# Automatic Construction of Cross-lingual Networks of Concepts from the Hong Kong SAR Police Department

Kar Wing Li and Christopher C. Yang
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
{kwli, yang}@se.cuhk.edu.hk

**Abstract.** The tragic event of September 11 has prompted the rapid growth of attention of national security and criminal analysis. In the national security world, very large volumes of data and information are generated and gathered. Much of this data and information written in different languages and stored in different locations may be seemingly unconnected. Therefore, *cross-lingual semantic interoperability* is a major challenge to generate an overview of this disparate data and information so that it can be analysed, searched. The traditional information retrieval (IR) approaches normally require a document to share some keywords with the query. In reality, the users may use some keywords that are different from what used in the documents. There are then two different term spaces, one for the users, and another for the documents. The problem can be viewed as the creation of a thesaurus. The creation of such relationships would allow the system to match queries with relevant documents, even though they contain different terms. Apart from this, terrorists and criminals may communicate through letters, e-mails and faxes in languages other than English. The translation ambiguity significantly exacerbates the retrieval problem. To facilitate cross-lingual information retrieval, a corpus-based approach uses the term co-occurrence statistics in *parallel* or *comparable corpora* to construct a statistical translation model to cross the language boundary. However, collecting parallel corpora between European language and Oriental language is not an easy task due to the unique linguistics and grammar structures of oriental languages. In this paper, the text-based approach to align English/Chinese Hong Kong Police press release documents from the Web is first presented. This article then reports an algorithmic approach to generate a robust knowledge base based on statistical correlation analysis of the semantics (knowledge) embedded in the bilingual press release corpus. The research output consisted of a thesaurus-like, semantic network knowledge base, which can aid in semantics-based cross-lingual information management and retrieval.

## 1. Introduction

In a string of fatal attacks that include the tragic event of September 11, a car bombing in Bali, and an explosion on a French oil tanker off the coast of Yemen, casualties of terrorism have increasingly become regular in daily news all over the globe. These events have prompted the rapid growth of attention of national security and criminal analysis. However, Osama bin Laden's al Qaeda terrorists are not the only threat. We also need to effectively predict and prevent other criminal activities. These include religious, racist and fascist terrorists, opportunistic crime, organized

crime (narcocriminial, Mafia, Russian mob, Triads, etc.), political espionage and sabotage, anarchists and vandals. An intelligent system is required to retrieve relevant information from the criminal records and suspect communications. The system should continuously collect information from relevant data streams and compare incoming data to the known patterns to detect the important anomalies. For example, historical cases of tax fraud can disclose patterns of taxpayers' behaviors and provide indicators for potential fraud. The customers' credit card data can reveal the patterns of transactions and help to detect credit card theft. It should also allow the user to retrieve what persons, organizations, projects, and topics are relevant to a particular event of interest, e.g. car bombing in Bali. However, information stored in the repositories is often fragmented and unstructured, especially on-line catalogs. Also, the man-made fog of deliberate deception militates against normal pattern learning from databases causes much crucial information and the knowledge underlying to be buried. Therefore this information has become inaccessible.

Developing systems that can retrieve relevant information have long been the goal of many researchers since important domain knowledge or information resides in the databases. Many information retrieval systems have been created in the past for medical diagnosis and business applications. The major difficulties to retrieve relevant information are the lack of explicit semantic clustering of relevant information and the limits of conventional keyword-driven search techniques (either full text or index-based)[2]. The traditional approaches normally require a document to share some keywords with the query. In reality, it is known that the users may use some keywords that are different from what used in the documents. There are then two different term spaces, one for the users, and another for the documents. How to create relationships for the related terms between the two spaces is an important issue. The problem can be viewed as the creation of a thesaurus. The creation of such relationships would allow the system to match queries with relevant documents, even though they contain different terms. Language boundaries is another problem for criminal analysis. In criminal analysis, we need to find out how to frame questions, or create search patterns, that would help an analyst. If the right questions are not posed, the analyst may head down a path with no conclusions. In addition, terrorists and criminals may communicate openly and less openly through letters, e-mails, faxes, bulletin boards, etc. in languages other than English. The translation ambiguity significantly exacerbates the retrieval problem. Use of every possible translation for a single term can greatly expand the set of possible meanings because some of those translations are likely to introduce additional homonomous or polysemous word senses in the second language. Also, the users can have different abilities for different languages, affecting their ability to form queries and refine results.

The human expertise to decompose an information need into the queries may take a man several years to acquire. However, knowledge-based systems aim to capture human expertise or knowledge by means of computational models. Knowledge acquisition was defined by Buchanan [10] as "the transfer and transformation of potential problem-solving expertise from some knowledge source to a program". The approach to knowledge elicitation is referred to as "knowledge mining" or "knowledge discovery in databases" [2]. The "knowledge discovery" approach is believed by many Artificial Intelligence experts and database researchers to be useful for resolving the information overload and knowledge acquisition bottleneck

problems. In this research, our aim is to generate a robust knowledge base based on statistical correlation analysis of the semantics (knowledge) embedded in the documents of English/Chinese daily press release issued by Hong Kong Police Department. The research output consisted of a thesaurus-like, semantic network knowledge base, which can aid in semantics-based cross-lingual information management and retrieval. Before the generation of the thesaurus-like, semantic network knowledge base, we firstly propose the text-based approach to collect the parallel press release documents from the Web.

## 2. Automatic Construction of Parallel Corpus

Cross-lingual semantic interoperability has drawn significant attention in recent criminal analysis as the information of criminal activities written in languages other English has grown exponentially. Since it is impractical to construct bilingual dictionary or sophisticated multilingual thesauri manually for large applications, the corpus-based approach uses the term co-occurrence statistics in *parallel* or *comparable corpora* to construct a statistical translation model for cross-lingual information retrieval. Many corpora are domain-specific. To deal with criminal analysis, we use the English/Chinese daily press release articles issued by Hong Kong SAR Police Department. Bates [1] stressed the importance of building domain-specific lexicons for retrieval purposes since a domain-specific, controlled list of keywords can help identify legitimate search vocabularies and help searchers "dock" on to the retrieval system. For most domain-specific databases, there appears to be some lists of subject descriptors (e.g., the subject indexes at the back of a textbook), people's names (e.g., author indexes), and other domain-specific objects (e.g., organizational names, procedures, location names, etc.). These domain-specific keywords can be used to identify important concepts in documents. In the criminal analysis world, the information can help the analyst to identify the people who belongs to which group or organization, uses what methods to conduct the criminal activities in where. In addition, the online bilingual newswire articles used in this experiment are dynamic. They provide a continuous large amount of information for relieving the lag between the new information and the information incorporated into a reference work. To continuously collect English/Chinese daily Police press release articles from the data stream, we investigate the text-based approach to align English/Chinese parallel documents from the Web. Parallel corpus can be generated using *overt translation* or *covert translation*. The overt translation [20] possesses a directional relationship between the pair of texts in two languages, which means texts in language A (source text) is translated into texts in language B (translated text)[25]. The covert translation [13] is non-directional, e.g. press release from the government, commentaries on a sports event broadcast live in several languages by a broadcasting organization.

There are two major approaches for document aligning, namely length-based and text-based alignment. The length-based makes use of the total number of characters or words in a sentence and the text-based approaches use linguistic information in the sentence alignment [9]. Many parallel text alignment techniques have been developed in the past. These techniques attempt to map various textual units to their translation

and have been proven useful for a wide range of applications and tools, e.g. cross-lingual information retrieval [18], bilingual lexicography, automatic translation verification and the automatic acquisition of knowledge about translation [22]. Translation alignment technique has been used in automatic corpus construction to align two documents [16].

There are three major structures of parallel documents on the World Wide Web, *parent page* structure, *sibling page structure*, and *monolingual sub-tree structure*[24]. Resnik [19] noticed that the parent page of the Web page may contain the links to different versions of the web page. The sibling page structure refers to the cases where the page in one language contains a link directly to the translated pages in the other language. The third structure contains a completely separate monolingual sub-tree for each language, with only the single top-level Web page pointing off to the root page of single-language version of the site. Parallel corpus generated by overt translation usually uses the parent page structure and sibling page structure. However, parallel corpus generated by covert translation uses monolingual sub-tree structure. Each sub-tree is generated independently [24]. The press release issued by the HKSAR Police Department is an example.

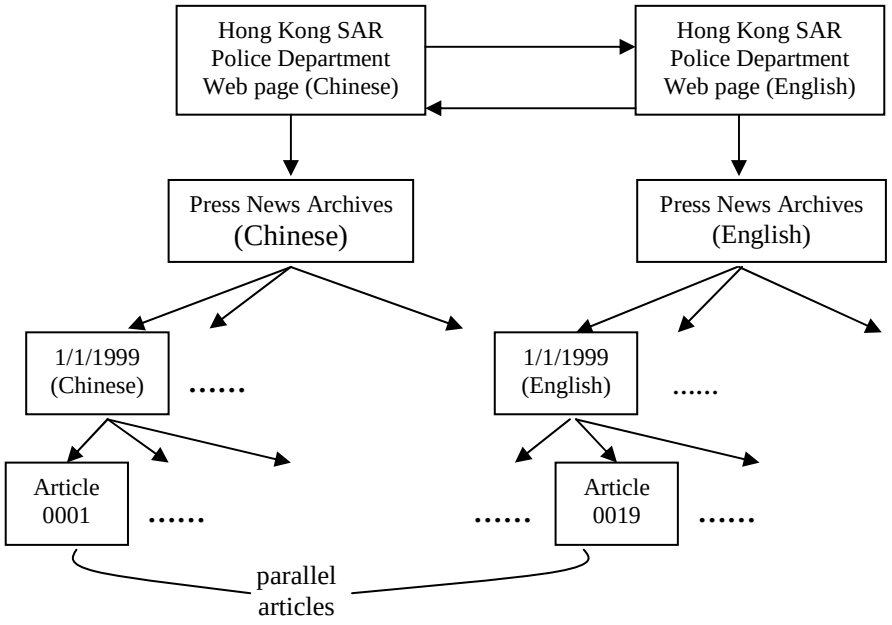


Figure 1. Organization of Hong Kong SAR Police Department's press release articles in the Hong Kong SAR Police Department Web site.

### 2.1 Title Alignment

Titles of two texts can be treated as the representations of two texts. Referring to He[11], the titles present “micro-summaries of texts” that contain “the most important focal information in the whole representation” and as “the most concise statement of

the content of a document". In other words, titles function as the condensed summaries of the information and content of the articles. In our proposed text-based approach, the longest common subsequence is utilized to optimize the alignment of English and Chinese titles [24]. Our alignment algorithm has three major steps: 1) alignment at word level and character level, 2) reducing redundancy, 3) score function.

An English title, *E*, is formed by a sequence of English simple words, i.e., $E = e_1 e_2 e_3 \ldots e_i \ldots$, where $e_i$ is the $i^{th}$ English word in *E*. A Chinese title, *C*, is formed by a sequence of Chinese characters, i.e., $C = char_1 char_2 char_3 \ldots char_q \ldots$, where $char_q$ is a Chinese character in C. An English word in E, $e_i$, can be translated to a set of possible Chinese translations, Translated($e_i$), by dictionary lookup. $Translated(e_i) = \{ T_{e_i}^1, T_{e_i}^2, T_{e_i}^3, \ldots, T_{e_i}^j, \ldots \}$ where $T_{e_i}^j$ is the $j^{th}$ Chinese translation of $e_i$. Each Chinese translation is formed by a sequence of Chinese characters. The set of the longest-common-subsequence (*LCS*) of a Chinese translation $T_{e_i}^j$ and *C* is $LCS(T_{e_i}^j, C)$. $MatchList(e_i)$ is a set that holds all the unique longest common subsequences of $T_{e_i}^j$ and *C* for all Chinese translations of $e_i$.

Based on the hypothesis that if the characters of the Chinese translation of an English word appears adjacently in a Chinese sentence, such Chinese translation is more reliable than other translations that their characters do not appear adjacently in the Chinese sentence. $Contiguous(e_i)$ is used to determine the most reliable translation based on adjacency. The second criteria of the most reliable Chinese translations, is the length of the translations. $Reliable(e_i)$ is used to identify the longest sequence in $Contiguous(e_i)$.

Due to redundancy, the translations of an English word may be repeated completely or partially in Chinese. To deal with redundancy, $Dele(x,y)$ is an edit operation to remove the $LCS(x,y)$ from *x*. *WaitList* is a list to save all the sequences obtained by removing the overlapping of the elements of $MatchList(e_i)$ and $Reliable(e_i)$. $MatchList(e_i)$ is initialized to $\varnothing$ and $Reliable(e_i)$ is initialized to $\varepsilon$. *Remain* is a sequence that is initialized as *C*, and $Reliable(e_i)$ are removed from *Remain* starting from the $e_1$ until the last English word. *WaitList* will also be updated for each $e_i$. When all $Reliable(e_i)$ are removed from *Remain*, the elements in *WaitList* will also be removed from *Remain* in order to remove the redundancy.

Given *E* and *C*, the ratio of matching is determined by the portion of *C* that matches with the reliable translations of English words in *E*. Given an English title, the Chinese title that has the highest *Matching_Ratio* among all the Chinese titles is considered as the counterpart of the English title. However, it is possible that more than one Chinese title have the highest *Matching_Ratio*. In such case, we shall also consider the ratio of matching determined by the portion of English title that is able to identify a reliable translation in the Chinese title.

### 2.2. Experiment

An experiment is conducted to measure the precision and recall of the aligned parallel Chinese/English documents from the HKSAR Police press releases using the text-based approach as described in Section 2.1. Results are shown on Table 1. The Hong

Kong SAR Police press releases are developed based on covert translation. From 1st January, 2001 to 31st October,2002, there are 2,698 press articles in Chinese and 2,695 press articles in English. There are only 2,664 pairs of Chinese/English parallel articles. Experimental result shows that the proposed text-based title alignment approach can effectively align the Chinese and English titles.

**Table 1**. Experimental results

| | Precision | Recall |
|---|---|---|
| Proposed text-based approach | 1.00 | 1.00 |

## 3. A corpus-based approach: Automatic cross-lingual concept space generation

The semantic network knowledge base approach to automatic thesaurus generation is also referred to as a concept space approach[4] because a meaningful and understandable concept space (a network of terms and weighted associations) could represent the concepts (terms) and their associations for the underlying information space (i.e., documents in the database). In terms of criminal analysis, recent terrorist events have demonstrated that terrorist and other criminal activities are connected, in particular, terrorism, money laundering, drug smuggling, illegal arms trading, and illegal biological and chemical weapons smuggling. In addition, hacker activities may be connected to these other criminal activities. Information in the concept space can be split into concepts and links. Concepts include real people, aliases, groups, organizations, companies (including bank and shells), countries, towns, regions, religious groups, families, attacks (hacker, terrorist), etc. The associated concepts in the concept space can provide links about the persons who generally remain hidden, unknown, and use aliases, who, in turn, belong to various groups and organizations, use banks, vehicles, phones, meet in various locations, conduct both criminal and non-criminal activities, and communicate openly and less openly through bulletin boards, e-mail, phone calls, letters, word-of-mouth, etc. – encrypted or not. It helps the analyst to detect the important anomalies. The cross-lingual concept space clustering model is originally suggested by Lin and Chen [15] and based on the Hopfield network. The cross-lingual concept space includes the concepts themselves, their translations as well as their associated concepts. The automatic Chinese-English concept space generation system consists of four components: 1)English phrase extraction; 2)Chinese phrase extraction; and 3) Hopfield network, and 4) Parallel Chinese/English Police press release corpus. The Chinese and English phrase extraction identifies important conceptual phrases in the corpora. The Hopfield network generates the cross-lingual concept space with the Chinese and English important conceptual phrases as input. A press release parallel corpus was dynamically collected from the Hong Kong Police website in order to get the relationship between Chinese terms and English terms.

### 3.1. Automatic English Phrase Extraction

Automatic phrase extraction is a fundamental and important phrase in concept space clustering. The clustering result will be downgraded significantly if the quality of

term extraction is low. Salton [21] presents a blueprint for automatic indexing, which typically includes stop-wording and term-phrase formation. A stop-word list is used to remove non-semantic bearing words such as the, a, on, in, etc. After removing the stop words, term-phrase formation that formulates phrases by combining only adjacent words is performed[4].

## 3.2. Chinese Phrase Extraction

Unlike English language, there are not any natural delimiters in Chinese language to mark word boundaries. In our previous work, we have developed the boundary detection [23] and the heuristic techniques to segment Chinese sentence based on the mutual information and significant estimation [5]. The accuracy is over 90%.

### 3.3.1 Automatic Phrase Selection

To generate the concept space, the relevance weights between the English and Chinese term phrases are first computed in order to select significant concepts from the collection.

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j} \times w_j\right) \quad (1)$$

Equation 1 shows how the combined weight of term $j$ in document $i$ is calculated. $tf_{ij}$ is the occurrence frequency of term $j$ in document $i$. $N$ is the total number of documents in the collection and $df_j$ is the number of documents containing term $j$. $w_j$ is the length of term $j$. For an English term, the length of it is the number of words in it. For a Chinese term, the length of it is the number of characters in it. The weight is directly proportional to the occurrence frequency of the term because it carries important idea if it appears in the document for many times. On the other hand, it is inversely proportional to the number of documents containing the term because the meaning carried by the term may be too general. For example, "Hong Kong" frequently appears in the collection of documents from HKSAR Police. It becomes a common term in the collection and does not carry specific meaning in any document of the collection. The length of term also plays an important role in the weight. It is known that a longer term carries more specific meaning. For example, name of places and organizations are often in multiple words (for English) or characters (for Chinese). Terms, which significantly represent a document, are selected for clustering. Based on the combined weights of terms that are calculated using Equation 1, a number of terms with the largest combined weights in each document are selected for clustering. The number is based on the average length of documents in the collection. For longer average length, more terms are selected for clustering. Terms with common meaning and not representative are filtered out.

### 3.3.2 Co-occurrence weight

After the calculation of $d_{ij}$, asymmetric co-occurrence function [2] is used to evaluate the relevance weights among concepts. For a pair of relevant term A and B, the

weight of the link from term A to term B and that of the link from term B to term A are different. This function gives a good description of natural thinking of human to terms. For example, "Ford" and "car" are relevant. When a person comes up with "Ford", he can think of "car". However, when a person comes up with "car", he may not think of "Ford". This example shows that two terms the associations between two terms are not symmetric. Therefore, we adopt the co-occurrence weight to calculate the relevance weights.

$$d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_j\right) \quad (2)$$

The co-occurrence weight, $d_{ijk}$ , in Equation 2 is the weight between term *j* and term *k* that are both exist in document *i* . $tf_{ijk}$ is the minimum between occurrence frequency of term *j* and that of term *k* in document *i* . The weight will be zero if either of term *j* or term *k* is not exist in the document. The calculation is similar to the calculation in Equation 1. Therefore, the co-occurrence weight is a measure of combined weight between term *j* and term *k*.

$$Weight(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times WeightingFactor(T_k) \quad (3)$$

$$Weight(T_k, T_j) = \frac{\sum_{i=1}^{n} d_{ikj}}{\sum_{i=1}^{n} d_{ik}} \times WeightingFactor(T_j) \quad (4)$$

Equation 3 shows the relevance weights from term *j* to term *k*. Equation 4 shows the relevance weight from term *k* to term *j*. Relevance weight measures the association between two terms in the collection. The combined weights and co-occurrence weights of terms in all documents are summed up to derive the global association between terms in the collection.

$$Weighting\ \ Factor\ \ (T_j) = \frac{\log \frac{N}{df_j}}{\log N} \quad (5)$$

$$Weighting\ \ Factor\ \ (T_k) = \frac{\log \frac{N}{df_k}}{\log N} \quad (6)$$

Equation 5 shows the weighting factor of term *j*. Equation 6 shows the weighting factor of term *k*. The weighting factor is used to penalize general terms. General terms always affect the result of clustering. A lot of terms associate with the general terms. During clustering, if a general term is activated, other terms associate with that general term will also be activated. Then, the size of that concept space will be large and the precision will unavoidably low. The weighting factor is a value between 0 and 1. It carries an idea of inverse document frequency. The more the documents contain the concept, the smaller the weighting factor.

### 3.3.3 The Hopfield Network Algorithm

Given the relevance weights between the extracted Chinese and English term phrases in the parallel corpus, we will employ the Hopfield network to generate the concept space. The Hopfield network models the associate network and transforms a noisy

pattern into a stable state representation. When a searcher starts with an English term phrase, the Hopfield network spreading activation process will identify other relevant English term phrases and gradually converge towards heavily linked Chinese term phrases through association (or vice versa). Term is represented by node in the network. The algorithm is shown below:

$$u_j(t+1) = f_s[\sum_{i=0}^{n-1} t_{ij} u_i(t)], 0 \le j \le n-1 \quad (7)$$

where $u_j(t+1)$ denotes the value of node $j$ in iteration $t+1$, $n$ is the total number of nodes in the network, $t_{ij}$ denotes the relevance weight from node $i$ to node $j$.

$$f_s(x) = \frac{1}{1 + \exp\left[\frac{-(x - \theta_j)}{\theta_o}\right]} \quad (8)$$

Equation 8 shows the continuous SIGMOID transformation function which normalizes any given value to a value between 0 and 1[4].

$$\sum_{j=0}^{n-1} [u_j(t+1) - u_j(t)]^2 \le \varepsilon \quad (9)$$

where ε was the maximal allowable difference between two iterations. ε measures the total change of values of nodes from iteration $t$ to $t+1$. After several iterations, more nodes are activated and nodes with strong connection to the target node are those with high values. Total change of values of nodes is evaluated at the end of iteration. When the change is smaller than a threshold, ε, the Hopfield network is converged and the iteration process stops. Once the network converged, the final output represented the set of terms relevant to the starting term. In our system the following values were used: $\theta_j = 0.1$, $\theta_o = 0.01$, ε=1.

## 4. Concept space evaluation

10 students of the Department of System Engineering and Engineering Management, The Chinese University of Hong Kong, were invited to examine the performance of concept space. The concept space is a robust and domain-specific Hong Kong Police press release thesaurus which contains 9222 Chinese/English concepts. The thesaurus includes many social, political, legislative terms, abbreviations, names of government departments and agencies. Each concept in the thesaurus may associate with up to 46 concepts. It is generated from 2548 parallel Hong Kong Police press release article pairs. The goal of this experiment is to capture meaningful conceptual association between concepts. The associations forms the basis for the decisions and inferences the user use when searching the criminal information of Hong Kong.

### 4.1 Experimental Design

Among these 10 graduate students, 5 subjects are Hong Kong students and the other 5 subjects came from Mainland China. They all have been living in Hong Kong for more than one year. They use their knowledge and experience on both the Hong Kong SAR Police system and the living environment in Hong Kong to evaluate the concept

space. 50 among 9222 concepts were randomly selected as the test descriptors. Twenty five among these 50 test descriptors are English concepts. The other 25 test descriptors are Chinese concepts. Each test descriptor together with its associated concepts were presented to the 10 subjects. A small portion (about 10% of total number of associated concepts for each test descriptor) of noise terms was added to reduce the bias generated by the subjects to the concept space. The experiment is divided into two phrases: recall phrase and recognition phrase. In the recall phrase, each subject (Hong Kong graduate students and graduate students from Mainland China) was asked to generate as many related terms as possible in response to each test descriptor presented. In the recognition phrase, the subjects needed to determine the associated concept either "irrelevant" or "relevant" to the test descriptor. Terms considered too general were to be ranked as “irrelevant”. This phrase tested the ability of subjects on recognition of relevant terms. If the subjects felt the definition of a concept needed to clarify or they wished to add comments on the concept, they were asked to write them on a piece of paper. After the experiment, we found that the subjects spent more time on recognition phrase than what they spent on recall phrase. This confirms the statement made by Chen et al. [3] that human beings are more likely to recognize than to recall. Apart from the 10 students, the 50 concepts in concept space were also carefully evaluated by two experimenters and no noise term was added in the case. One of them is a graduate student of the Department of System Engineering and Engineering Management. The other is a graduate student of the Department of Translation. They both have been living in Hong Kong for more than 10 years. They also have done research on Chinese to English translation and English to Chinese translation for more than two years. Since there is no tailored bilingual thesaurus for Hong Kong government press release articles, the experimental result provided by these two senior subjects is treated as a benchmark or human verified thesaurus in comparison with the result provided by the 10 subjects. The additional associated concepts provided by the 10 subjects in the recall phrase were examined by the two senior judges before treating them as relevant terms.

### 4.2 Experimental Result

We adopted the concept recall and concept precision for evaluation based on the following equations:

$$\text{Concept Recall} = \frac{\text{Number of Retrieved Relevant Concepts}}{\text{Number of Total Relevant Concepts}} \tag{10}$$

$$\text{Concept Precision} = \frac{\text{Number of Retrieved Relevant Concepts}}{\text{Number of Total Retrieved Concepts}} \tag{11}$$

The ***number of Retrieved Relevant Concepts*** represented the number of concepts in the concept space judged as "Relevant". The ***number of total relevant concepts*** includes the concepts in the concept space judged as "Relevant", the additional relevant concepts provided. The ***number of total retrieved concepts*** represented the number of concepts suggested by the concept space and the human verified thesaurus.

### 4.3 Evaluation provided by 10 graduate subjects

The 10 graduate students provided 12 to 73 new associated concepts during the experiment. The analysis is listed in Table 3. It is interested to note that all the Hong Kong graduate subjects have been living in Hong Kong for at least six years but the graduate subjects from Mainland China have been living in Hong Kong around one year. So, the Hong Kong graduate subjects are more familiar with Hong Kong Police system and they added more new concepts to the concept space. In addition, the Hong Kong graduate students added more English concepts to the concept space than that of the graduate students from Mainland China. This confirms that even though the first language of all these graduate students is Chinese, the working language for the Hong Kong graduate students is English.

**Table 3**. The statistics of new associated concepts added by the 10 graduate students

| | New associated concepts | Range |
|---|---|---|
| On average, each graduate student generated | 44.2 | 12 to 73 |
| For each test descriptor, the average number of associated concepts generated by the students | 0.884 | |
| For each test descriptor, a graduate provided | | 0 to 8 new concepts |
| Total number of new associated concepts provided by the 10 graduate students | 442 | |

**Table 4.** Precision and recall

| | Precision | Recall |
|---|---|---|
| 10 graduate students | 0.835 | 0.795 |
| 2 experimenters | 0.86 | 0.83 |

**Table 5.** The new concepts added by the 10 graduate students

| | Chinese concept added | English concept added |
|---|---|---|
| 10 graduate students | 222 | 220 |

Hong Kong is a bilingual community. Even though the Police concept space contains many technical, political and geographical English vocabularies, the Hong Kong graduate students frequently encounter these terms in their daily life. As a result, the Hong Kong graduate students naturally added more English terms into the concept space. This observation also appears in Welsh and English community [7].

Also, even though Chinese technical terms do exist, they may not common use. Therefore, the Hong Kong graduate may have limited Chinese technical vocabulary even where Chinese is their first language and use English terms when necessary. As a result, the Hong Kong graduate subjects judged more English concepts to be relevant and added more English terms into the concept space. On the other hand, the graduate students from Mainland China have a higher degree of Chinese fluency than that of Hong Kong graduate students. Also, they know more Chinese translations of those English technical vocabularies in Mainland China. These cause them to add more Chinese concepts. We also observe some associated concepts are judged as irrelevant because the associated concepts do not show the clear association with their test descriptor. For example, one of associated concepts for the test descriptor "走私

活動" (smuggling) is "Mr Mark Steeple" (施德博) because the Chief Inspector Anti-smuggling Task Force in Hong Kong is Mr Mark Steeple. Another associated concept is "Mirs Bay" (大鵬灣) because of the recent trend of smuggling by small craft in the Mirs Bay area. However, all the graduate students do not have a prior knowledge of these and judged them as irrelevant. Since the corpus is a dynamic resource, it is not surprise that the students do not have a prior knowledge. For criminal analyst, the information is important for identifying the recent trend of smuggling by small craft in the Mirs Bay area. In addition, one of the associated concepts for "Golden Bauhinia Square" (金紫荊廣場) are "警察" (Police). We know that the flag raising ceremony began promptly at 8 a.m. with the Flag Raising Parade at the twin flagpoles at Golden Bauhinia Square. The flag party, provided by the Hong Kong Police Force comprised a Senior Inspector of Police, four flag raisers. Without knowing this, the subjects only read the concept space and judged that there is no clear association between "Police" and "Golden Bauhinia Square". The phenomenon displays that the clustering process using Hopfield network induces the relevant concepts based on the contents of documents. Apart from this, as we know, a lexical item (word) in a sentence may be a concept in one language[12], where *concept* is a recognizable unit of meaning in any given language [11]. A concept represented by a word in one language may be translated into a word, two words, a phrase, or even a sentence in another language [11]. A concept in one language can be a broader concept encompassing some narrower concepts, and the translation of such a concept may result in an altered concept in another language. In contrast, a narrower concept in one language may be translated as a broader concept in another language. Such relationship is known as *generic-specific* relationship[12]. For example, the word "China" is modified to be a specific word "京" (Beijing), a city of China. Omission, addition, and deviation are also common phenomena. For example, "Closure" corresponds to "停止服務" in some cases. "Closure" is translated to "關閉" by dictionary, but it refers to "停止服務 (stop service)" in some cases (deviation). Therefore, *conceptual alternation* may occur in translation. This also causes the judges to judge some associated concepts to be irrelevant. Nida[11] explains that conceptual alteration is caused by three major reasons: 1) no two languages were completely isomorphic, 2) different languages might have different domain vocabulary; and 3) some languages were more rhetorical than other languages. Courtial and Pomian[6] argued that searches performed in the realms of science and technology frequently involve association of concepts that lie outside the traditional associations represented in thesauri. Associative networks gleaned through textual analysis, they argued, facilitated innovation by making obvious associations that would otherwise be impossible for humans to find on their own. In early research, Lesk[14] found little overlap between term relationships generated through term associations and those presented in existing thesauri. This term relationship is especially important for criminal analysis. The associated concepts in the concept space can provide links about the persons who generally remain hidden, unknown, and use aliases, who, in turn, belong to various groups and organizations, use banks, vehicles, phones, meet in various locations, conduct both criminal and non-criminal activities, and communicate through bulletin boards, e-mail, phone calls, letters, word-of-mouth, etc. – encrypted or not. Ekmekcioglu, Robertson and Willet [8] tested retrieval performances for 110 queries on a database

of 26,280 bibliographic records using four approaches. Their result suggested that the performance may be greatly improved if a searcher can select and use the terms suggested by a co-occurrence thesaurus in addition to the terms he has generated[4].

### 4.4 Translation ability of the concept space

The 46683 associated concepts were also examined. For those test descriptors associating with two relevant associated concepts, 47.64% of these associated concepts are Chinese concepts and 52.36% of these associated concepts are English concepts. Among these 9222 test descriptors, 87.7% of them obtain their translations from the associated concepts. It shows that the concept space generated through Hopfield network can effectively recognize the translations of a concept in a parallel corpus.

## 5. Conclusion

The tragic event of September 11 has prompted the rapid growth of attention of national security and criminal analysis. In the national security world, very large volumes of data and information are generated and gathered. Much of this data and information written in different languages and stored in different locations may be seemingly unconnected. Therefore, *cross-lingual semantic interoperability* is a major challenge to generate an overview of this disparate data and information so that it can be analyzed, shared, searched. To effectively predict and prevent criminal activities, an intelligent system is required to retrieve relevant information from the criminal records and suspect communications. The system should continuously collect information from relevant data streams and compare incoming data to the known patterns to detect the important anomalies. However, *information retrieval (IR)* systems present two main interface challenges: first, how to permit a user to input a query in a natural and intuitive way, and second, how to enable the user to interpret the returned results. A component of the latter encompasses ways to permit a user to comment and provide feedback on results and to iteratively improve and refine results. As we know, the *vocabulary difference problem* has been widely recognized: users tend to use different terms for the same information sought. Also, in terms of criminal analysis, the man-made fog of deliberate deception militates against normal pattern learning from databases cause much crucial information and the knowledge underlying to be buried. As a result, an exact match between the user's terms and those of the indexer is unlikely. An advanced tool is required to understand the user's needs. *Cross-lingual information retrieval* brings an added complexity to the standard *IR* task. Users can have different abilities for different languages, affecting their ability to form queries and interpret results. This highlights the importance of automated assistance to refine a query in cross-lingual information retrieval.

This article has presented a bilingual concept space approach using Hopfield network to relieve the vocabulary problem in national security information sharing, using the Hong Kong Police press release bilingual pairs as an example. The concept space allows the user to interactively refine a search by selecting concepts which have been automatically generated and presented to the user. This allows the user to descend to the level of actual objects in a collection at any time. By observation,some

information may be seemingly unconnected but actually information can help the analyst to identify the important anomalies, such traffic accidents frequently happen at a particular location. Since the press release collection is dynamically generated, the subjects may not have a full prior knowledge. However, experimental result shows the precision and recall for the bilingual concept space are over 78% in all cases. Among these 9222 test descriptors, 87.7% of them obtain their translations from the associated concepts. It shows that the concept space generated through Hopfield network can effectively recognize the translations of a concept in a parallel corpus.

## References


1. Bates, M. J. "Subject access in online catalogs: A design model". *Journal of the American Society for Information Science*, 37,357-376. (1986)
2. Chen, H., Lynch, K. J., "Automatic construction of networks of concepts characterizing document database" *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 5, pp. 885-902, Sept-Oct (1992)
3. Chen, H., Schatz, B., Ng, T., Martinez, J., Kirchhoff, A., Lin, C., "A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 771-782, August (1996)
4. Chen, H., Ng, T., Martinez, J., Schatz, B., "A Concept Space Appraoch to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System" In *Journal of The American Society for Information Science*, 48(1):17--31. (1997)
5. Chien, L. F., "PAT-Tree-BASED Keyword Extraction for Chinese Information Retrieval", In *Proceedings of ACM SIGIR*,pp.50-58, Philadelphia, PA,1997.
6. Courtial, J. P. and Pomian, J. “A system based on associational logic for the interrogation of databases”, In *Journal of Information Science*, 13,91-97,1987
7. Cunliffe, D., Jones, H., Jarvis, M., Egan, K., Huws, R., Munro, S., “Information Architecture for Bilingual Web Sites”. In *Journal of The American Society for Information Science*, 53(10):866--873. 2002
8. Ekmekcioglu, F. C., Robertson, A. M. and Willett, P. “Effectiveness of query expansion in ranked-output document retrieval systems”, In *Journal of Information Science*, 18, 139-147,1992.
9. Fung, P. and McKeown, K. (1997) " A technical word- and term-translation aid using noisy parallel corpora across language groups". In *Machine Translation* 12: 53-87.
10. Hayes-Roth, F., Waterman, D. A. and Lenat, D. (1983) "Building Expert Systems". Reading, MA:Addison-Wesley.
11. He, S. "Translingual Alteration of Conceptual Information in Medical Translation: A Cross-Language Analysis between English and Chinese," *Journal of the American Society for Information Science*, Vol. 51, No. 11,2000, pp.1047-1060.
12. Larson, M. L. Meaning-based translation: A guide to cross-language equivalence. Lanham, MD: University Press of American
13. Leonardi, V., "Equivalence in Translation: Between Myth and Reality," *Translation Journal*, Vol. 4, No.4, 2000.
14. Lesk, M. E. (1969) “Word-word associations in document retrieval systems”, In *American Documentation*, 20(1),27-38,1969.

15. Lin, C. H., Chen, H., "An Automatic Indexing and Neural Network Approach to Concept Retrieval and Classification of Multilingual (Chinese-English) Documents" *IEEE Transactions on Systems, Man and Cybernetics*, vol 26, no.1, pp. 75-88, Feb 1996
16. Ma X. and Liberman M. (1999) "BITS: A Method for Bilingual Text Search over the Web". In *Machine Translation Summit VII*, September 13th, 1999, Kent Ridge Digital Labs, National University of Singapore.
17. Oard, D. W., & Dorr, B. J. (1996). *A Survey of Multilingual Text Retrieval*. UMIACS-TR-96-19 CS-TR-3815.
18. Oard, D. W. (1997). Alternative approaches for cross-language text retrieval. In Hull D, Oard D, (Eds.) ,1997 *AAAI Symposium in Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence.
19. Resnik P. "Mining the Web for Bilingual Text," 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), College Park, Maryland, June, 1999.
20. Rose, M. G. (1981). Translation Types and Conventions. In *Translation Spectrum: Essays in Theory and Practice*, Marilyn Gaddis Rose, Ed., State University of New York Press, pp.31-33.
21. Salton, G. (1989) ***Automatic Text Processing***. Addison-Wesley Publishing Company, Inc., Reading, MA, 1989.
22. Simard, M. (1999) "Text-translation Alignment: Three Languages Are Better Than Two". In *Proceedings of EMNLP/VLC-99*. College Park, MD.
23. Yang, C. C., Luk, J., Yung, S., Yen, J., (2000) "Combination and Boundary Detection Approach for Chinese Indexing, " In *Journal of the American Society for Information Science, Special Topic Issue on Digital Libraries*, vol.51, no.4, March, 2000, pp.340-351.
24. Yang, C. C. and Li, K. W. "Automatic Construction of English/Chinese Parallel Corpora," *Journal of the American Society for Information Science and Technology*, vol.54, no.7, May, 2003.
25. Zanettin, F,. "Bilingual comparable corpora and the training of translators," Laviosa, Sara. (ed.) *META, 43:4, Special Issue. The corpus-based approach: a new paradigm in translation studies*: 616-630, 1998.