Syllable-based Thai-English Machine Transliteration

Chai Wutiwiwatchai and Ausdang Thangthai
National Electronics and Computer Technology Center
Pathumthani, Thailand
{chai.wutiwiwatchai,ausdang.thangthai}@nectec.or.th

Abstract

This article describes the first trial on bidirectional Thai-English machine transliteration applied on the NEWS 2010 transliteration corpus. The system relies on segmenting source-language words into syllable-like units, finding unit's pronunciations, consulting a syllable transliteration table to form target-language word hypotheses, and ranking the hypotheses by using syllable n-gram. The approach yields 84.2% and 70.4% mean F-scores on English-to-Thai and Thai-to-English transliteration. Discussion on existing problems and future solutions are addressed.

1 Introduction

Transliteration aims to phonetically transcribe text in source languages with text in target languages. The task is crucial for various natural language processing research and applications such as machine translation, multilingual text-to-speech synthesis and information retrieval. Most of current Thai writings contain both Thai and English scripts. Such English words when written in Thai are mainly their translations. Without official translation forms, transliterations often take place.

Thai-English machine transliteration and related research have been investigated for many years. Works for Thai word romanization or Thai-to-English transliteration are such as Charoenporn et al. (1999), Aroonmanakun and Rivepiboon (2004). Both works proposed statistical romanization models based on the syllable unit. Generating Thai scripts of English words are mainly via automatic transcription of English words. Aroonmanakun (2005) described a chunk-based n-gram model where the chunk is a group of characters useful for mapping to Thai transcriptions. Thangthai et al. (2007) proposed a method for generating Thai phonetic transcriptions of English words for use in Thai/English text-to-speech synthesis. The CART learning machine was adopted to map English characters to Thai phonetics. As our literature review, a general algorithm for bi-directional Thai-to-English and English-to-Thai transliteration has not been investigated.

The NEWS machine transliteration shared task has just included Thai-English words as a part of its corpus in 2010, serving as a good source for algorithm benchmarking. In this article, a Thai-English machine transliteration system is evaluated on the NEWS 2010 corpus. The system was developed under intuitive concepts that transliteration among Thai-English is mostly done on the basis of sound mimicking of syllable units. Therefore, the algorithm firstly segments the input word in a source language into syllable-like units and finding pronunciations of each unit. The pronunciation in the form of phonetic scripts is used to find possible transliteration forms given a syllable translation table. The best result is determined by using syllable n-gram.

The next section describes more details of Thai-English transliteration problems and the Thai-English NEWS 2010 corpus. The detail of proposed system is given in Section 3 and its evaluation is reported in Section 4. Section 5 discusses on existing problems and possible solutions.

2 Thai-English Transliteration

As mentioned in the Introduction, the current Thai writing often contains both Thai and English scripts especially for English words without compact translations. Many times, transliterations take place when only Thai scripts are needed. This is not only restricted to names but also some common words like “computer”, “physics”, etc.

The Thai Royal Institute (http://www.royin.go.th) is authorized to issue official guidelines for Thai transcriptions of foreign words and also romanization of Thai words, which are respectively equivalent to English-to-Thai and Thai-to-English transliteration. Romanization of Thai words is based on sound transcription. Thai con-

Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 66–70,
Uppsala, Sweden, 16 July 2010.

sonant and vowel alphabets are defined to map to roman alphabets. Similarly, English-to-Thai transliteration is defined based on the phonetic transcription of English words. However, in the latter case, an English phoneme could be mapped to multiple Thai alphabets. For example, the sound /k/ could be mapped to either "ก", "ข", "ค", or "ฆ". Moreover, the guideline reserves for transliterations generally used in the current writing and also transliterations appeared in the official Royal Institute dictionaries, even such transliterations do not comply with the guideline.

Since the guidelines are quite flexible and it is also common that lots of Thai people may not strictly follow the guidelines, ones can see many ways of transliteration in daily used text. To solve this ambiguity, both the official guidelines and statistics of usage must be incorporated in the machine transliteration system.

The Thai-English part of NEWS 2010 corpus developed by the National Electronics and Computer Technology Center (NECTEC) composes of word pairs collected mainly from 3 sources; press from the Thai Royal Institute, press from other sources, and the NEWS 2009 corpus. The first two sources, sharing about 40% of the corpus, mostly contain common English words often transliterated into Thai and the transliteration is almost restricted to the Royal Institute guidelines. The rest are English names selected from the NEWS 2009 corpus based on their frequencies found by the Google search. Such English names were transliterated into Thai and rechecked by linguists using the Royal Institute transliteration guideline.

3 Proposed Transliteration System

Our proposed model is similar to what proposed by Jiang et al. (2009), which introduced translation among Chinese and English names based on syllable units and determined the best candidate using the statistical n-gram model. The overall structure of our model is shown in Figure 1.

3.1 Syllabification and letter-to-sound

An input word in the source language is first segmented into syllable-like units. It is noted that there are some cases where segmented units are not really a syllable. For examples, "S" in the word "SPECTOR" might actually be pronounced as a single consonant without vowel. The Thai word "เสนอ"/s-a n-ɛ:/ is unbreakable as the letter expressed for the first syllable /s-a/ is enclosed in the letters of the second syllable /n-ɛ:/. These cases are considered exceptional syllables.

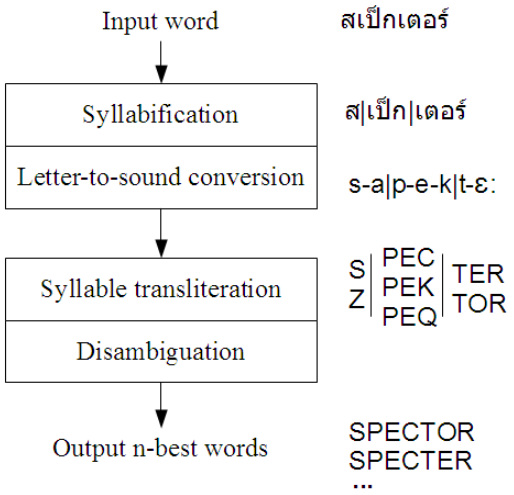

Figure 1. The overall system architecture.

In the Thai-to-English system, syllabification of Thai words is a part of a Thai letter-to-sound conversion tool provided by Thangthai et al. (2006). It is performed using context-free grammar (CFG) rules created by Tarsaku et al. (2001). The CFG rules produce syllable-sequence hypotheses, which are then disambiguated by using syllable n-gram. Simultaneously, the tool provides the phonetic transcription of the best syllable sequence by using a simple syllable-to-phone mapping. Figure 1 shows an example of an input Thai word "สเป็กเตอร์" which is segmented into 3 syllables "ส|เป็ก|เตอร์" and converted to the phonetic transcription defined for Thai "s-a|p-e-k|t-ɛ:".

In the English-to-Thai system, a simple syllabification module of English words is created using the following rules.

1) Marking all vowel letters "a, e, i, o, u",
   e.g. L[o]m[o]c[a]t[i]v[e], J[a]nsp[o]rt
2) Using some rules, merging consonantal letters surrounding each vowel to form basic syllables,
   e.g. Lo|mo|ca|ti|ve, Jan|sport
3) Post-processing by merging the syllable with "e" vowel into its preceding syllable e.g. Lo|mo|ca|tive, and re-segmenting for syllables without vowel letters, e.g. mcdo|nald to mc|do|nald, sport to s|port

Letter-to-sound conversion of English words can actually be conducted by several public tools like Festival (http://www.cstr.ed.ac.uk/projects/ festival/). However, the tool does not meet our re-

quirement as it could not output syllable boundaries of the phonetic sequence and finding such boundaries is not trivial. Instead, a tool for converting English words to Thai phonetic transcriptions developed by Thangthai et al. (2007) is adopted. In this tool, the CART learning machine is used to capture the relationship among alphabets and English phone transcriptions of English words and Thai phone transcriptions. Since the Thai phonetic transcription is defined based on the syllable structure, the syllable boundaries of phonetic transcriptions given by this tool can be obtained.

3.2 Syllable transliteration and disambiguation

In the training phase, both Thai and English words in pairs are syllabified and converted to phonetic transcriptions using the methods described in the previous subsection. To reduce the effect of errors caused by automatic syllabification, only word pairs having equal number of syllables are kept for building a syllable transliteration table. The table consists of a list of syllable phonetic transcriptions and its possible textual syllables in both languages. An n-gram model of textual syllables in each language is also prepared from the training set.

In the testing phase, each syllable in the source-language word is mapped to possible syllables in the target language via its phonetic transcription using the syllable transliteration table described above. Since each syllable could be transliterated to multiple hypotheses, the best hypothesis can be determined by considering syllable n-gram probabilities.

4 Experiments

The Thai-English part of NEWS 2010 were deployed in our experiment. The training set composes of 24,501 word pairs and two test sets, 2,000 words for English-to-Thai and 1,994 words for Thai-to-English are used for evaluation. All training words were syllable segmented and converted to phonetic transcriptions using the tools described in the Section 3.1. Since the CFG rules could not completely cover all possible syllables in Thai, some words failed from automatically generating phonetic transcriptions were filtered out. As mentioned also in the Section 3.1, only word pairs with equal number of segmented syllables were kept for training. Finally, 16,705 out of 24,501 word pairs were reserved for building the syllable transliteration table and for training syllable 2-gram models.

Table 1 shows some statistics of syllables collected from the training word pairs. Since the Thai-English word pairs provided in NEWS 2010 were prepared mainly by transliterating English words and names into Thai, it is hence reasonable that the number of distinct syllables in Thai is considerably lower than in English. Similarly, the other statistics like the numbers of homophones per syllable phonetic-transcription are in the same manner.

Total no. of syllables	39,537
Avg. no. of syllables per word	2.4
No. of distinct syllables	4,367 (Thai) 6,307 (English)
No. of distinct syllable phonetic-transcriptions	1,869
Avg. no. of homophones per syllable phonetic-transcription	2.3 (Thai) 3.4 (English)
Max. no. of homophones per syllable phonetic-transcription	16 (Thai) 38 (English)

Table 2. Some statistics of syllables extracted from the training set.

As seen from the Table 1 that there could be up to 38 candidates of textual syllables given a syllable phonetic transcription. To avoid the large search space of syllable combinations, only top-frequency syllables were included in the search space. Table 2 shows transliteration results regarding 4 measures defined in the NEWS 2010 shared task. Both experiments on English-to-Thai and Thai-to-English transliteration are non-standard tests as external letter-to-sound conversion tools are incorporated.

Measure	**Eng-to-Thai**	**Thai-to-Eng**
ACC in Top-1	0.247	0.093
Mean F-score	0.842	0.707
MRR	0.367	0.132
MAP_{ref}	0.247	0.093

Table 2. Transliteration results based on the NEWS 2010 measurement.

5 Analysis and Discussion

There are still some problematic issues regarding the transliteration format including hyphenation and case sensitivity in the test data. Ignoring both problems leads to 0.5% and 8.3% improvement on the English-to-Thai and Thai-to-English tests respectively. Figure 2 illustrates the distribution of test words and error words with respect to the word length in the unit of syllables. More than 80% of test words are either 2 or 3 syllables. It can be roughly seen that the ratio of error words over test words increases with respect to the length of words. This is by the fact that the whole word will be considered incorrect even if only a syllable in the word is wrongly transliterated. Out of 3,860 syllable units extracted from all error words, over 57% are correctly transliterated.

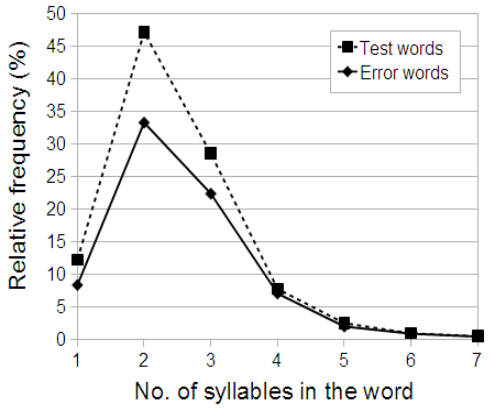

Figure 2. The distribution of test words and error words with respect to the word length.

Another issue largely affecting the system performance is as mentioned in the Section 2 that the Thai Royal Institute's guideline is somewhat flexible for multiple ways of transliteration. However, the corpus used to train and test currently provides only one way of transliteration. Improving the corpus to cope with such transliteration flexibility is needed. In developing the Thai-English NEWS 2010 transliteration corpus, some foreign names are difficult to pronounce even by linguists. Errors in the corpus are then unavoidable and required further improvement.

Many algorithms could be conducted to help improve the system accuracy. First, the current system uses only syllable n-gram probabilities to determine the best result without considering how likely the target syllable is close to the source syllable. For example, the source syllables "BIKE" and "BYTE" are transliterated to Thai as "ไบค์"and "ไบท์" respectively. Both Thai transliterated syllables are pronounced in the same way as /b-ai/. It can be seen that both syllables "BIKE" and "BYTE" can be linked to both "ไบค์" and "ไบท์". Selecting the best syllable takes only the syllable n-gram into account without considering its right transliteration. Direct mapping between source and target syllables could solve this problem but leads to another problem of unseen syllables. A better way is to incorporate in the search space another score representing the closeness of source and target syllables. As the example, the syllable "BIKE" is closer to "ไบค์" than to "ไบท์" as the letter "K" is normally pronounced like "ค" /k/, not "ท" /t^h/. We have tried incorporating such knowledge by introducing a syllable similarity score in the search space. Given a pair of source and target syllables, the syllable similarity score is the number of consonants having the same sound like "K" and "ค" divided by the total number of consonants in the syllable. Unfortunately, this approach could not yield any improvement currently as many syllable pairs happened to have the same similarity score. A better definition of the score will be conducted in the future work.

6 Conclusion

The Thai-English part of NEWS 2010 transliteration corpus was briefly described and its use in building a Thai-English machine transliteration system was reported. The system is based on transliteration of syllable units extracted from the whole input word. Within the space of candidate transliterated syllables, the best output was determined by using the statistical syllable n-gram model. There are many issues left for further improvement. First, possible transliterations of each word should be added to the corpus. Second, the system itself could be improved by e.g. incorporating better syllabification approaches, defining a better syllable similarity score, and comparing with other potential algorithms. Finally, as the Thai-to-English part of the transliteration corpus is actually back-transliteration of English-to-Thai, it is interesting to extend the corpus to cope with real-use Thai-to-English word pairs.

Acknowledgments

The authors would like to thank the Thai Royal Institute and Assoc. Prof. Dr. Wirote Aroonmanakun from the Faculty of Arts, Chulalongkorn University, who help supply parts of the Thai-English NEWS 2010 transliteration corpus.

References

Ausdang Thangthai, Chatchawarn Hansakunbuntheung, Rungkarn Siricharoenchai, and Chai Wutiwiwatchai. 2006. *Automatic syllable-pattern induction in statistical Thai text-to-phone transcription*, In Proc. of INTERSPEECH 2006, pp. 1344-1347.

Ausdang Thangthai, Chai Wutiwiwatchai, Anocha Ragchatjaroen, Sittipong Saychum. 2007. *A learning method for Thai phonetization of English words*, In Proc. of INTERSPEECH 2007, pp. 1777-1780.

Thatsanee Charoenporn, Ananlada Chotimongkol, and Virach Sornlertlamvanich. 1999. *Automatic romanization for Thai*, In Proc. of the Oriental COCOSDA 1999, Taipei, Taiwan.

Wirote Aroonmanakun and Wanchai Rivepiboon. 2004. *A unified model of Thai word segmentation and romanization*, In Proc. of the 18th Pacific Asia Conference on Language, Information and Computation, Tokyo, Japan, pp. 205-214.

Wirote Aroonmanakun. 2005. *A chunk-based n-gram English to Thai transliteration,* In Proc. of the 6th Symposium on Natural Language Processing, Chiang Rai, Thailand, pp. 37-42.

Xue Jiang, Le Sun, Dakun Zhang. 2009. *A syllable-based name transliteration system*, In Proc. of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pp. 96–99.