VIETNAMESE-ENGLISH CROSS-LANGUAGE INFORMATION RETRIEVAL (CLIR) USING BILINGUAL DICTIONARY

Nguyen Han Doan
Hewlett-Packard Company, USA

ABSTRACT: *Web content is growing each day explosively. A case study performed in 2001 [9] suggested that 70 percent of internet content is in English, but only about 44 percent of Internet users are native English speakers. These numbers are expected to change but English language is still expected to play a dominate role. To gain access to English digital documents from a search query written in Vietnamese language, we propose a Cross Language Information Retrieval (CLIR) technique which takes a query and translate it into phrases, as query text, to retrieve relevance English search result. The technique employs web query logs to arrive at statistical information regarding patterns of words usage. The information then helps to eliminate translation ambiguities by selecting a proper word sense (meaning) for translation. The proposed work also is concerning about the structure of translated query, in posing translated words in a certain order, to obtain relevance search result.*

Keywords: *Cross-Language Information Retrieval, Ambiguity resolution, Word Segmentation, Query translation.*

1.INTRODUCTION

Today, popular web search engines such as Google or Yahoo return search result to users. In general cases, the expectation is that query text and indexed documents are both written in the same language. This is referring as monolingual searching. Due to a lack of content in a localized language such as Vietnamese, a user might expect to query a search engine with a localized text and expect to see a result returns in English. This is a challenge which researchers in the Information Retrieval (IR) community, in recent years, are trying to tackle. According to Mirna Adriani et al.[5], CLIR approaches can be classified into (1) translating the queries into the language of the target documents or (2) translate the documents into the language of queries. Translation of the web documents, by machine or human being, into the same language as a source query is a very expensive task. Commercial machine translation (MT) program such as "Systran" has improved drastically over the years. However, it might not able to produce the expected result as an expert translator. The difficulties are due to word sense disambiguation, complexity in building linguistic rules, and handling of idioms, semantic, and pragmatic knowledge. Translation of queries into a language of target documents can be included the following approaches (1) applying machine translation to queries, (2) using parallel corpora, or (3) using bilingual dictionaries. Bilingual dictionary based approach is attractive due to its simplicity and increasing availability of online machine readable dictionary, but also it has proved to be the most robust for the web as short queries that are typically entered by users [3]. An estimate average length of web queries was 2.30 words [10] and 3.18 characters for Chinese web searches [3]. The author study of a sample of 1045 Vietnamese search queries, generated from Google add word tool, with result focused tailored specifically to Vietnamese language, found to have an average of 3.9 tokens. Dictionary-based method also has its own challenges. One can not take a translation of word by word to compose a query text. This will result in incompatible meaning therefore yielding an undesirable search result. Another challenge with dictionary based method is the

elimination of ambiguity by choosing the right word for translation. Since Vietnamese lexical category is broader than English, a Vietnamese word can map to various English words. Hai [7] referred to this condition as categorical mismatch. This is a well-known issue discussed in [3, 5]. Finally, constructing of translated words together to arrive at a meaningful of query text does require a support of a word interdependency concept on how to best formulate a text – it is better to phrase a query as AB as suppose to BA even though grammatically both are correct. A query AB would get a more relevance result since the same text is also appeared in documents more frequently. The web users' experience, derived from search query log, also indicates the same activities. That is, statistically, a majority of users is likely to query with AB as suppose to BA. Our approach exploits the data from such knowledge, query log, to assist in composing translated queries.

In section 2, we describe a brief survey of relevant work done by researchers in the areas of CLIR. Serving the background of the work, section 3 provides a short introduction to Vietnamese word structure. Section 4 outlines our approach to the problem of Vietnamese-English cross language search. Section 5 illustrates the use of the technique through an example. Section 6 discusses an experiment and initial result. Finally, section 7 concludes the paper with a summary.

2.RELATED WORKS

Early work in CLIR included Salton [13] described a usage of small thesauri; cross-language retrieval was nearly as good as mono-lingual retrieval. The work was focusing on German/English dictionary. Parallel corpora has been proposed by Davis and Dunning [14]. The method is based on the vector space model and involves the linear transformation of the representation of a query in one language to its corresponding representation in another language. Sheridan and Ballerini [15] discussed a technique to translate queries using parallel corpora where term similarity thesaurus built by comparing the occurrences of term in match up documents written Italian and German.

Query translation using bilingual dictionaries has been studied by David A. Hull and Gregory Grefenstette [16]. The result showed that word-by-word translation result in poor performance due to ambiguity of translation. Lisa Ballesteros and W. Bruce Croft [6] suggested that failure to take advantage of translation multi-term concepts as phrases greatly reduces the effectiveness of dictionary translation. In experiments where query phrases were manually translated [17], performance improved by up to 25% over automatic word-by-word query translation.

Asian Pacific (AP) languages CLIR studies have also used bilingual dictionary based approach [1]. The work described an English-Chinese cross-language retrieval experiments at Berkeley for TREC-9 CLIR track. Thai-English CLIR experimental work was described by Jaruskulchai [11]. The study showed that simple dictionary mapping technique was 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. This is a phenomenon occurs where a lexical mapping from a source language to a target does not produce a correspond word. Le et al [7], in their work with Vietnamese-English Machine Translation, referred this condition as a lexical gap. Nguyen et al [2] conducted a preliminary experiment of a cross-lingual Vietnamese-English retrieval system to determine feasibility for retrieving documents between the languages. The work used a simple language transformation algorithm by looking up word

in a bilingual dictionary for translation. Phrasal translation consideration was not included in the study. There was no discussion on handling word sense ambiguity and lexical gap.

3.VIETNAMESE WORD CHARACTERISTICS

Vietnamese language is an isolated one [7]. Its word never changes its form. As such its grammar highly relies on word order and sentence structure rather than morphology (in which word changes through inflection). English language uses morphology to express tense. Vietnamese uses grammatical particles or syntactic constructions. For example:

Vietnamese	English
đã tới	Arrived

The particle indicates that the action is past tense.

Vietnamese words are grouped into single, compound and reduplicative [4,7]. Single word has a single syllable. Compound words have two or more syllables. Most of them are two-syllable words. Many single words can come together to form a new compound words to provide a specific meaning or to form a new meaning (e.g. máy/machine, bay/fly, máy bay/air plane). Since a word can be derived from other words, it is difficult to determine a word boundary. Even though the white space is used in Vietnamese language, it can not be used as word segmentation for recognition of a word. Of course, native speakers do not have any issue with the language; they can converse or write words without any difficulty. However, for a language processing by a computer, this is a problem since the white space can not be used to separate words because a word might be included more than just a single syllable (morpheme). Reduplicative word is another type of lexical units. It usually consists of two syllables, in which only one, or even none, has some meanings, the other one is just a variant of the sound of other (e.g. đỏ/red - đo đỏ/ reddish). This type of lexical unit causes difficulties for a computer to recognize word boundary condition.

According to [7], there are two categories for word classification. Lexical word points to real thing or action such as "nhà" (house), "uống" ("drink"). Grammatical word does not describe any real thing or action, but has grammatical role in building sentences, such as "bởi vì" ("because"), "đây" ("here"). In general, lexical word includes noun, pronoun, verb, and adjective. Grammatical word consists of adverb, preposition, conjunct, and auxiliary and interjection. In our proposed work, grammatical words are treated as noise words and will not be translated for searching. To a search engine such as Google, noise words rarely help narrow a search and can slow search results. It ignores common words and characters [18]. Recognizing this condition, from practical perfectives, our work is only focusing on the translation and formulation of key concepts as phrases. Thus, noise words are translated in our current frame work because these words will be filtered out by search engine eventually.

4.VIETNAMESE-ENGLISH CROSS LANGUAGE SEARCH TECHNIQUE

In this section, we describe a technique to perform Vietnamese to English cross language search. The technique is based on a dictionary based approach where segmented Vietnamese word is translated and formulated as phrases. Translation ambiguities are resolved statistically by neighborhood words co-occurrence. The concept is base on a key observation in [6] that the correct translation of query words should co-occur in target language documents and incorrect translation should tend not to co-occur. Given the possible target equivalents of two source

words, we can deduce the most likely translation by looking at the pattern of co-occurrence for each possible pair definition.

Several assumptions are made here about the applicability of the proposed technique. (1) Search engine query log (i.e. derive from a web server log) is made available in order that Vietnamese segmented words can be discovered by checking the occurrence patterns of terms appearing in the corpus. (2) Query log comprises of unigram and bigram words can occur independently in the query log. (3) A Vietnamese-English bilingual dictionary is available online (refer to http://vdict.com) so that word lookup can be performed quickly.

The proposed Vietnamese-English cross-language technique is depicted in Fig. 1 below. The Query Pre-processing processes on Vietnamese search query. It removes the noise words i.e. "bởi vì", "trên", "những". It recognizes proper names of people via a known listing. The word segmentation module slices the query text into individual Vietnamese words using statistical data derived from occurrence patterns of terms appeared in the corpus. The Dictionary-based Word Translation looks up translated English word(s) from a given Vietnamese word. The Phrasal Formulation & Ambiguity Resolution formulates translated words into individual phrases. Whenever an ambiguity situation is arising, a resolution technique is employed to resolve it. This is done by looking at the previous pattern of previous co-occurrence pair words. Finally, the translated query is formulated and submitted to an English Search Engine to find desired documents.

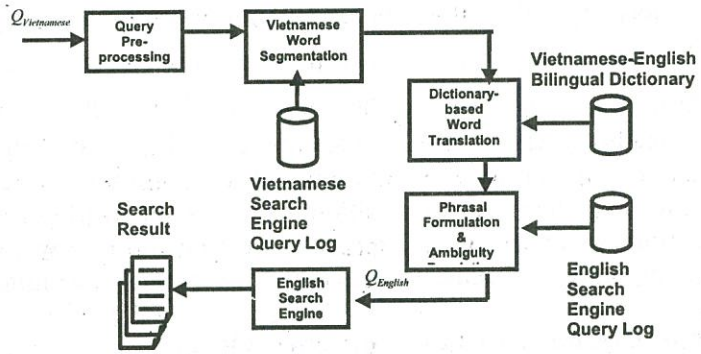


Figure 1. Processing of Vietnamese English Cross Language Search Retrieval.

4.1.Vietnamese Word Segmentation

Prior any word can be looked up in a dictionary for an English translated word, a word must be identified. For languages such as English in which words are separated by blank space, it is simple to identify such words. However, with Vietnamese text, even though blank space is available, it can not be used deterministically to decide word boundary because a word can be constructed from multiple terms. There are quite a few of researches in this area. In this paper, we will not attempt to examine the area. Interested audiences could refer to [4,12] for more information. Both corpus-bases statistical methods and dictionary-based methods have been developed to break sentence into words. If one has a Vietnamese dictionary, one could use a dictionary to match the text against the dictionary to determine the validity of a word. Corpus-base approach utilizes large amount of Vietnamese text and attempt to discover words by checking the occurrence patterns of terms appearing in the corpus.

We approach the segmentation problem by exploiting user search query log. Our method is similar to [1] using for Chinese word segmentation. It breaks a sentence into unigrams and

bigrams by maximizing the probability of the sentence with an assumption that unigrams and bigrams occur independently in the query log. For a segmented sentence S = $w_1\ w_2 ... w_m$, if we assume words occur independently, then the probability of the sentence S can be described as follows:

$$P(S) = P(w_1\ w_2 ... w_m) \quad (1)$$

$$P(S) = P(w_1)\,P(w_2) ... P(w_m) = \prod_{i=1}^{m} P(w_i) \quad (2)$$

Given term t $\in \Sigma^*$ and query q= $t_1\ t_2 \ldots t_m$, find a sequence of words $w_1\ w_2\ w_k \ldots w_n$ such that, for each i $\in$ 1..m-1, $w_k = t_i\ t_{i+1}$ if $P(t_i\ t_{i+1}) > P(t_i)P(t_{i+1})$. Otherwise, $w_k\ w_{k+1} = t_i\ t_{i+1}$. In the above, $P(t_i)$ describes the probability of one single term (unigram word) appears in the corpus. It is estimated by $P(t_i) = N(t_i)/N$. $P(t_i\ t_{i+1})$ depicts the probability of two terms (bigram word) estimated by $P(t_i\ t_{i+1}) = N(t_i\ t_{i+1})/N$, where $N(t_i)$ is the number of times the term, t_i, occurs in the query log, $N(t_i\ t_{i+1})$ is the number of time the bigram $t_i\ t_{i+1}$ occurs in the query log and N is the total number of times that any single term and any two-terms occurs in the query log. There is a situation where both $t_{i-1}\ t_i$ and $t_i\ t_{i+1}$ suggest two separate words $w_k\ w_{k+1}$ respectively. Under such condition, one of a word must be chosen. A higher probability will be used to determine the best candidate leaving one of remained term as a single unigram word. The main idea of algorithm 1 below is to scan a query text from left to right. It compares probabilities of biagram with unigram words. The outcome determines if the two terms should appear as a single word or two words separately. We propose a Greedy algorithm to segment a Vietnamese search query as follows:

Algorithm 1: Word Segmentation for Vietnamese Query:

Input: $t_i \ldots t_m$ - A sequence of individual terms appears in query text, q

Output: S - Buffer storage for an output sentence with segmented word(s)

```
S = {}
For I = 1 to M - 1:
    If P(t_i) * P(t_{i+1}) > P(t_i t_{i+1}):
        If t_{i-1} t_i or t_i not In S:
            S = S + t_i + t_{i+1} // 2 new words added
        Else:
            S = S + t_{i+1}  // 1 single word added
    Else:
```

```
If t_{i-1} t_i In S: // conflict condition
        If P(t_i t_{i+1}) > P(t_{i-1} t_i):
                S = S - t_{i-1} t_i  // remove previous bigram
                S = S + t_{i-1}  // make single word
                S = S + t_i t_{i+1}  // add new bigram
        Else:  // treat as one new word
                S = S + t_{i+1}
Elseif t_i in S:  // new bigram word from t_i
        S = S - t_i
        S = S + t_i t_{i+1}
Else: // new bigram with t_i not appear previously
        S = S + t_i t_{i+1}
```

4.2.Dictionary-Based Word Translation

In this step, we proceed to lookup segmented words against a Vietnamese-English bilingual dictionary. If there is a match, we obtain all suggestions provided. If a word is not found in a dictionary and a word contains 2 terms, we will lookup individual term for possible translation. If translation is not available, we mark the word as not translatable and retain it as part of the query text use for Phrasal Formulation and Ambiguity Resolution. Non-translated words also contain proper names or terms that are numbers.

4.3.Query Translation Model

We will utilize of the work of Marcello and Ying [8,3] by using Hidden Markov Model (HMM) as a disambiguation technique. Given a Vietnamese query ($v_1, v_2, ...v_n$), each English translation candidate set E is a sequence of words ($e_1, e_2, ...e_n$). A probability model P(E) = P($e_1, e_2, ...e_n$) is estimated the maximum likelihood (ML) of each sequence of words. The English translation E with the highest P(E) among all possible translation set is selected. With a bigram HMM with window size 1, P($e_1, e_2, ...e_n$) is estimated as follows:

$$\mathbf{P}(e_1, e_2, ...e_n) = \mathbf{P}(e_1) \prod_{a=2}^{n} \mathbf{P}(e_a \mid e_{a-1}) \qquad (3)$$

where P(e) is the probability of term e occurring within the corpus, and P(e,e') is the probability of term e' occurring after term e within the corpus. P(e) can be computed as follows:

$$\mathbf{P(e) = f(e) / N} \qquad (4)$$

where f(e) is the frequency of term e and N is the number of terms occurring in the corpus. There are many events x such that f(x) = 0. This is an undesired condition because it gives probability of 0 to unseen words. Thus smoothing technique is required to provide a better way to estimate and assign the probability to that unseen event. We compute P(e|e') by using the absolute discounting smoothing method described in [8]:

$$\mathbf{P(e|e') = max\ \{} \frac{f(e,e')-\beta}{N}\mathbf{, 0\} + \beta\ P(e)\ P(e')} \qquad (5)$$

where f(e,e') is the number of co-occurrences appearing in the corpus. Further simplifying of the equation by choosing β with a value of 1, we have the following:

$$\mathbf{P(e|e') = max\ \{} \frac{f(e,e')-1}{N}\mathbf{\ , 0\} + P(e)\ P(e')} \qquad (6)$$

This equation computes the likelihood of e co-occurring with e'.

4.4.Query Translation Algorithm for Phrasal Formulation & Ambiguity Resolution

The query translation model presents a concept for translation of source query into the target language. The model evaluates at all possible translations to arrive at a best solution. This is a not practical solution as search throughput performance is an important requirement for web searching. We need to keep the choice set manageable in order to obtain a desire throughput. We propose to use a heuristic search strategy called Beam search. This strategy is not complete nor is it optimal. However, it is efficient in both time and space. Beam search is like Breadth-first search in that it progresses level by level. Unlike Breadth-first search, however, Beam search moves downward only through the best w nodes at each level; the other irrelevant nodes are ignored. We use w=2 in our current work. At the end of the search, a probability model P(E) = $P(e_1,e_2,...e_n)$ is used to estimate the maximum likelihood (ML) of each sequence of words for the two best translations. The English translation E with the highest P(E) among all possible translation set is selected. The main idea of algorithm 2 is to step through the chains of translated English words to selectively choose at most 2 possible translations at each level of evaluation. The probability of the better of 2 final translation paths determines which translation to be used the next phase. To eliminate ambiguity condition, the system will utilize word bigram frequency to select the best candidate. For example, to choose "roast" or "grill" for bread, we choose "grill bread" because P(bread | grill) is greater than P(bread | roast). This knowledge is derived from our log data. The heuristic beam search strategy is described as follows:

Algorithm 2: Beam Search for Query Translation:

Input: E - A sequence of translated English terms E = $[e_1,e_2,...e_n]$

Output: Path - Buffer storage for 2 possible translation paths. Each path contains a sequence of translated English words

```
Create a queue Path
Set Path = {}
Path = {Start node: e1 }
Until a first path of Path ends with a last translated term
OR Path is empty
```

```
    . Extract the first term, e, from Path
   Extend ALL PATHS one step to all neighborhood terms, e',
creating N new paths
       Sort all paths by distance freq(e,e') and freq(e',e)
   Discard all but W = 2 paths where freq(e,e') or freq(e',e)
has the highest value
       Push each remaining new path to the back of Path
    If Path != {}:
       Extract path1 and path2 from Path
       Compute probability model P(path1) and P(path2) via eq
(6)
       Choose a path with the highest P value
    endIf
```

There are 3 heuristic rules used in conjunction with the algorithm 2 above. They are:

(1) When comparing path (e,e') and (e,e''), if f(e,e') = f(e,e''), we compute p(e' | e) and p(e'' | e). The highest value is used to include in **Path**.

(2) Whenever insert a term e' in front of e, if there is a pre existing term e'', in front e, ordering the terms e' and e'' by their matching frequency with e, from low to high. The idea is to avoid disrupting an ordering of a more important existing pair of words.

(3) When f(e,e') = f(e',e) = 0, due to unseen words, expand to a larger window size of 2, when possible. Include (e,e') to **Path** if there is an existing term, e'', appears to the right hand side of e where f(e',e'') > 0. Otherwise, include (e'e) to **Path.** This idea is to promote a generation of a phrasal text whenever possible.

4.5.Query Syntax Formulation

To help improve recall performance, when a translated query text contains terms that are ambiguous, due to unavailability of a word translation or by a partial translation of a bigram word, again due to an absence of a word translation, a search query will be expanded with these words in addition to translated contextual related words. The search syntax, for a translation text, will have the following syntax:

Query Text = (Translated contextual words) [(Partial translated words) OR (Untranslatable words)] (7)

Google interprets any whitespace user types as an automatic Boolean AND. To specify a Boolean OR condition, the word OR must be specified in capital.

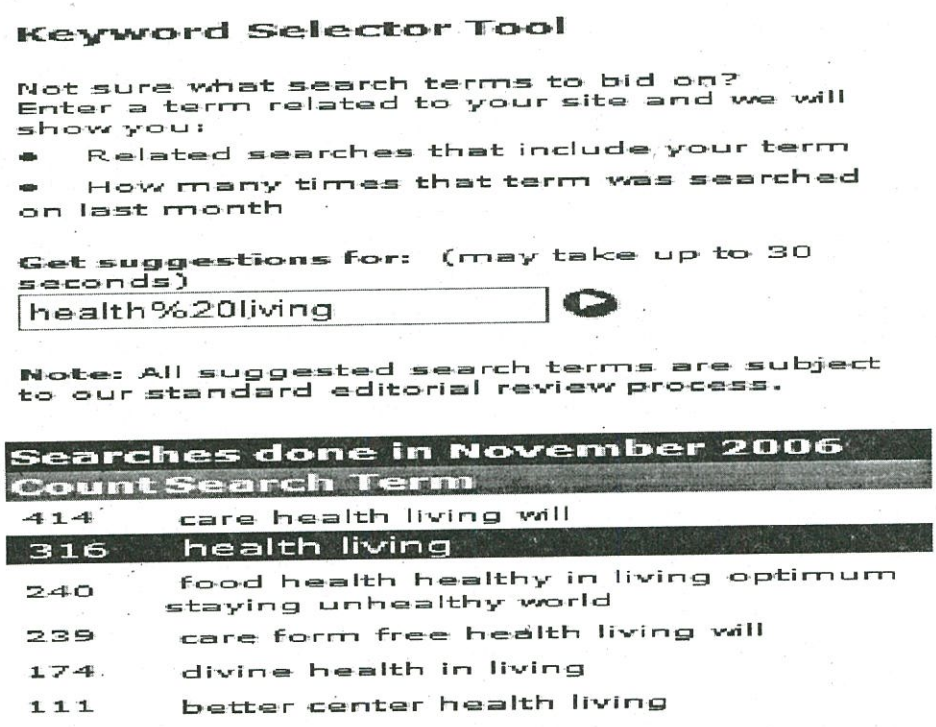
5.EXPERIMENTATION AND INITIAL RESULT

The purpose of this experiment is to understand the effectiveness of the proposed technique. We are also identified potential issues facing ahead. We tested the algorithms on a small set of text obtained randomly from various subjects in Vnexpress (http://vnexpress.net). A sample of test queries is shown in Table 1.

Table 1: Sample of test queries and their translations for Vietnamese-English CLIR method

ID	Query Text	Translated Queries	Comments
Q1	Giá xe ôtô tại Việt Nam	price xe ôtô Vietnam OR xe ôtô	Vdict's Dictionary does not have a translation for xe ôtô (Automobile)
Q2	Thuế thu nhập cá nhân	personal income tax	Good relevancy
Q3	Trung tâm bán cây kiểng	center sell plant kiểng OR "cây kiểng" OR plant	No translation for "cây kiểng" (bonsai). With query expansion syntax, the system is able to get some acceptable relevancy.
Q4	Phát triển phần mềm StarSoft	develop part soft starsoft OR starsoft OR phần mềm OR part OR soft	Vdict's Dictionary does not have a translation for "phần mềm" (Software)
Q5	Máy nướng bánh mì	machine grill bread	Good relevancy
Q6	Gửi tiết kiệm ngân hang	send economize bank	Vdict's Dictionary does not have a good translation for "tiết kiệm" (saving)
Q7	Máy nghe nhạc MP3/MP4	machine mp3/mp4 hear music	Good relevancy
Q8	Những công nghệ máy tính của tương lai	calculator future industry	Vdict's Dictionary does not have a good translation for "máy tính" (Computer). It suggests calculator instead.
Q9	Sức khoẻ đời sống	health living	Good relevancy - eliminates ambiguities
Q10	Lịch sơn phủ cho năm mới	year calendar sơn phủ new OR sơn phủ OR paint OR endow	Vdict misrecognizes "Lịch sơn phủ"

In Figure 2, to resolve words ambiguities (Algorithm 2), the statistics of English words usage, is obtained from: http://inventory.overture.com/d/searchinventory/suggestion



Keyword Selector Tool

Not sure what search terms to bid on? Enter a term related to your site and we will show you:

- Related searches that include your term
- How many times that term was searched on last month

Get suggestions for: (may take up to 30 seconds)

health%20living

Note: All suggested search terms are subject to our standard editorial review process.

Searches done in November 2006

Count	Search Term
414	care health living will
316	health living
240	food health healthy in living optimum staying unhealthy world
239	care form free health living will
174	divine health in living
111	better center health living

Figure 2. Frequency of bigram "health living" on November 2006 logged by Overture Internet Search Tool

This experiment showed:

- The "Phrasal Formulation & Ambiguity Resolution" is able to recognize correct words sense. For example, for a Vietnamese query "Máy nướng bánh mì", the system chooses "grill bread" instead of "roast bread". See Q7.

- The system is able to produce excellent result if the bilingual dictionary provides sufficient translation. See Q2, Q7, and Q11. The system exploits past user behavior, in query log, in choosing "health living" instead of "living health". Thus, it is able to improve search relevancy. See Q11.

- The bilingual dictionary (http://vdict.com/) is not able to produce translation for common words such as "phần mềm" or "tiết kiệm". See Q4 and Q8. In order to improve the search relevancy, there is a need to further improving the online bilingual dictionary to provide better translation.1

- For a condition where there is an inaccessibility of a translation word, with a query expansion method, the system might able to still produce acceptable result, improving recall with a loss on precision, providing there is a sufficient translation of partial term is found. Without partial translation and query expansion, the system will produce very few results. See Q4.

[1] Per site copy right, Vdict.com's dictionary data are collected, recompiled and reformatted from various sources, including Jdict by Ho Ngoc Duc, FOLDOC by Denis Howe from Imperial College London and Wordnet by Princeton University.

- Miss-recognize of proper noun, i.e. Lịch Sơn Phủ, will cause degradation of system performance. See Q6. Having proper nouns define up front will help to improve search relevancy.

- The time complexity for Algorithm 1 is linearly basing on the number of tokens (or syllables) processed. On the average, for web queries, data extracted from "Google Ad Tool" showed that the numbers of tokens included are roughly 4 tokens (see Section 1). The frequency lookup function for unigram and bigram words, from log files, is a constant time due to the usage of memory "Hash Map" function. It is available in most programming languages such as Java or Python. Algorithm 2, where Beam search is used for translation paths resolution, in the worst case, is O(Bm), where B is the beam width, and m is the maximum depth of any path in the search tree. In our case, we are only evaluating 2 possible translations at each level of the tree. Thus, the time complexity is estimated as O(m).

6.CONCLUSION AND NEXT STEPS

We have proposed a technique for Vietnamese-English Cross-language Information Retrieval (CLIR) using bilingual dictionary. Basic ideas and methodologies have been discussed in previous sections along with an example. We have also performed a small experiment where the proposed technique showed good promises. The work is still an ongoing effort and currently being implemented. To conclude the paper, we would like to point out where further research efforts are required:

Implementation concerns. The Beam search is where the most computational effort is spent in the technique. We will look into a possibility in incorporating English grammatically rules for "noun phrase" recognition in order to trim out unnecessary paths.

Performance analysis. We would like to experiment with more search queries. We will use Recall and Precision as basic measurements to measure the effectiveness of the approach.

System tuning. Depending on the outcome of the result performance, we might consider adjusting the window size for query translation algorithm. Marcello's smoothen parameter, β, might be adjusted.

KỸ THUẬT TÌM THÔNG TIN NGÔN NGỮ CHÉO NHAU SỬ DỤNG TỪ ĐIỂN HAI NGÔN NGỮ TIẾNG VIỆT VÀ TIẾNG ANH

Nguyễn Hân Đoàn
Công ty Hewlett-Packard, USA

TÓM TẮT: *Nội dung của web đã tiếp tục phát triển mỗi ngày. Một nghiên cứu ở năm 2001 cho rằng 70 phần trăm nội dung web được viết bằng tiếng Anh, nhưng chỉ có 44 phần trăm người dùng có thể nói được tiếng Anh thông suốt. Những con số này có thể thay đổi nhưng tiếng Anh vẫn được coi là ngôn ngữ chính sử dụng trên web. Để tìm kiếm những nội dung viết bằng tiếng Anh, chúng tôi đề nghị một kỹ thuật "Cross Language Information Retrieval (CLIR)". Kỹ thuật này thông dịch những từ tiếng Việt sang tiếng Anh với dạng nhóm từ (phrases) để kiếm được những nội dung quan trọng. Kỹ thuật này dùng "web log" để kiếm được những thống kê về sự sử dụng của tù tiếng Anh. Thông tin này dùng để loại bỏ những từ không cần thiết hoặc không rõ ràng (ambiguous) trong quá trình phiên dịch bằng cách lựa*

chọn nghĩa đúng trong lúc dịch. Các đề xuất của bài báo cũng liên quan đến cấu trúc của các câu truy vấn đã được dịch, trong việc sắp xếp thứ tự các từ đã được dịch, để đưa ra được kết quả thích hợp.

REFERENCES

[1]. Aitao Chen, Hailing Jiang, and Fredric Gey. *English-Chinese Cross-Language IR using Bilingual Dictionaries*. Ninth Text REtrieval Conference (TREC-9). Gaithersburg, MD. pp. 517-522, (2000).

[2]. Van Be Hai Nguyen, Ross Wilkinson, Justin Zobel. *Cross-language Retrieval in English and Vietnamese.* In AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association of Artificial Intelligence, March (1997).

[3]. Ying Zhang, Phil Vines. *Using the Web for Translation Disambiguation,* Proceedings of NTCIR-5 Workshop Meeting, December 6-9, (2005).

[4]. Le An Ha. *A method for word segmentation in Vietnamese*, Proceedings of the Corpus Linguistics 2003 conference. University centre for computer corpus research on language. Lancaster University (UK), pp. 28-31, (2003).

[5]. Mirna Adreani, C.J. Van Rijsbergen. *Term similarity-based Query Expansion for Cross-Langauge Information Retrieval*, Proceedings of Research and Advanced Technology for Digital Libraries, Third European Conference, ECDL'99, pp. 311-322, (1999).

[6]. Lisa ballesteros, W. Bruse Croft. *Resolving Ambiguity for Cross-language Retrieval.* Proceedings of ACM SIGIR, pp. 64-71, (1998).

[7]. Le Manh Hai, Asanee Kawtrakul, Yuen Poovorawan. *Phrasal transfer model for Vietnamese-English Machine Translation*, Workshop on Multilingual Information Processing (NLPRS'97), (1997).

[8]. Marcello Federico, Nicola Bertoldi. *Statistical Cross-Language Information Retrieval using N-Best Query Translations.* In Proceedings of the 25th ACM Conference on Research and Development in Information Retrieval (SIGIR'02), pp 167-174, (2002).

[9]. Ted Smalley Bowen, *Technology Research News.* http://www.trnmag.com/Stories/2001/112101/English_could_snowball_on_Net_112101.html

[10]. Tessa Lau and Eric Horvitz, *Patterns of Search: Analyzing and Modeling Web Query Refinement,* In Proceedings of the Seventh International Conference on User Modeling. ACM Press, (1998).

[11]. Chuleerat Jaruskulchai: *Dictionary-Based Thai CLIR: An Experimental Survey of Thai CLIR,* Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, (2001).

[12]. Dinh Dien, Hoang Kiem, Nguyen Van Toan, *Vietnamese Word Segmentation.* Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, November 27-30, (2001).

[13]. Salton, G., *Automatic Processing of Foreign language Documents*, Journal of American Society for Information Science, 21: pp. 187-194, (1970).

[14]. Davis, M, and Dunning, T, E., *A TREC Evaluation of Query-Translation Methods for Multi-Lingua Trex Retrieval,* In NIST Special Publication: The 4th Text Retrieval Conference (TREC-6), (1997).

[15]. Sheridan, P., and Ballerini, J. P., *Experiments in Multilingual Information Retrieval using the SPIDER System*, In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, August (1996).

[16]. Hull, D. A, Grefenstette, G., *Querying Across Languages: A dictionary-based approach to Multilingual Information Retrieval*, In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (1996).

[17]. Lisa Ballesteros and W. Bruce Croft, *Dictionary-based methods for cross-lingual information retrieval,* In proceedings of the 7th International DEXA Conference on Database and Expert Systems application, (1996).

[18]. *Why won't Google let me search for numbers or words like "how" and "the"?* http://www.google.com/support/bin/answer.py?answer=981