New Techniques in Thai-English Transliterated Words Searching, Applied to Our New Webservices Platform for Tourism (WICHAI)

Chochiang K., Hanna F., Betbeder M. L., Lapayre J. C.*
CNRS Femto-ST Institute – UMR 6174 University of France-Comte, Besancon, France.

* Corresponding author. Tel.: +33.81.66.65.15; email: jc.lapayre@femto-st.fr
Manuscript submitted December 24, 2015; accepted May 11, 2016.
doi: 10.17706/jcp.12.5.408-415

Abstract: This research proposes a new technique in Thai-English transliterated words searching by using the similar-sounding words database with similarity and relatedness calculations. The proposed search engine is used at the core of the "WICHAI" platform for tourism. The newly proposed method can help tourists, who do not speak the Thai language, to find relevant and accurate keywords on the Internet with better flexibility. The calculations use, successively, consonants and vowels comparisons based on the Royal Thai General System of Transcription (RTGS) and iterations with similar characters. Similarity calculations using 4 algorithms with a threshold of 0.6 were computed in order to obtain the most appropriate words while maintaining the level of word-meaning understanding.

Key words: Thai language, transliteration, ontology, similarity.

1. Introduction

Tourism is an important business section in many countries including Thailand. One of the major obstacles for tourists who wish to find information about a desired destination on the Internet is the language barrier. Some words, phrases and accents are only used in some specific social situations and in regions that have their own local dialect. Tourists may get some suggestions for unseen destinations from the locals and might not be able to find further information from the Internet. This is due to language barriers and particularly to the variation of sounds and pronunciation in the native language. For example, when tourists hear from locals about "Namtok Kathu", which is a famous waterfall in Phuket, they will search for it on the Internet using their Latin alphabet keyboards. Tourists will most likely search for "Namtork Kathu", "Numtoc Katu", "Numtok Kratuu" or "Narm tog Kathu", depending on their transliteration capabilities, instead of searching for the correct word "Namtok Kathu". Therefore, they will not be able to find the desired destination. The new proposed algorithm helps solving this type of problems by enhancing the system's transliterated words searching techniques.

In the globalization era, most people can connect to each other via Internet anytime and anywhere. Most search engines use keywords query, which is identical and relevant to conceptual search [1]. This means that the way the user spells a word affects the search results and thus a spelling mismatch would give inaccurate results. Unfortunately, keywords for touristic attractions in Thailand are often expressed in Thai language such as the name of touristic sites, museums, restaurants or plaza. These keywords cannot be translated into English such as "วัดจันทร์" ("Wat Chan"), which is used by locals to indicate the "Chan temple".

When translated to English, the keyword "Wat Chan" becomes "Moon Temple" and when a tourist searches the translation he will not be able to find the desired results. Therefore, tourists have to search only in transliterated keywords. Currently, word transliteration from Thai to English has fixed guidelines and rules but, due to the language barrier, these guidelines cannot be used by tourists (because they cannot speak nor understand the Thai language). Thereby, tourists may not find suitable and accurate information on touristic attractions on the Internet without the help of a personal travel guide or native Thai speakers.

Our algorithm enhances the effectiveness of search engines when transliterated words are used. This is done by adding, organizing and classifying characters into groups of consonants and vowels. A search engine using our algorithm can match and find correct or nearest and most related words as much as possible even when misspellings (with acceptable basis) are present in the search. This approach can provide more flexibility to find information on search engines. This is a powerful tool, which can reduce the search time and provide more flexibility for tourists in Thailand.

2. Transliteration for Thai Language

The consonants and vowels mapping between Thai and English are an essential key for this paper. Phonetic form (transcription) was written to make a Thai word readable for tourists who do not speak Thai. The transcription between languages is very difficult due to do phonetics distinction between vowels and consonants. The transcription is indeed composed of two parts: transcription of consonants and transcription of vowels. Basic Thai textual syllables can be represented in the form of initial consonant, a vowel, a final consonant or a tone [2]. Most of transcription problems come from character-sound mapping ambiguity, functional consonants, character ordering and implicit vowels [3]. Writing words pronunciation can be done precisely but the problem is that many special symbols are introduced and the basic knowledge of phonetics is needed [4]. Therefore, this research focuses on transliterating Thai words directly character by character. It is called the transliteration method.

The mapping of consonants and vowels used in this research is based on two resources. The first is The Royal Thai General System of Transcription [5]. The second is the similarity table. RTGS is the official system used to render Thai words into the Latin alphabet. It is published by the Royal institute of Thailand, which is authorized to issue official guidelines for Thai transcriptions of foreign words and also Romanization of Thai words. Thai Romanization could be performed on the basis of transcription, transliteration or both. The purpose of RTGS is to enable a user to read and write Thai words using the Latin alphabets that are the closest to the original Thai ones (an extract of RTGS Table is given in Fig. 1). The standard rule to convert transliterated words uses the principle of Romanization by transcription method (called Romanized characters) as proposed by the Royal Institute. Romanization of Thai words is based on sound transcription. Both consonants and vowels must be spelled and sorted grammatically correctly. Otherwise, the search engine may not find the required information. The work presented in this paper reuses the results of a research paper written in Thai [6]. Then we apply to it phonetics mapping of consonants and vowels between Thai and English. We created the similarity table, which organizes the groups of similar consonants and vowels together. Moreover, the number of similar characters is increased in order to gain more comprehensibility as shown in the similarity table part of Fig. 1. The similar characters in consonant and vowel parts are compatible with users who do not have any knowledge of linguistics. For example, to search for the word "poe", users can search with one of the following transliterations: "per", "pur", "phoe", "pher" or "phur". This table is not officially used as a standard alternative. However, it will be useful in reducing search limitations.

Prayut and Somboon [7] presented an algorithm for Thai-English cross-language transliterated word retrieval using phonetic codes retrieval based on a modified soundex coding. The soundex system is fast

and usually matches names that it should find, but often causes errors such as incorrectly matched names that do not have actually similar sounds. Prasitjutrakul *et al*. [6] proposed Thai- English transliterated word encoding for cross-language retrieval system using back propagation neural network. This encoding uses the network output to obtain a list of phonetic codes of a word. The experimental results use the K-fold cross validation technique. Canasai *et al*. [8] proposed machine transliteration which is composed of two components: the first is a group of simple strategies for generating training examples based on character alignment while the second component provides discriminative training based on the Margin Infused Relaxed Algorithm.

RTGS Table

Consonant				Vowel		
No.	Thai Alphabet	Romanized Character		No.	Thai character	Romanized Character
		Initial consonant (sound)	Final consonant (sound)			
1	ก	k	k	1	ะ, า, ั	a
	ข ฃ ค ฅ ฆ	kh		2	-ิว	io
2	จ ฉ ช ฌ	ch	t	3	-ี	i
3	ซ ส ศ ษ [สร ทร]	s	t		-ี	i
4	ป	p	p	4	เ-ียว	iao
	ผ พ ภ	ph				
5	ฝ ฟ	f	p	5	ไ-, ใ-, ไ-ย, ั-ย, -าย	ai
...	...	...	...	...	...	...

Similar Table

Consonant				Vowel		
No.	Thai Alphabet	Romanized Character		No.	Thai character	Romanized Character
		Initial consonant (sound)	Final consonant (sound)			
1	ก	ck, g, x, c, q	ck, g, x, c, q	1	ะ, า, ั	u, ar
	ข ฃ ค ฅ ฆ			2	-ิว	iu, ew, eua
2	จ ฉ ช ฌ	j, x	j, x	3	-ี	-
3	ซ ส ศ ษ [สร ทร]	s, z	s, z		-ี	ee, ei, ea, ey
4	ป	bh	-	4	เ-ียว	ieo, eaw, iow, iau, iew, iaw
	ผ พ ภ					
5	ฝ ฟ	-	f	5	ไ-, ใ-, ไ-ย, ั-ย, -าย	ie, uy
...	...	...	...	...	...	...

Fig. 1. Extract of RTGS table and similarity table (http://members.femto-st.fr/kitsiri-chochiang/).

The best performance was obtained for English to Thai and English to Hebrew. Yoshiki *et al.* [9] presented a phonetic similarity measurement method across eight Asian languages based on the Romanization, International Phonetic Alphabet (IPA). This algorithm measures similarity between a pair of language via the Levenshtein distance algorithm. Lin *et al*. [10] proposed a similarity-based framework for backward transliteration. The Widrow-Hoff rule is applied in order to obtain automatically phonetic similarities from a corpus.

3. Methodology

3.1. WICHAI Platform

WICHAI (WebservICes platform for pHuket tourism bAsed on ontologIes) is a platform for searching touristic attractions, activities and destinations in Phuket (Thailand) based on three ontologies as shown in Fig. 2. The word WICHAI means “research” in Thai language. This platform enables visitors to search for information easily and correctly. The input of the system is the words (English words or Transliterated words), which represent the touristic attractions searched by tourists. The output of the platform shows the relevant information including map, keywords, contents, contacts, pictures, etc. This platform is composed of a website, a language processing module and ontologies. The website is a key component for users to access the main functions. It provides various services such as maps, searching with English words, searching with transliterated words, etc. The website of the platform connects to the search engine. The language processing represents the main linguistic processing for English words and Transliterated words using interoperability between RTGS and similarity table. The transliterated words searching is the part that guarantees a high searching efficiency when using the platform. The last part is the ontologies, it is composed of three ontologies, which are PhuketTourism, Language and UserProfile. The first ontology, PhuketTourism, represents the various information available in Phuket such as attractions, activities, events, etc. Next, the Language ontology is the collection of relevant words, synonyms and data types to be applied with the algorithm to create the list of similar words. Finally, the UserProfile ontology stores details

of users such as the favorite sport, the travel time, etc. It is used as initial data for search mapping.

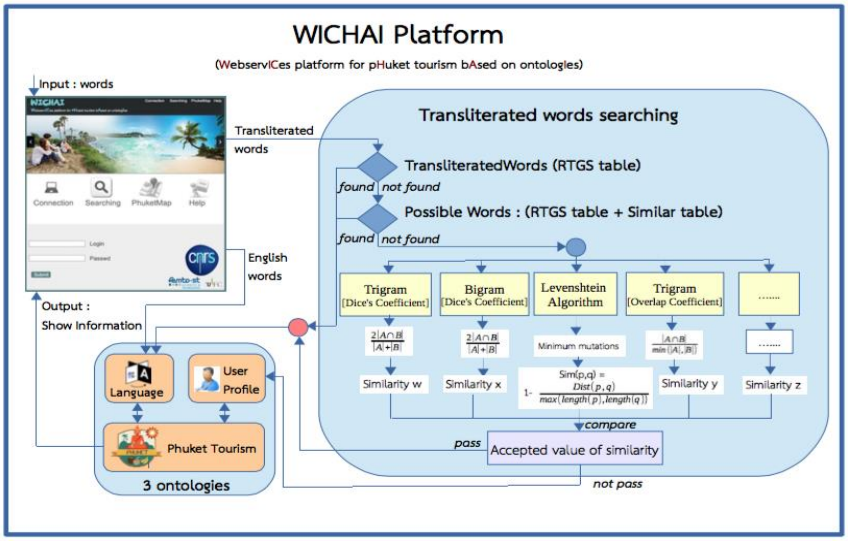

Fig. 2. WICHAI platform.

3.2. Methodology

Thai alphabet consists of 44 consonants. This research classified them into 14 groups of consonants-sound and 21 groups of vowels-sound. Each syllable in the Thai alphabet must contain both vowels and consonants. For a consonant, the transliteration differs depending on its location in the syllable. Thai consonants can be initial consonants or final consonants. The initial consonants are required in all syllables. An initial consonant might be pronounced differently when used as a final consonant. For example, the word "ดาด" [TH] can be transliterated as "DAT"[TW]. We notice that the consonant "ด" is pronounced as "D" when it is an initial consonant while it is pronounced as "T" when it is a final consonant.

Consonants in Thai cannot be pronounced without vowels. Each vowel begins with either a, e, i, o or u. Vowels in Thai are very complicated. RTGS defined the Romanized characters for consonant and vowel sounds in Thai. But it is very difficult for users who do not speak Thai. Therefore, the similarity table is applied on a set of characters that have similar pronunciation. Adding additional similar characters to a group makes it more easy and flexible to use.

From Fig. 3, the user input goes first through the Romanized Processing step, which is used to extract relevant characters in the RTGS table. If the word is correct (a match is found) then the algorithm stops. Otherwise, the Romanized and Possiblewords Processing are used to extract characters to create the word from the RTGS table and the similarity table. The output of this step is a list of possible words if the entered word was correct. The list of possible words contains the matched words between the two tables. In contrary, if the Romanized and Possiblewords Processing fails, the output contains a list of wrong words that will be next passed to the 4 algorithms step. These algorithms calculate use the Trigram_Dice (1), Bigram_Dice (1), Trigram_Overlap (2), and the Levenshtein algorithm (3). This is very useful when we have errors resulting from wrong keyboard input as they are often of the same kind as the allowed edit operations.

Definitions:

A: the *cardinality* of A, denoted |A| counts how many elements are in A

B: the *cardinality* of B, denoted |B| counts how many elements are in B

Dice's coefficient: It is a statistic approach used to compare the similarity of two samples. It can be used

to measures the number of common n-grams (bi-gram, tri-gram,...)

$$Sim(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

Overlap coefficient: The overlap coefficient is a similarity measure related to the Jaccard index that measures the overlap between two data sets, and is defined as the size of the intersection divided by the smaller size of the two sets:

$$Sim(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)} \tag{2}$$

Levenshtein Distance: It is a string metric for measuring the difference between two sequences (A and B). It calculates the least number of edit operations that are necessary to transform one string to the other.

$$Sim(A,B) = \frac{Dist(A,B)}{\max(length(A),length(B))} \tag{3}$$

On the other hand, 4 algorithms calculate the similarity index of its two arguments based on the edit distance Equation (1)-(3): "0" means that the strings are entirely different while "1" means that they are identical.

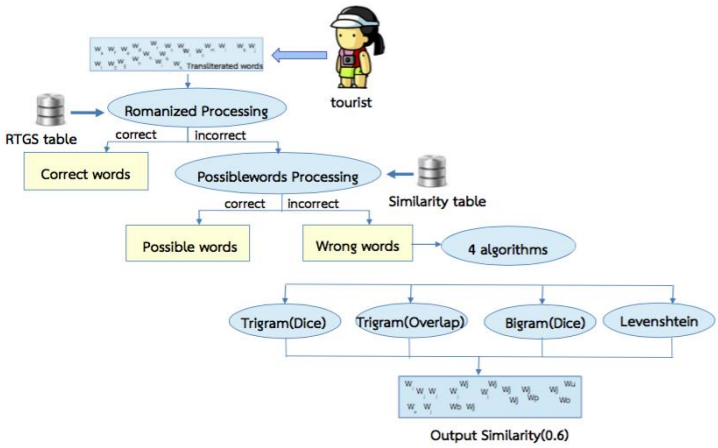

Fig. 3. Overview of the transliterated words searching.

4. Experiment

The data set used in our research is a list of Thai words that covers the 14 groups of consonants and the 21 groups of vowels. There are different patterns with a total of 114 syllables. The experiments were conducted on a group of 18 students (from different domains) in the University of France-Comte, Besancon, France. They all come from different nationalities but do not have any knowledge of Thai linguistic. This experimental group has the same profile as tourists in Thailand because they come from different countries and they all speak the English language. The words we use in our data set consist of initial consonants, vowels and final consonants (optional). In our experiments we compare on the basis of syllables and not only the whole word. This helps us to understand with which consonants or vowels a foreign user has more problems. This process requires an in-depth analysis of alphabets. These experiments resulted in 2043 syllables to analyze: 114 (initial data set syllables) * 18 (number of students undergoing the experiments) - 9 (missing data or incomplete words) = 2043 syllables.

We use 4 algorithms to find the nearest output using similarity with the suitable threshold. This

threshold is consistent with the results proposed by [11]: Authors presented the quality of similarity calculations. Using a similarity threshold of 0.6 we can successfully separate relevant and irrelevant words. Thus, we calculate using a similarity threshold of 0.6 with Trigram(Dice), Trigram(Overlap), Bigram(Dice) and Levenshtein as shown in Table 1.

Table 1. Average Similarity for 4 Algorithms

Algorithms	Trigram(Dice)	Trigram(Overlap)	Bigram(Dice)	Levenshtein
Average similarity	0.62	0.64	0.61	0.75

From Table 1, Levenshtein gets the highest value of average similarity. So, we present, in Table 2, the information from each step using the Levenshtein algorithm (Romanized processing, Romanized and Possiblewords processing and Romanized, Possiblewords and calculate_similarity (0.6)).

As shown in Table 2, the Romanized processing step extracted only 511 syllables which represents 25.01% of the data set (2043 syllables). The combination between the Romanized and the Possiblewords processing can give more correct output by extracting 826 syllables which represents 40.43% of the data set.

Table 2. Results Comparison for All the Steps

Method	**Number of Syllables(words)**			**Percentage(%)**		
	Correct	**Incorrect**	**Total**	**Correct**	**Incorrect**	**Total**
Romanized processing	511	1532	2043	25.01	74.99	100.00
Romanized and Possiblewords processing	826	1217	2043	40.43	59.57	100.00
Romanized, Possiblewords and calculatesimilarity(0.6)	1742	301	2043	85.27	14.73	100.00

We performed an in-depth analysis of resulting syllables separating them into initial consonants, vowels and final consonants, as shown in Table 3.

Table 3. Results from Romanized and Possiblewords Processing

Method		**Correct**		**Incorrect**	**Total**
		Romanization	**Possiblewords**		
initial	Number of Syllables	1268	470	305	2043
	Percentage	62.07	23.00	14.93	100.00
vowel	Number of Syllables	1021	152	870	2043
	Percentage	49.98	7.44	42.58	100.00
final	Number of Syllables	517	32	304	853
	Percentage	60.61	3.75	35.64	100.00

To summarize, Table 2 shows that using Romanized processing, Possiblewords processing and similarity calculation with a threshold of 0.6 gives the most efficient results 85.27% (1742 datasets of 2043). The result is excellent for users who lack of Thai language skills but they can still search with Thai words (transliterated words) using a Latin keyboard.

5. Conclusion

In this paper, we proposed a word searching algorithm based on the "WICHAI" platform. Thai-English transliterated words searching using the similarity table can improve search efficiency. This algorithm uses both RTGS table and similarity table in order to extract the possible words and match user requirement. Similarity calculations with the Trigram(Dice), Trigram(Overlap), Bigram(Dice) and Levenshtein algorithms were applied. The Levenshtein algorithm using the threshold of 0.6 is the best and gives the maximum matching for relevant words in Thai-English transliteration. This work is useful for tourists in Thailand who do not speak Thai. Using our platform, they can search all the words they hear from locals

easily using a Latin keyboard anytime and anywhere. They can obtain information with more flexibility by using this newly proposed algorithm to overcome the language barrier and enjoy their stay in Thailand.

Our first experiments gave interesting results. For the future work, we have developed a webpage (http://kizzyy.seniorproject-te.com/test_css/indexen.html) and a database to conduct experiments on a large scale of population to obtain more significant results. We plan to conduct our experiments on two populations of people who speak the English language (native and non native speakers) in order to be able to compare the performance of information retrieval and add more characters in the same group (Fig. 1). In addition, we will calculate the similarity using other algorithms such as Hamming, JaroWinkler, ... This will enable us to find the best similarity calculation algorithm to use for Thai-English transliterated words searching.

References

[1] Jatsada, S., & Suphakit, N. (2013). A method for measuring keywords similarity by applying jaccard's, n-gram and vector space. *Lecture Notes on Information Theory, 1(4)*, 159-164.

[2] Wutiwiwatchai, C., & Thangthai, A. (2010). Syllable-based thai-english machine transliteration. *Proceeings of the Named Entities Workshop* (pp. 66-70).

[3] Ausdang, T., Chatchawarn, H., Rungkarn, S., & Wutiwiwatchai, C. (2006). Automatic syllable-pattern induction in statistical thai text-to-phone transcription. *Interspeech-ISCA, Pittsburgh*, PA, USA, 1344-1348.

[4] Charoenporn, T., Chotimongkol, A., & Sornlertlamvanich, V. (1999). Automatic romanization for Thai. *Proceeding of the 2nd International Workshop on East-Asian Language Resources and Evaluation* (p. 4).

[5] RTGS. (1999). The royal Thai general system of transcription. From http://www.royin.go.th

[6] Somchai, P., Tasanawan, S., & Boonserm, K. (2000). Thai-English transliterated word encoding cross-language retrieval using neural networks. *Proceedings of National Computer Science and Engineering Conf*.

[7] Suwanvisat, P., & Prasitjutrakul, S. (1998). Thai-English cross-language transliterated word retrieval using soundex technique. *Proceedings of the National Computer Science and Engineering Conf.,* Bangkok, Thailand.

[8] Kruengkrai, C., Charoenporn, T., & Sornlertlamvanich, V. (2011). Simple discriminative training for machine transliteration. *Proceedings of Named Entities Workshop* (pp. 28-32).

[9] Yoshiki, M., Ohnmar, H., & Shigeaki, K. (2011). Cross-language phonetic similarity measure on terms appeared in Asian language. *International Journal of Intelligent Information Processing*, *2(2)*, 9-21.

[10] Lin, W.-H., *et al.* (2002). Backward machine transliteration by learning phonetic similarity. *Proceedings of the 6th Conference on Natural Language Learning* (pp. 1-7). Association for Computational Linguistics.

[11] Carlos, A. H., Francisco, N. A, Viviane, K., & Moreira, O. (2007). Simeval: A tool for evaluating the quality of similarity functions. *Proceedings of the 26th Intern. Conf. on Conceptual Modeling* (pp. 71-76). ER 2007 Auckland, New Zealand.

Kitsiri Chochiang is a PhD student at Femto-st CNRS Institut of Franche-Comte University (France). She works especially in the Department of Computer Science for complex systems. Her current research focuses on ontology, linguistic and web service. From this work she developed an interest ontology-based reasoning and its applications in tourism using Thai-English transliteration to search effectively.

Fouad Hanna is a PhD student at Femto-st CNRS Institut of Franche-Comte University (France). He works especially in the Department of Computer Science for complex systems. His principal domain of research is the distributed systems and especially managing data consistency in collaborative environments. He is also interested in working with ontologies for collaborative distributed platforms.

Marie-Laure Betbeder is Associate Professor at Femto-st CNRS Institut of Franche-Comte University (France). She works especially in the Department of Computer Science for Complex Systems.

For the last ten years, she has been involved in the domains of technology enhanced learning and computer supported collaborative learning. Her early works were focused on the taylorability of computer artifacts as a support for learning environments. She is now working on the use of ontologies for the CSCW domain (computer supported collaborative work).

Jean-Christophe Lapayre is Professor at Femto-st CNRS Institut of Franche-Comte University (France). He works especially in the Department of Computer Science for complex systems. At the Biomedical Superior Institute of Engineers (ISIFC) of Franche-Comte, he is the Responsible for the 3rd year.

In research, his general field is the distributed systems and his present research interests are on ontologies used in distributed collaborative platforms.