable. Therefore, the model may be confused, as the English onset is matched to the Thai nucleus, and the English nucleus is matched to the Thai onset.

4 Proposed Augmented System

The proposed system (Figure 4) augmented the joint source-channel model with a vowel-onset transposition step to regulate the syllable structure.

During training, for Thai syllables with onset-nucleus inversion, the vowel-onset transposition step swaps the location of the leading vowel and the onset consonant (see next page). This swapping ensures that the nucleus always occurs after the onset in a syllable, and vowels that belong to the same nucleus are adjacent in the Thai script.

During decoding, Thai entries that have undergone vowel-onset transposition are reverted back

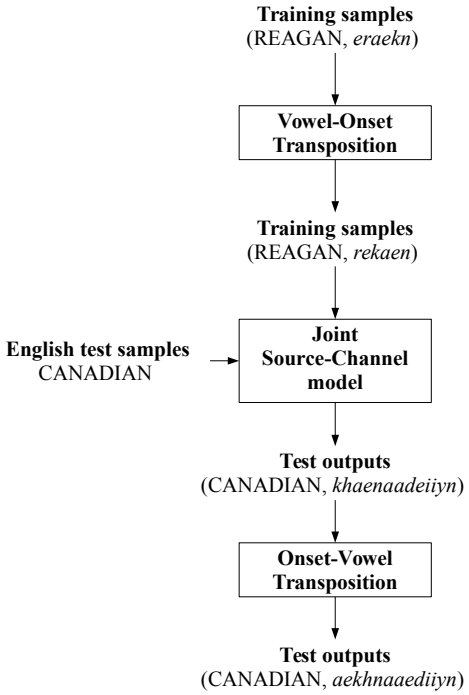

Figure 4: Augmented System. English words are capitalized, Thai words are italicized RTGS.

input : Thai word
output: Regulated Thai word
while *not end of word* **do**
  if *character is leading vowel* **then**
    swap position of vowel and onset;
    go to character after leading vowel;
  else
    go to next character;
  end
end

Vowel-Onset Transposition

via onset-vowel transposition.

5 Experiments

We used the TOP-1 metric (Banchs et al., 2015) for performance comparison between the baseline and the augmented systems. As denoted in Table 1, 75% of the NEWS2016 training set was used for training, the remaining 25% of the NEWS2016 training set was used for tuning, and the NEWS2016 dev set was used for testing. A 6-gram joint source-channel model was used for Baseline joint S-C and Augmented 1. Reordering option *'msd-bidirectional-fe'* in Moses was used for the Baseline pbSMT system as it yielded the best TOP-1 metric.

From Table 2, on the NEWS2016 test set, the augmented system (Augmented 1) achieves a 29% relative improvement over the pbSMT baseline system (Baseline pbSMT), and a 23% relative improvement over the joint source-channel baseline system (Baseline joint S-C).

To observe the performance of the augmented system with the full training data, we trained another separated augmented system (Augmented 2), using both the training set and the dev set data as denoted in Table 1. A 6-gram joint source-channel model was also used for this augmented system. Despite a simpler setup due to exploiting phonology knowledge, this system achieves comparable performance to that of systems reported in (Nicolai et al., 2015) and (Finch et al., 2015) for English to Thai transliteration task.

System	Set-up
Baseline pbSMT Baseline joint S-C Augmented 1	Train = 75% NEWS 2016 training Tune = 25% NEWS 2016 training
Augmented 2	Train = NEWS 2016 training Tune = NEWS 2016 dev

Table 1: Data partitioning for the different systems. (pbSMT: phrase-based statistical machine translation, joint S-C: joint source-channel)

System	Dev Set	Test Set
Baseline pbSMT	0.3117	0.111650
Baseline joint S-C	0.3662	0.117314
Augmented 1	0.4015	0.144013
Augmented 2	NA	0.155340

Table 2: Results on NEWS 2016 shared task in terms of TOP-1 accuracy.

6 Discussion

Besides the augmented system, we explored a rule-based approach and a phonology-augmented statistical approach for the English to Thai transliteration task. Phonology-augmented statistical approach for English to Vietnamese transliteration has been proposed in (Ngo et al., 2015). These two approaches have inspired the vowel-onset transposition strategy for the joint source-channel model. Despite being similar, our proposed approach takes into account the complex relationship between phonology and orthography, which was not considered in (Ngo et al., 2015). We are working on generalizing this relationship into a statistical model, and also to address other peculiar characteristics of Thai script.

References

Rafael E. Banchs, Min Zhang, Xiangyu Duan, Haizhou Li, and A. Kumaran. 2015. Report of NEWS 2015 Machine Transliteration Shared Task. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop*, page 10.

Maximilian Bisani and Hermann Ney. 2002. Investigations on joint-multigram models for grapheme-to-phoneme conversion. In *INTERSPEECH*.

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

Ananlada Chotimongkol and Alan W. Black. 2000. Statistically trained orthographic to sound models for Thai. In *INTERSPEECH*, pages 551–554.

Andrew Finch, Lemao Liu, Xiaolin Wang, and Eiichiro Sumita. 2015. Neural Network Transduction Models in Transliteration Generation. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop*, page 61.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2010. Integrating joint n-gram features into a discriminative training framework.

Brett Kessler and Rebecca Treiman. 1997. Syllable structure and the distribution of phonemes in english syllables. *Journal of Memory and Language*, 37(3):295–311.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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*, pages 177–180. Association for Computational Linguistics.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2015. Data representation methods and use of mined corpora for Indian language transliteration. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop*, page 78.

Peter Ladefoged and Keith Johnson. 2014. *A course in phonetics*. Cengage learning.

Haizhou Li, Min Zhang, and Jian Su. 2004. A joint source-channel model for machine transliteration. In *Proceedings of the 42nd Annual Meeting on association for Computational Linguistics*, page 159. Association for Computational Linguistics.

Sudaporn Luksaneeyanawin. 1992. Three-dimensional phonology: a historical implication. In *Proceedings of the Third International Symposium on Language and Linguistics: Pan Asiatic Linguistics*, volume 1, pages 75–90.

Hoang Gia Ngo, Nancy F. Chen, Binh Minh Nguyen, Bin Ma, and Haizhou Li. 2015. Phonology-Augmented Statistical Transliteration for Low-Resource Languages. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Garrett Nicolai, Bradley Hauer, Mohammad Salameh, Adam St Arnaud, Ying Xu, Lei Yao, and Grzegorz Kondrak. 2015. Multiple System Combination for Transliteration. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop*, page 72.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Royal Institute. 1999. Principles of romanization for thai script by transcription method.

D. Smyth. 2002. *Thai: An Essential Grammar*. Essential grammar. Routledge.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*, volume 2002, page 2002.

Jess Toms and Francisco Casacuberta. 2006. Statistical phrase-based models for interactive computer-assisted translation. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 835–841. Association for Computational Linguistics.

Ian H Witten and Timothy C Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *Information Theory, IEEE Transactions on*, 37(4):1085–1094.

Moira Yip. 2002. *Tone*. Cambridge University Press.