Thai-English Cross-Language Transliterated Word Retrieval using Soundex Technique [*]

Prayut Suwanvisat
Graduate Student
g40psw@cp.eng.chula.ac.th

Somchai Prasitjutrakul
Assistant Professor
somchaip@chula.ac.th

Department of Computer Engineering
Chulalongkorn University
Bangkok 10330, Thailand
Tel : (66-2)-218-6981
Fax : (66-2)-218-6955

Abstract

This paper presents an algorithm for Thai-English cross-language transliterated word retrieval. The algorithm enables retrieval of documents containing either the English keywords or the corresponding English-to-Thai transliterated words. This is done by retrieving documents based on phonetic codes of keywords rather than the keywords themselves. The phonetic coding is based on the Soundex coding of Odell and Russell where the encoding table is slightly modified to incorporate Thai characters and the code is extended to unlimited length. Experimental results showed that a high recall and precision of more than 80% can be achieved especially when the phonetic codes are longer than four.

1. Introduction

Text retrieval has become one of the most essential tools for managing information as computer-generated documents get published and computers get connected locally and globally, especially with the advent of the World Wide Web and CD-ROM. The effectiveness of any text retrieval system is commonly measured in terms of precision and recall [3] where precision is the ratio of the number of relevant documents retrieved over the total number of documents retrieved, and recall is the ratio of relevant documents retrieved for a given query over the number of relevant documents for the query in the textbase. One problem arises when a user enter query keywords that are in one language where the documents to be managed are in another. For example, searching for documents containing "*ALEXANDER*" does not return documents containing "*อเล็กซานเดอร์*" (the corresponding Thai transliterated word). This problem causes recall of the retrieval to be lower than it should be.

Here we are interested in Thai-English cross-language transliterated word retrieval. Cross-language information retrieval is defined as the retrieval of documents when the language in which the documents are expressed is not the same as the language in which the queries are expressed [7]. It is very common in Thai documents that most of English proper nouns and technical terms appearing in the documents are either in English or transliterated into Thai. Using bilingual dictionaries as thesauri of the retrieval system does not solve the problem since most of the transliterated words are not found in the dictionaries [4]. Therefore, querying using one language will miss documents containing corresponding keywords in the other language if the system does not support the cross-language feature.

There are previous researches working on the problem. The algorithms presented in [5] and [6] transcribe English and Japanese words into intermediate codes and use exact code matching during retrieval. Since transliterating the two languages back and forth loses some information, two corresponding words may not be exactly matched . Whereas the algorithm in [2] encodes each Katakana word into a phonetic string representation and uses partial matching with English words. Two words are considered to be in transliteration relation when the number of matched characters is more than a certain threshold. The algorithm uses a depth-first search which trends to take longer time than a straightforward matching so that some heuristics are incorporated to reduce search time. [8] presents an algorithm for encoding English word to a set of possible Thai sounds using a set of encoding tables and rules. The encoding tables are not fully elaborated and no details on effectiveness of the methods are reported.

In this paper, we present an algorithm supporting Thai-English cross-language transliterated word retrieval. In other words, the system enables retrieval of documents containing either the English keywords or the corresponding English-to-Thai transliterated words. By slightly modifying coding table and algorithm of the Odell and Russell's Soundex code [1], a higher recall on document for cross-language retrieval queries on transliterated words is achieved with good precision. The rest of the paper is organized as follows. Section 2 explains the Odell and Russell's Soundex coding and algorithm. Section 3 presents our new modified encoding

[*] This research was supported by the Thai Government Research Fund.

algorithm with experimental results presented in Section 4. Then the paper is concluded in Section 5.

2. Odell and Russell's Soundex Algorithm

M. K. Odell and R. C. Russell designed a system to encode names based on their pronunciation so that names that sound alike would have the same phonetic code. Their system is called Soundex [1]. The system is based on the idea that English names can be distinguished based only on the consonants. It constructs the phonetic code by converting each letter (ignoring the leftmost letter) into a numeric code using the coding table shown in Table 1. Then all the zeros are removed and any runs of the same digits are reduced to one digit. The final Soundex code is the first (left-most) letter of the word followed by the first three digits of the converted code. For example, ALEXANDER is converted to A425.

Table 1. Soundex Coding Table

Letter	*Numeric code*
A E I O U H W Y	0
B F P V	1
C G J K Q S X Z	2
D T	3
L	4
M N	5
R	6

The Soundex system is fast and usually matches names that it should find, but often causes false hit i.e., incorrectly matches names that are not actually sound alike.

3. Our Proposed Encoding Method

There are 44 consonants in Thai which can be categorized into 21 phonetic groups as shown in Table 2 [9]. To incorporate Thai letters into the Soundex coding table, we can assign an English letter to each group having the similar phonetic, then the 21 groups are further grouped into seven groups according to phonetic similarity of the letters in the Odell and Russell's Soundex coding table as shown in Table 3. We propose to modify the original Soundex coding (shown in Table 4) as follows :

- use numeric representation for the first letter rather than the letter itself in the code. This is due to the fact that there are many cases where more than one English letter can be mapped to the same Thai letter [10], e.g., V and W are mapped to ว. as shown as the note #2 in Table 4. Therefore we need to introduce three more numbers (7, 8, 9) in the table.

- add another numeric code for the Thai letter ง since this is the only one left unencoded from the Thai consonant. The transliteration to ง is normally from the letters NG or NK e.g. KING is transliterated to คิง. Therefore its corresponding code is 52 (5 is for N, and 2 is for both G and K).

- extend the code length to be unlimited rather than of length four in the original Soundex code. (This is for enchancing precision but potentially will reduce recall of retrieval.) And also set the minimum limit of code length to be *k*, i.e, that is to only considered words having code length of length not less than *k*, where *k* is a parameter to be determined later.

Note that all Thai vowels and tones are all ignored as being done in English for the Soundex coding. (Remember that we mainly concern with the English to Thai transliteration.)

Table 2. The 21 phonetic groups of Thai consonants

ก	ฎ ด	ฝ ฟ
ข ฃ ค ฅ ฆ	ฏ ต	ม
ง	ฐ ฑ ฒ ถ ท ธ	ร
จ	ณ น	ล ฬ
ฉ ช ฌ	บ	ว
ซ ศ ษ ส	ป	ห ฮ
ญ ย	ผ พ ภ	อ

Table 3. The seven similarly phonetic groups according to Soundex code table

English	*Thai*
A E I O U H W Y	อ ห ฮ ว ญ ย
B F P V	บ ฝ ฟ ป ผ พ ภ ว
C G J K Q S X Z	ข ฃ ค ฅ ฆ ฉ ช ฌ ก จ ซ ศ ษ ส
D T	ฎ ด ฏ ต ฐ ฑ ฒ ถ ท ธ
L	ล ฬ
M N	ม ณ น
R	ร

Table 4. The new modified Thai/English coding table

English	*Thai*	Code	Note
A E I O U H W Y	อ	0	#1
B F P V	บ ฝ ฟ ป ผ พ ภ ว	1	
C G J K Q S X Z	ข ฃ ค ฅ ฆ ฉ ช ฌ ก จ ซ ศ ษ ส	2	
D T	ฎ ด ฏ ต ฐ ฑ ฒ ถ ท ธ	3	
L	ล ฬ	4	
M N	ม ณ น	5	
R	ร	6	
A E I O U	อ	7	#2
H	ห ฮ	8	#2
W	ว	1	#2
Y	ย ญ	9	#2
	ง	52	

#1 : for the first (leftmost) letter of the word
#2 : for the second letter and the rest of the word

4. Experimental Results

We implemented the encoding algorithm presented in the previous section by slightly modifying the encoding table and algorithm in the Soundex coding function in [1]. Then the algorithm is tested using a set of 1,902 pairs of English and English-to-Thai transliterated words which are mostly proper names (brand names, country names, scientist and mathematician names, etc.) and technical terms in science, mathematics, and chemistry obtained from [10], [11], [12], and [13].

We ran experiments by having all the words (and their corresponding phonetic codes presented in the previous section) stored in the database and then querying all of the words, one by one, to measure recall and precision of the retrievals. The experiments were tested repeatedly by varying the minimum limit of code length, *k* presented in the previous section, in order to show how the minimum limit of code length affects effectiveness of the retrieval. The experimental results are shown in Figure 1. In the set of tested words, the number of words whose code length is more than seven accounts for only 1% so we do not include the results in the plots which may mislead the interpretation.

From the plot in Figure 1, we can notice that recall of the retrieval is very high around 90% and then starts to drop slowly as the minimum limit of code length increases. This is in the opposite behavior for the precision which starts at around 45% and then climbs sharply to around 80% as we increase the minimum limit of code length. This observed behavior is the nature of recall and precision which grows in the opposite direction. However, in this particular problem and proposed coding algorithm, the recall slowly declines as precision gets sharply improved by increasing the minimum limit of code length especially after $k > 4$. So the proposed algorithm is generally effective for words having code length greater than 4, i.e., for words of length longer than approximately seven characters (this is concluded by an observation from the length of tested words and their corresponding code lengths).

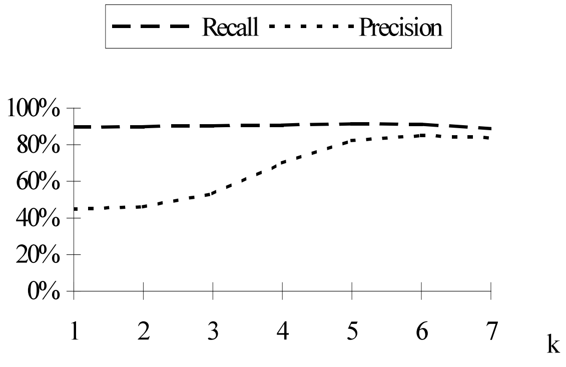

Figure 1 Plots of recall and precision against the minimum code length

5. Conclusion

In this paper, we presented an algorithm for Thai-English cross-language transliterated word retrieval. The retrieval is done by using phonetic codes retrieval based on a modified Soundex coding rather than searching for the words themselves. The system enables retrieval of documents containing either the English keywords or the corresponding English-to-Thai transliterated words. Words that are generally get retrieved by this approach are proper nouns and technical terms. In the proposed algorithm, we modify the Soundex coding algorithm of Odell and Russell by slightly modifying the encoding table and extending the code to unlimited length. Experimental results showed that a high recall and precision of more than 80% can be achieved especially when the phonetic codes are longer than four.

References

[1] A. Binstock and J. Rex, *Practical Algorithms for Programmers*, Addison Wesley, 1995.

[2] N. Collier, A. Kumano, and H. Hirakawa, "Acquisition of English-Japanese proper nouns from noisy-parallel newswire article using Katakana matching", *Proc. of the Natural Language Processing Pacific Rim Symposium 1997*, Phuket, Thailand, Dec. 2-4, pp. 309-320.

[3] W. Frakes, "Introduction to Information Storage and Retrieval System", *Information Retrieval : Data Structures & Algorithms*, W.B. Frakes and R. Baeza-Yates ed., Prentice Hall, 1992.

[4] K. Knight and J. Graehl, "Machine Transliteration", *Annual Meeting of the Association for Computational Linguistics (ACL-97/EACL-97)*.

[5] A. Kumano, "Building a technical term dictionary with Katakana-English Matching", *Gengoshorigakai - Annual Conf. of the Japanese Association for Natural Language Processing*, (in Japanese) Japan, March, pp.221-223.

[6] Y. Matsuo and S. Shirai, "Using pronunciation to automatically extract bilingual word pairs", *Shizengengoshori*, (in Japanese), November, pp.101-106.

[7] D. Oard and B. Dorr, "A Survey of Multilingual Text Retrieval", *Technical Report UMIACS-TR-96-19 CD-TR-3615*, University of Maryland, College Park, April 1996.

[8] S. Ongroongruang, R. Prongsirivattana, and V. Jantarasukree, "English to Thai Word Retrieval Using Sound Index", *Proc. 2nd SNLP'95*, Bangkok Thailand, Aug. 2-4, 1995, pp. 407-413.

[9] P. Prapapitayakorn and et. al., *รู้จักภาษาไทย* Odien Bookstore, 1976

[10] Royal Academy, *หลักเกณฑ์การทับศัพท์* (Transliteration Guideline), 1992.

[11] Royal Academy, *ศัพท์วิทยาศาสตร์* (Science Dictionary), 1993.

[12] Royal Academy, *ศัพท์คณิตศาสตร์* (Mathematics Dictionary), 1997.

[13] Royal Academy, *หนังสือเรียนวิชาเคมี เล่ม 1 หลักสูตรมัธยมปลาย 2524* (Chemistry Book 1: High School Level 1981), 1987.