Dictionary-based Thai CLIR: Experimental Survey of Thai CLIR

Jaruskulchai Chuleerat
Department of Computer Science
Faculty of Science
Kasetsart University
fscichj@ku.ac.th

Abstract

This paper describes our work, which participated in the Cross-Language Information Retrieval (CLIR) at the Cross-Language Evaluation Forum. Our objectives for this experiment have three folds. Firstly, the coverage of the Thai-bilingual dictionary was evaluated when translating queries. Secondly, whether the segmentation process has effected the CLIR. Lastly, this research investigates the query formations techniques. Since this is the first international experimental in CLIR, our approach used dictionary-based technique to translate Thai queries into English queries. Four runs are submitted to the CLEF: (a) single mapping translation with manual segmentation, (b) multiple mapping translation with manual segmentation, (c) single mapping translation with automatic segmentation and (d) Single mapping with query enhancing with the Thai thesaurus words.

The retrieval effectiveness is worse than our expected. The simple dictionary mapping technique is unable to achieve the retrieval effectiveness, although the dictionary lookup gave very good high percentage of mapping word. The words from the dictionary lookup are not specific terms but each is mapped to a definition or meaning of that term. Furthermore, Thai stopword, stemmed word and word separation have effected in Thai CLIR.

1. Introduction

Most of the CLIR research community believes that CLIR would be useful for people who do not speak a foreign language well. Unfortunately, some of the Thai CLIR hasn't evaluated their results with proper data. Thus, we participate in the Cross-Language Evaluation Forum (CLEF) as an opportunity for us to better understanding the issues in the research of the Cross-Language Information Retrieval (CLIR). We performed four Thai-English cross language retrieval runs. Our approach to the CLIR was to translate the Thai topics into English by using dictionary mapping. The bilingual dictionaries are LEXiTRON [8], and Seasite [6]. These two dictionaries were compiled by the Software and Language Engineering Laboratory, National Electronics and Computer Technology Center (NECTEC) [1] and Northern Illinois university [6].

The objectives of these four runs are follows: to survey the available of Machine Readable Dictionary (MRD) and the coverage of the vocabulary, to investigate the query formation techniques, to explore the possibility of automatic translation, and to enhance query by using Thai Thesaurus.

According to the Thai CLIR's objective, the four official runs are the single dictionary mapping, multiple dictionary mapping, manual and automatic segmentation, query expansion using Thai thesauruses.

The shareable or public MRDs are LEXiTRON from Software and Language Engineering Laboratory, National Electronics and Computer Technology Center and Seasite from Northern Illinois University were used in our experiment.

The rest of this paper is as fellows. Section 2 briefly reported the related fields in the Thai text information retrieval. Summaries of related Thai CLIR resources are given. Thai CLIR experimental design is described in section 4. Experimental results are presented in the last section.

2. Related Works

In this section, the Thai computer processing and the Thai Natural Language process is briefly discussed for understanding the current technology, which play an importance in the CLIR.

2.1 Computer Processing of Thai

Attempts to work with the Thai language on the computer started when computers were frist introduced into the country more than four decades. There are no any special characters to separate words from phrase and sentences in the Thai writing system. To overcome this problem, artificial intelligent, natural language processing, and computational linguistic are exhaustively studies. The accomplishment of these studies established of

machine translation project by the National Electronics and Computer Technology Center (NECTEC) in 1980 [10]. Additionally, a number of research output has been commercially promoted, for example, hand-held electronic dictionaries and translators from English to Thai (Pasit) [9], the Thai spelling checking software and the word segmentation programs.

2.2 Thai Text Retrieval

Most of Thai Text Retrieval system is always coordinated with the segmentation algorithms. Automatic extraction keyword from the documents is nontrivial task. Trie Structure along with dictionary based word segmentation are proposed in [15] to solve the unknown words. However, only the indexing process is presented, there is no report on the retrieval effectiveness. The work done in [2] was more contributed in the information retrieval method. The paper presented a number of comparison in segmentation process, the indexing techniques and term weighting system for Thai text retrieval. Three methods of indexing are proposed, ngram-based, word-based and rule-based. When applying term weight system, the segmentation process does not much effect the retrieval performances. All the performance metric is tested on the Thai news. The collection size is about 8 MB and 4800 documents. Additionally, the environment for testing the hypothesis used SMART text retrieval system from [16]. The other indexing technique, the signature file, has been proposed for indexing for Thai Text [17]. This paper studied the number of bit for representing the each document signature and the test collection was from Thai Holy Bible.

2.3 Works in Thai CLIR

There are some Thai research papers [3, 11-12], which presented their work in the area of CLIR. All of their techniques are based on the transliterated words. The research paper in [11] presented transliterated word encoding algorithm and creating 5000 Thai English personal names. Then, the retrieval process is against with this database. This paper claimed that the CLIR effectiveness is 69 and 73% in precision and recall. The second paper [3] is also from the same research lab to achieve a better precision and recall over 80% in the CLIR. Their CLIR model retrieved document containing either the English or Thai transliterated words using phonetic codes for keywords and the phonetic coding is based on Soundex coding of Odell and Russell. Their result of experiment is compared with the Thai-English transliterated words which are collected from Royal Academy in transliteration Guideline, Science Dictionary, mathematics Dictionary, Chemistry Boook1: High School Level. Most of those words are proper nouns, and technical terms. The last paper [12] also presented the transliteration from Thai to English for solving the loan words. This paper are more emphatic solving loan word problems such as non-native accent, information losing and orthographic translation. There are two processes to identify load word. First, the explicit unknown words are recognized by mapping with the Thai dictionary. Secondly, the hidden unknown words, which are composed of one or more known words, are identify by frequency checking. However, it is unclear how these algorithms are applied to work with CLIR.

In Asian CLIR research, the dictionary-based method is the well-known method and the query translation strategy is employed. The work done in [14], also employed the dictionary-based method for Indonesian-English Cross-Language Text Retrieval. The local-feedback techniques are applied to expand the queries terms for improving the retrieval effectiveness. Their research is conducted on TREC's data. Chen and his colleges worked on the Japanese English cross language. They stated the segmentation problem of Japanese language, which contain a number of technical terms. To increase the number vocabulary, the parallel corpus is employed. They stated that the retrieval effectiveness of CLIR is effect by the coverage of term in the dictionary

3. Resource available for Thai CLIR

The most important resource for the CLIR is bilingual dictionary. In our survey of the bilingual electronic dictionary, a number of Thai-English bilingual electronic dictionaries are found, for examples:- the Thai internet education project [7], an Online Thai Dictionary (Seasite) [6], and LEXiTRON [8]. Only the last two dictionaries are able to get the whole electronic form. However, the Seasite dictionary need to be reencoded since the original encoding system is different from the current system. The total number of words in each system is 16,060 and 11,188 words from LEXiTRON and Seasite respectively. The electronic format of Thai thesaurus is not available to share for public. We prepared our own Thai thesaurus from [13]. Around 20,000 Thai thesaurus words are collected and used in this research.

Another important resource is the Thai segmentation program. Processing of Thai language has been working for more than 3 decades. The free resource for breaking phrase or sentence into words is the wordbreak from NECTEC [1] and from University of Massachusetts [2]. In [2], wordbreak, which is from NECTEC, gave the best the effectiveness of segmentation process.

Therefore, in our initiative CLIR research, we deployed the LEXiTRON, and Seasite for checking the coverage of the number of vocabulary used in the automatic query translation. Additionally, to be able to automatically translation, the segmentation process needs to be verified and the wordbreak (Swath) from the NECTEC is deployed in Thai CLIR research.

4. Experimental Design

The Thai CLEF is aimed at the bilingual task. English documents are retrieved from the Thai topics.
Since Thai language is not an official language in the CLEF, no topic is provided by CLEF. Thus, the CLEF's English topics were chosen and translated by manual into two types of Thai queries. One is segmented Thai queries by human and another is like normal Thai writing system. The Swath's NECTEC is used to break phrase or sentence of the unsegmented Thai queries into words. The disadvantage of manual translation is that it relies on human judgment and may be bias. Then we apply the dictionary mapping techniques to translate the Thai queries back to English queries. In the dictionary lookup process, if any words are able to lookup, the process will leave that word from the topic. Thus, the concept terms or relevance terms may not include the topics. The four official runs, which are rely on the query formation, are as follows.

(a) Single Mapping: The bilingual Thai-English definition trend to give several senses or meanings. Thus, English queries are translated by using single dictionary mapping, and only the first map is selected for translation.
(b) Multiple Mapping: Since the first map was not always to give the right translation. This second run, English queries are replaced with all meaning found in the dictionary.
(c) Single Mapping and Segmentation: The unsegmented Thai queries are segmented using the NECTEC wordbreak program and single mapping is applied for the query translation.
(d) Query Expansion: Thai thesaurus words are added to the single mapping queries. The process of expanded query terms is done before translation.

For testing the coverage of the number of terms in electronic dictionary, SEASITE dictionary is used to translate English queries. Figure 1 shows our experimental design. The SMART system from Cornell University [16] is used to measure the retrieval effectiveness of our Thai CLIR. In all runs, stop words and stemming were applied to query and text collection. The term weight was applied to the document collection.

Documents collection, the Los Angeles Time of 1994, is indexed using the SMART vector model. English query is indexed based on the long query format, or on the descriptions, <DESC> marked tag. Although SMART is based on the vector model, we do not modify the original topics. When a query was sent to the system, the 1,000 highest-ranked records are returned.

The dictionary terms of the dictionary mapping algorithm are loaded into MySQL database. The mapping algorithm is deployed using Java technology and running on Linux Environment.

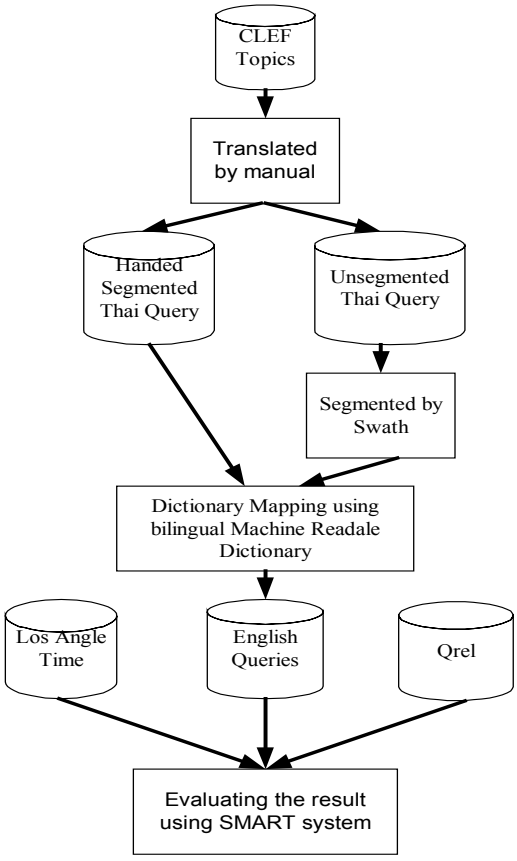

Figure 1. The Thai CLIR experimental

4. Results

We have learned from [17], the retrieval effectiveness of CLIR which is based on the dictionary mapping, will drop about half. For Thai CLIR, the retrieval results are worse when the Thai-CLIR was tested with CLEF's text collection. This principle of Thai CLIR has been experimented with ZIFF's TREC collection. The retrieval results dropped about 40% [18]. This section reports the Thai CLIR results.

(Result in the paper is slightly changed from what we had submitted to CLEF. We submitted the wrong data of the first runs, single mapping)

4.1 The Coverage of the Vocabulary of the Dictionary

Using Dictionary	LEXiTRON	SEASITE
Total Words	2864	2864
Total Topics	50	50
Total Word Found	2778	2746
Total Word Not-Found	86	118
Average Word Per Topic	58	58
Average Word Found Per Topic	55	54
Average Word Not-Found Per	2	3

Figure 2. The Effective of Dictionary

Recall Level	Mono (Eng)	Man Lex.	Man. Sea	Query Exp.	Mul. Lex.	Seg.
at 0.00	0.2932	0.0041	0.0065	0.0015	0.0014	0.0120
at 0.10	0.2111	0.0016	0.0054	0.0007	0.0004	0.0054
at 0.20	0.1583	0.0012	0.0040	0.0004	0.0002	0.0016
at 0.30	0.1207	0.0011	0.0039	0.0003	0.0000	0.0014
at 0.40	0.1035	0.0010	0.0034	0.0000	0.0000	0.0014
at 0.50	0.0864	0.0008	0.0032	0.0000	0.0000	0.0012
at 0.60	0.0680	0.0008	0.0031	0.0000	0.0000	0.0007
at 0.70	0.0574	0.0005	0.0031	0.0000	0.0000	0.0007
at 0.80	0.0422	0.0005	0.0031	0.0000	0.0000	0.0007
at 0.90	0.0335	0.0005	0.0031	0.0000	0.0000	0.0007
at 1.00	0.0294	0.0005	0.0031	0.0000	0.0000	0.0007
Rel_ret	589	41	54	26	24	46
Avg:	0.1094	0.0012	0.0038	0.0003	0.0002	0.0026

Table 1. Retrieval Effectiveness of Thai CLIR (The frist four runs are the official runs)

Recall Level	Mono (Eng)	Man Lex.	Man. Sea	Mul. Lex.	Seg.
at 0.00	0.2932	0.0529	0.0113	0.0529	0.0311
at 0.10	0.2111	0.0237	0.0075	0.0237	0.0071
at 0.20	0.1583	0.0145	0.0071	0.0145	0.0064
at 0.30	0.1207	0.0109	0.0060	0.0109	0.0049
at 0.40	0.1035	0.0086	0.0060	0.0086	0.0041
at 0.50	0.0864	0.0052	0.0057	0.0052	0.0031
at 0.60	0.0680	0.0035	0.0050	0.0035	0.0023
at 0.70	0.0574	0.0027	0.0043	0.0027	0.0018
at 0.80	0.0422	0.0015	0.0043	0.0015	0.0018
at 0.90	0.0335	0.0015	0.0043	0.0015	0.0018
at 1.00	0.0294	0.0015	0.0043	0.0015	0.0018
Rel_ret	589	231	81	227	115
Avg:	0.1094	0.0115	0.0115	0.0082	0.0060

Table 2. Retrieval Effectiveness after modify Thai Topics (Additional runs, not submitted to CLEF)

As mention in section 3, the number of term in LEXiTRON dictionary is more cover than SEASITE. However, the effectiveness of both dictionaries is almost the same and over 90% of words are found. Figure 2 shows the characteristic of Thai queries. When applying the single mapping techniques, it turns out that SEASITE can retrieval a little bit better than LEXiTRON, which is opposite to our experiment in [18]. However, this number is not significant achievement.

4.2 The Effect of Segmentation Algorithm in CLIR

As mentioned in section 3 for automatically query translation, the effectiveness of the current technology for segmenting phrase or sentences into words needed to verify. In our experiment, it is not clear that the segmentation has effected in the CLIR. Though it has been reported in [18] that the segmentation has effected in the CLIR, we are unable to prove in the experiment. In our discussion, the different is of the Topic translation techniques from English to Thai. Our first experiment, the researcher is lean on dictionary to translate the English to Thai and try to break words according to the mapping of word in dictionary. Additionally, some unofficial report stated that terms in the Thai-bilingual electronic dictionary are the smallest term with meaning. Thus, some manual segmented words cannot found in dictionary. However, the percentage of number of word found is quite high. Therefore, we compared the original query and the translation query, we found that only 15 percent of matching words. It means that the words, which are found in the dictionary, are not relevance to the search terms.

Table 2 shows significant different between manual and automatic segmentation after modified Thai topics (see discussion in next section).

4.3 The Effectiveness of CLIR

Table 1 shows the results of our techniques. In all techniques cannot achieve the retrieval effectiveness. The query expansion by adding Thai thesaurus terms is not only increase the recall/precision but the retrieval effectiveness also drop.

Comparing with our previous results [18], in which the retrieval effectiveness is around 40% of the monolingual, there is many different in the design process and can be summarized as fellows:

1. The manual translation techniques from English to Thai: As mention in section 4.2, our previous translation technique is based on the vocabulary in the dictionary terms.
2. Query length: Our previous query topics are translated from the <TITLE> tag. The Thai keywords from the <TITLE> tag are more relevance to the retrieval system since the translation process was biased. In this experiment, the query topics are translated from <DESC> tag and avoiding consult the dictionary. Although, there are more terms, most of the terms are not specific to query or more general. As we learned from expanding the query topics with Thai thesaurus, it does not increase the retrieval effectiveness.

3. Thai stopword and Thai stemming: Relatively few intensive studies in Thai stopword and stemming have been reported. Some Thai stopwords are reported in [2]. It is not clear whether Thai language has stemming property. The words 'การ:kan:when prefix to a noun, it indicate action' and 'ความ:kwam:prefix to an adjunctive indicate stats, condition' are two Thai prefix. Removing or not removing has effected in Retrieval effectiveness and is required language knowledge judgment. Since removing the Thai prefix, the meaning of the stem word may not relate to the original meaning. Therefore adding these stem words will degrade the retrieval effectiveness.

Thus, to prove the above issues, the Thai topics are modified by human judgment, some Thai stopwords are removed from the topics, and choosing the search terms which can be found in the dictionary. At this time the retrieval effectiveness is improved 1.5 of the unremoved Thai stopword query (see Table 2). The number of relevance retrieval increase to 40% of monolingual Though, the retrieval effectiveness still cannot achieve as of other CLIR, which deployed dictionary mapping techniques.

Results in Table 2 brought back our confident. It showed that the number of terms in LEXiTRON is more coverage than SEASITE. Segmentation process still is the critical issue in CLIR. However, adding the Thai thesaurus terms still cannot improve the retrieval effectiveness. There has some changed in average precision for individual query.

4.4 Implementation of Thai CLIR

The algorithm of Thai CLIR has implemented and opened for publicly try and the web site is http://www.cs.sci.ku.ac.th/~ThaiIr/CLIR/demo. The demonstrated web site receives Thai keywords from users and then translate using single dictionary mapping. The result of translation is sent back and allow user for selecting the English keywords. Then, query is sent to Google or Altravista for searching English web pages. Furthermore, the results of cross language retrieval may be translated from English to Thai by Pasit. This part of the demonstration program was supported by the NECTEC.

5. Discussion

The Thai CLIR faced the same problems as other MRD CLIR based. The fundamental problems of the MRD CLIR based are as follows: phrase translation, polysemy translation, and the coverage of dictionary. The phrase translation is very critical for Thai CLIR. Some of Thai words may be classified into sentence or phrase. Therefore, the phrase translation will be dependent on the segmentation process. The researching of segmented algorithm in Thailand can be classified into two types. The first preferable segmentation is based on the longest matching. The second research group will segment text into the smallest word. Theoretically, these smallest words can be formed a new word. However, electronic dictionary is collected based on the first approach.

We also have learned that doing research in the area of CLIR only knowledge from the information retrieval but also requires knowledge and resource from Machine Learning. Although the Machine Learning project has been activated more than 4 decades, the resources from the MRD still very limited. The limiting of resources is regarded from the uncompatible or not ready to disseminate to public use. Thus, there exists a need to accelerate the research area. Especially, it needs to set up data format for the electronic dictionary for reusing the dictionary.

We have learned from our demonstration the Thai-CLIR, users quite satisfy the Thai-CLIR system. Unfortunately, in research experimental, not all query translation techniques can achieve. This raises awareness in the Thai-CLIR area. The basic infrastructure of Thai-CLIR needs to be stimulated and urgently needed to further develop.

6. Future Work

In this initiative CLIR research, the fundamental of CLIR research has been established. A number of research techniques to enhance CLIR performance is of solving disambiguate terms, detecting the transliterated word, local feedback.

References

[1] ____, Thai Wordbreak Insertion Services, National Electronics and Computer Technology Center, URL:http://ntl.nectec.or.th/services/wordbreak/ (download in June, 2001)

[2] Jaruskulchai C., An Automatic Indexing for Thai Text Retrieval, Ph.D. Thesis, George Washington University, U.S.A., Aug 1998.

[3] Suwanvisat P. and Prasijutrakul S., Thai-English Cross-Language Transliterated Word Retrieval Soundex Technique, NCSEC2000.

[4] Pirkola Ari, The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-language Information Retrieval, SIGIR'98

[5] Mirna Adriani, Dictionary-based CLIR for the CLEF Multilingual Track.

[6] ____, Online Thai English Dictionary, Northern Illinois University,

www.seasite.niu.edu/Thai/home_page/online_thai_dictionaries.htm (download in June 2001)
[7] ____, The Thai Internet Education Project, http://www.cyberc.com/crcl/ehelp/base.htm (doug@crcl.chula.edu: Contract person, download in June, 2001)
[8] ____, LEXiTRON, Thai<->English Dictionary, Software and Language Engineering Laboratory, National Electronics and Computer Technology Center, http://www.links.nectec.or.th/lexit/lex_t.html (download in June, 2001)
[9] ____, Parsit, Information Research and Development Division, National Electronics and Computer Technology Center, http://www.links.nectec.or.th/services/parsit/index2.html (download in June, 2001)
[10] Sophonpanich Kalaya, The R&D Activities of MT in Thailand, The National Electronics and Computer Technology Center, Bangkok, Thailand.
[11] Suwanvisat Prayut and Prasitjutrakul Somchai, Transliterated Word Encoding and Retrieval Algorithms for Thai-English Cross-Language Retrieval
[12] Kawtrakul A., Deemagarn A., Thumkanon C., Khantonthong N and McFetridge Paul., Backward Transliteration for Thai Document Retrieval, Natural Language Processing and Intelligent Information System Technology, Research Laboratory, Dept. of Computer Engineering, Kasetsart University, Bangkok, Thailand.
[13] Yuen Phuwarawan and team, Thai Thesaurus, in Thai, Ed publisher.
[14] Adriani M., and Croft Bruce, The Effectiveness of a Dictionary-Based Technique for Indonesian-English Cross-Language Text Retrieval, Center for Intelligent Information Retrieval, Computer Science Department, University of Massachusetts, USA.
[15] Kanlayanawat W., and Prasitjutrakul S., Automatic Indexing for Thai Text with Unknown Words using Trie Structure, Department of Computer Engineering, Chulalongkorn University.
[16] SMART, ftp.cs.cornell.edu/pub/smart/smart.11.0.tar.z
[17] Charoenkitkarn, N., and Udomporntawee, R. Optimal Text Signature Length for Word Searching on Thai Holy Bible(in Thai). Proceeding of Electrical Engineering Conference, KMUTT, Bangkok, November 1998, 549-552.
[18] Sripimonwan V. and Jaruskulchai C., Cross-Language Retrieval from Thai to English (in Thai), to be summated to The Fifth National Computer Science and Engineering Conference, Thailand.