Automatic Crosslingual Thesaurus Generated From the Hong Kong SAR Police Department Web Corpus for Crime Analysis

Kar Wing Li and Christopher C. Yang
Department of Systems Engineering and Engineering Management, Room 116, Ho Sin Hang Engineering Building, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, The People's Republic of China.
E-mail: yang@se.cuhk.edu.hk

For the sake of national security, very large volumes of data and information are generated and gathered daily. Much of this data and information is written in different languages, stored in different locations, and may be seemingly unconnected. Crosslingual semantic interoperability is a major challenge to generate an overview of this disparate data and information so that it can be analyzed, shared, searched, and summarized. The recent terrorist attacks and the tragic events of September 11, 2001 have prompted increased attention on national security and criminal analysis. Many Asian countries and cities, such as Japan, Taiwan, and Singapore, have been advised that they may become the next targets of terrorist attacks. Semantic interoperability has been a focus in digital library research. Traditional information retrieval (IR) approaches normally require a document to share some common keywords with the query. Generating the associations for the related terms between the two term spaces of users and documents is an important issue. The problem can be viewed as the creation of a thesaurus. 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. The problem is expanded to crosslingual semantic interoperability. In this paper, we focus on the English/Chinese crosslingual semantic interoperability problem. However, the developed techniques are not limited to English and Chinese languages but can be applied to many other languages. English and Chinese are popular languages in the Asian region. Much information about national security or crime is communicated in these languages. An efficient automatically generated thesaurus between these languages is important to crosslingual information retrieval between English and Chinese languages. To facilitate crosslingual 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. In this paper, the text-based approach to align English/Chinese Hong Kong Police press release documents from the Web is first presented. We also introduce 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 crosslingual information management and retrieval.

Introduction

In a string of fatal attacks that include the tragic events of September 11th, a car bombing in Bali, and an explosion on a French oil tanker off the coast of Yemen, casualties of terrorism have increasingly become daily news items all over the globe. These events have prompted the rapid growth of national security and criminal analysis. However, the threat of terrorist attacks are not restricted to the Middle East or North America, but target alerts of terrorist attacks are also frequently raised in places such as Japan, Taiwan, and Singapore.

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. 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.

The major difficulties to the retrieval of relevant information are the lack of explicit semantic clustering of relevant information and the limits of conventional keyword-driven

Accepted February 26, 2004

© 2004 Wiley Periodicals, Inc. • Published online 2 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20118

search techniques (either full-text or index-based; Chen & Lynch, 1992). The traditional approaches normally require a document to share some common keywords with the query. Practically, users may use keywords that are different from the indexed keywords in the documents. There are two different term "spaces," one for the users, and another for the documents. Generating the associations for the related terms between the two spaces is an important issue. The creation of such relationships would allow the system to match queries with relevant documents, even though they contain different terms.

Language boundaries are major problems for criminal analysis, especially for international crimes. 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 homonymous 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.

In this study, 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 releases issued by the Hong Kong Police Department. The research output consisted of a thesaurus-like, semantic network knowledge base, which can aid in semantics-based crosslingual information management and retrieval. Before the generation of the thesaurus-like, semantic network knowledge base, the text-based approach to collect the parallel press release documents from the Web is first presented.

Automatic Construction of Parallel Corpus

Crosslingual semantic interoperability has drawn significant attention in recent criminal analysis as the information of criminal activities written in languages other than English has grown exponentially. Since it is impractical to construct a 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 crosslingual information retrieval.

Many corpora are domain-specific. To deal with criminal analysis, we use the English/Chinese daily press release articles issued by the Hong Kong SAR Police Department. Bates (1986) 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 belonging to a certain group or organization, where they conduct their criminal activities, and what methods they use. In addition, the online bilingual newswire articles used in this experiment provide a continuous flow of a 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 investigated the text-based approach to align English/Chinese parallel documents from the Web.

There are two major approaches for document aligning, namely length-based and text-based alignment. The length-based approach makes use of the total number of characters or words in a sentence and the text-based approach uses linguistic information in the sentence alignment (Fung & McKeown, 1997).

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., crosslingual information retrieval (Oard, 1997), bilingual lexicography, automatic translation verification, and the automatic acquisition of knowledge about translation (Simard, 1999). The translation alignment technique has been used in automatic corpus construction to align two documents (Ma & Liberman, 1999).

Given a text and its translation, an alignment is a segmentation of two texts such that the n^{th} segment of texts is the translation of the n^{th} segment of the other (Simard, Foster, & Isabelle, 1992). In other words, alignment is the process of finding relations between a pair of parallel documents.

Parallel corpus can be generated using *overt translation* or *covert translation*. The overt translation (Rose, 1981) 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) (Zanettin, 1998). The covert translation (Leonardi, 2000) is nondirectional. Multilingual documents expressing the same content in different languages are generated by the same source (Ebeling, 1998), e.g., a press release from the government, commentaries on a sports event broadcast live in several languages by a broadcasting organization.

There are three major structures of parallel documents on the World Wide Web, *parent page structure, sibling page structure*, and *monolingual sub-tree structure* (see Fig. 1). Resnik (1999) noticed that if a Web page has been written in many languages, the parent page of the Web page can contain the links to different versions of the Web page. For example, in a Web page, there are two anchor texts such as A_1 and A_2. A_1 is linked to Language 1 version and A_2 is linked to Language 2 version as shown. Another phenomenon is "sibling" pages, where the page in one language contains a

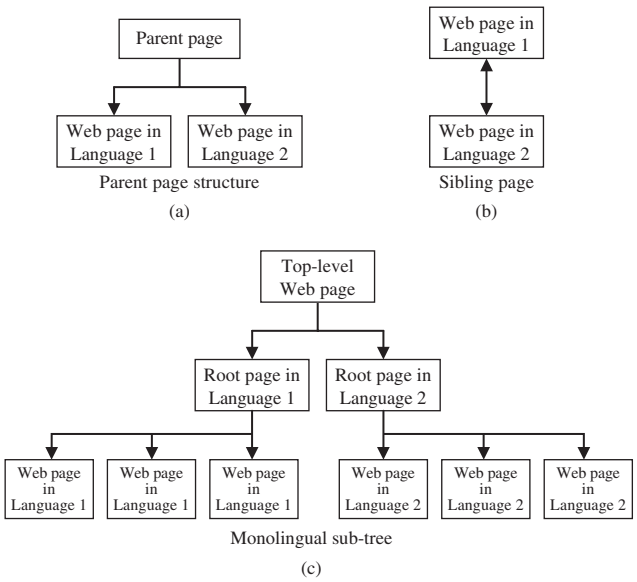

FIG. 1. Structures of parallel documents on the World Wide Web.

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 (Resnik, 1999). 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. The press release issued by the HKSAR Police Department is an example (Fig. 2).

Title Alignment

According to the Collins Cobuild dictionary, if you align something, you "place it in a certain position in relation to something else, usually along a particular line or parallel to it." A textual alignment usually signifies a representation of two texts, which are mutual translations in such a way that the reader can easily see how certain segments in the two languages correspond (Macklovitch & Hannan, 1996).

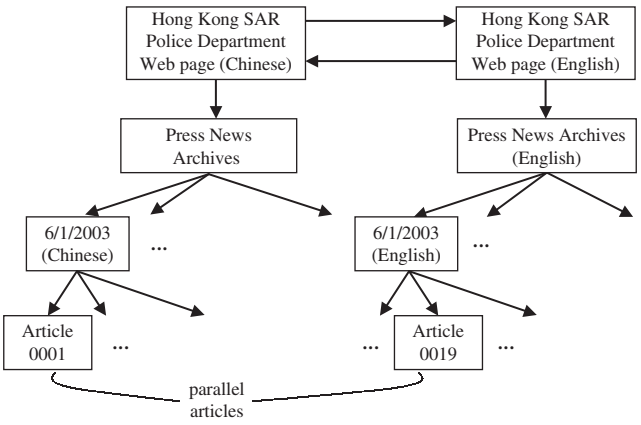

FIG. 2. Organization of Hong Kong SAR Police Department's press release articles in the Hong Kong SAR Police Department Web site.

Titles of two texts can be treated as the representations of two texts. Referring to He (2000), 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.

There are two major approaches for alignment of sentences or titles of different languages, namely length-based and text-based alignment. The length-based approach makes use of the total number of characters or words in a sentence and the text-based approach uses linguistic information in the sentence alignment (Fung & McKeown, 1997). The length-based approach works well with a clean input, such as the Canadian parliamentary proceedings; however, for cases where the languages originate from a different family of languages, such as Asian–European language pairs (Fung, 1995; Wu, 1994), these algorithms do not perform well. Text-based algorithms use lexical information across the language boundary to align sentences. Word pairs or token pairs are identified between sentences in different languages. The pair of sentences with the highest probability is aligned. However, the existing techniques have not resolved the redundancy problem not provide details for identifying the most reliable word pairs.

In our proposed text-based approach, the longest common subsequence is utilized to optimize the alignment of English and Chinese titles. The longest common subsequence (LCS) is commonly exploited to maximize the number of matches between characters of two sequences. Our alignment algorithm (Yang & Li, 2003a) has three major steps: (a) alignment at word level and character level, (b) reducing redundancy, and (c) score function.

Alignment at word level and character level. An English title, E, is formed by a sequence of English simple words, i.e., $E = e_1e_2e_3 \ldots e_i \ldots$, 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 look-up. $Translated(e_i) = \{T^1_{e_i}, T^2_{e_i}, T^3_{e_i}, \ldots, T^j_{e_i}, \ldots\}$ where $T^j_{e_i}$ 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 $T^j_{e_i}$ and C is $LCS(T^j_{e_i}, C)$. $MatchList(e_i)$ is a set that holds all the unique longest common subsequences of $T^j_{e_i}$ and C for all Chinese translations of e_i.

$$MatchList(e_i) = \bigcup_j LCS(T^j_{e_i}, C) \tag{1}$$

If there is no common subsequence of $T^j_{e_i}$ and C, $MatchList(e_i) = \varnothing$ and no reliable translation of e_i can be found in C. If there is at least one common subsequence of $T^j_{e_i}$

and C, we determine the most reliable translation based on the adjacency and length of Chinese translations found in C.

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 whose characters do not appear adjacently in the Chinese sentence. $Contiguous(e_i)$ is used to determine the most reliable translation based on adjacency.

$$Contiguous(e_i) = \{x \mid x \in MatchList(e_i) \text{ and all the characters of } x \text{ appear adjacently in } C\} \quad (2)$$

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 $Contiguous(e_i)$.

$$\text{Reliable}(e_i) = \begin{cases} \arg\max_{x \in Contiguous(e_i)} |x| & \text{if Contiguous}(e_i) \neq \varnothing \\ \arg\max_{x \in MatchList(e_i)} |x| & \text{Otherwise} \end{cases} \quad (3)$$

Resolving redundancy. 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 ε.

$$WaitList = DELE(WaitList, Reliable(e_i)) \cup DELE(MatchList(e_i) \backslash Reliable(e_i), Reliable(e_i)) \quad (4)$$

$$\text{where} \quad DELE(X, y) = \bigcup_{i=1}^{n} \text{Dele}\,(x_i, y)$$

x_i is the i^{th} element of X

Remain is a sequence that is initialized as C, and $Reliable(e_i)$ are removed from *Remain* starting from the e_i 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* to remove the redundancy.

Score function. 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.

$$Matching\_Ratio\,(E, C) = \frac{|C| - |Remain|}{|C|} \quad (5)$$

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 has the highest *Matching_Ratio*. In such a case, we shall also consider the ratio of matching determined by the portion of the English title that is able to identify a reliable translation in the Chinese title.

$$Matching\_Ratio^*(E, C) = \frac{\sum_i R(e_i)}{|E|} \quad (6)$$

$$\text{where} \quad R(e_i) = \begin{cases} 0 & \text{if Reliable}(e_i) = \varnothing \\ 1 & \text{otherwise} \end{cases}$$

If more than one Chinese title has the highest *Matching_Ratio* for the English title, E, the Chinese title with the lowest value of $|Matching\_Ratio(E, C) - Matching\_Ratio^*(E, C)|$ is considered as the counterpart of E.

An experiment was conducted to measure the precision and recall of the aligned parallel Chinese/English documents from the HKSAR Police press releases using our proposed text-based approach. The Hong Kong SAR Police press releases are developed based on covert translation. From 1st January, 2001 to 31st October, 2002, there were 2698 press articles in Chinese and 2695 press articles in English. There were only 2664 pairs of Chinese/English parallel articles. We obtained 100% precision and recall in HKSAR Police documents alignment. Experimental results show that the proposed text-based title alignment approach can effectively align the Chinese and English titles, especially the released documents from the HKSAR Police Department.

A Corpus-Based Approach: Automatic Crosslingual Concept Space Generation

Current research in crosslingual information retrieval can be divided into two major approaches: *controlled vocabulary* and *free text*. In the controlled vocabulary approach, documents are manually indexed using a predetermined vocabulary and queries from users use terms drawn from the same vocabulary. However, it imposes the limitation of the user-employed vocabulary and the selection of thesaurus highly affects the performance of the retrieval. The free text approach does not limit the usage of vocabulary; it uses the words that appear in the documents. The free text approach can be further categorized into a knowledge-based approach and corpus-based approach. The knowledge-based approach employs an *ontology* or *dictionary*. The corpus-based approach overcomes the limitation of the knowledge-based approach by making use of the statistical information of term usage in parallel or comparable corpora to construct an automatic thesaurus. In this work, we employ the corpus-based approach where the parallel corpus is automatically collected from the Web.

The semantic network knowledge base approach to automatic thesaurus generation is also referred to as a concept space approach (Chen, Ng, Martinez, & Schatz, 1997) 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 banks and shells), countries, towns, regions, religious groups, families, attackers (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 noncriminal 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 crosslingual concept space clustering model is developed based on the Hopfield network (Lin & Chen, 1996; Yang & Li, 2003b). The crosslingual 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
3. The Hopfield network
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 crosslingual 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 Web site to get the relationship between Chinese terms and English terms.

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 (1989) presents a blueprint for automatic indexing, which typically includes stop-wording and term-phrase formation. A stop-word list is used to remove nonsemantic-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 (Chen et al., 1997).

Chinese Phrase Extraction

Unlike the English language, there are not any natural delimiters in Chinese language to mark word boundaries. In our previous work, we developed the boundary detection (Yang, Luk, Yung, & Yen, 2000) and the heuristic techniques to segment Chinese sentences based on mutual information and significant estimation (Chien, 1997). Our accuracy is over 90% (Yang & Li, 2003c).

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

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j} \times w_j\right) \tag{7}$$

Equation 7 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 an important idea if it appears in the document 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, the names 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 7, 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.

Co-occurrence weight. After the calculation of d_{ij}, the asymmetric co-occurrence function (Chen & Lynch, 1992) is used to evaluate the relevance weights among concepts. For a pair of relevant terms 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 in terms of "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 the associations between two terms are not

symmetric. Therefore, we adopt the co-occurrence weight to calculate the relevance weights.

$$d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_j\right) \quad (8)$$

The co-occurrence weight, d_{ijk}, in Equation 8 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 terms j or term k is not exist in the document. The calculation is similar to the calculation in Equation 7. Therefore, the co-occurrence weight is a measure of combined weight between term j and term k.

$$Weight(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times Weighting\ Factor(T_k) \quad (9)$$

$$Weight(T_k, T_j) = \frac{\sum_{i=1}^{n} d_{ikj}}{\sum_{i=1}^{n} d_{ik}} \times Weighting\ Factor(T_j) \quad (10)$$

Equation 9 shows the relevance weights from term j to term k. Equation 10 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.

$$Weighting\ Factor(T_j) = \frac{\log\frac{N}{df_j}}{\log N} \quad (11)$$

$$Weighting\ Factor(T_k) = \frac{\log\frac{N}{df_k}}{\log N} \quad (12)$$

Equation 11 shows the weighting factor of term j. Equation 12 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. If a general term is activated during clustering, other terms associated with that general term would also be activated. Then, the size of that concept space will be large and the precision will be unavoidably low. The weighting factor is a value between 0 and 1. It carries an idea of inverse document frequency. The more number of documents contain the concept, the smaller the weighting factor.

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). The term is represented by a node in the network. The algorithm is shown below:

$$u_j(t+1) = f_s\left[\sum_{i=0}^{n-1} t_{ij} u_i(t)\right], \quad 0 \le j \le n-1 \quad (13)$$

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.

$$f_s(x) = \frac{1}{1 + \exp\left[\frac{-(x-\theta_j)}{\theta_o}\right]} \quad (14)$$

Equation 14 shows the continuous SIGMOID transformation function, which normalizes any given value to a value between 0 and 1 (Chen et al., 1997).

$$\sum_{j=0}^{n-1} [u_j(t+1) - u_j(t)]^2 \le \varepsilon \quad (15)$$

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 (determined experimentally) were used: $\theta_j = 0.1$, $\theta_o = 0.1$, $\varepsilon = 1$.

Concept Space Evaluation

Ten students from the Department of System Engineering and Engineering Management in the Chinese University of Hong Kong were recruited to examine the performance of concept space. The automatically generated concept space being evaluated 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 was to capture meaningful conceptual association between concepts. The associations form

the basis for the decisions and inferences the user uses when searching the criminal information of Hong Kong.

Experimental Design

Among these 10 graduate students, 5 subjects were Hong Kong students and the other 5 subjects came from Mainland China. All of them had been living in Hong Kong for more than one year. They used their knowledge and experience of both the Hong Kong SAR Police system and the living environment in Hong Kong to evaluate the concept space.

Then 50 among 9222 concepts were randomly selected as the test descriptors. Twenty-five among these 50 test descriptors were English concepts. The other 25 test descriptors were Chinese concepts. Each test descriptor together with its associated concepts was presented to the 10 subjects. A small portion (about 10% of the 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 was divided into two phases: recall phase and recognition phase. In the recall phase, 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 phase, 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 phase tested the ability of subjects on recognition of relevant terms. If the subjects felt the definition of a concept needed to be clarified 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 the recognition phase than on the recall phase. This confirms the statement made by Chen et al. (1996) that human beings are more likely to recognize than to recall.

Apart from the 10 student evaluators, two experimenters also carefully evaluated the 50 selected concepts in concept space, however, no noise term was added in this case. One of the experimenters was a graduate student in the Department of System Engineering and Engineering Management. The other was a graduate student in the Department of Translation. Both of them had been living in Hong Kong for more than 10 years. They had conducted research on Chinese to English translation and English to Chinese translation for more than 2 years and were familiar with the corpora in the government domain including information about crime. Since there was presently no tailored bilingual thesaurus for Hong Kong government press release articles, the experimental result provided by these two expert subjects was treated as a benchmark or human-verified thesaurus in comparison with the results provided by the 10 subjects. The additional associated concepts provided by the 10 subjects in the recall phase were examined by the two senior judges before treating them as relevant terms.

Experimental Result

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

$$\text{Concept Recall} = \frac{\text{Number of Relevant Concepts That Are Retrieved}}{\text{Total Number of Relevant Concepts}} \quad (16)$$

$$\text{Concept Precision} = \frac{\text{Number of Relevant Concepts That Are Retrieved}}{\text{Total Number of Retrieved Concepts}} \quad (17)$$

The *number of relevant concepts that were retrieved* represented the number of concepts in the automatically generated concept space that were judged as "Relevant." The *total number of relevant concepts* included the concepts in the automatically generated concept space that were judged as "Relevant" and the additional relevant concepts provided by subjects. The *total number of retrieved concepts* represented the number of automatic generated concepts in the concept space.

In the recall phase of the experiment, each subject generated 12 to 73 new concepts for 50 test descriptors. The total number of new concepts generated by all subjects was 442. The total number of new Chinese concepts was 222 and the total number of new English concepts was 220. The difference was insignificant. For each descriptor, the 10 subjects generated 0 to 8 new concepts. The average number of new concepts generated for each test descriptor by each subject was 0.884.

Table 1 presents the precision and recall of automatically generated concept space evaluated by the 10 subjects and the 2 experimenters. The differences in the precision and recall evaluated by the 10 subjects and the 2 experimenters were not significant although the precision and recall evaluated by the 2 experimenters were higher.

Translation Ability of the Concept Space

The experimenters further evaluated all the concepts in the automatically generated thesaurus. For 9222 concepts in the automatic thesaurus, there were in total 46683 associated concepts. We categorized the test descriptors into five categories: (a) the test descriptors associated with 2 concepts, (b) the test descriptors associated with 3 to 5 concepts, (c) the test descriptors associated with 6 to 9 concepts, (d) the test descriptors associated with 10 to 13 concepts, and (e) the test descriptors associated with 14 to 32 concepts. Table 2

TABLE 1. Precision and recall.

	Precision	Recall
10 Subjects	0.835	0.795
2 Experimenters	0.86	0.83

TABLE 2. Distribution of the five categories of test descriptors.

Category	Distribution	Cumulative distribution
1	14.46%	14.46%
2	26.55%	41.01%
3	31.25%	72.26%
4	18.38%	90.64%
5	9.36%	100%

presents the distribution of the test descriptors. Figure 3 presents the average number of English and Chinese associated concepts for the test descriptors in the five categories.

Table 2 shows that more than half of the test descriptors (72.26%) were associated with 10 or less concepts. Only 9.36% test descriptors (Category 5) were associated with a concept space of over 14 concepts. After examination of the test descriptors in Category 5 together with their concept space, we observed that the test descriptors in Category 5 were more general than the test descriptors in other categories. For instance, the concept "油麻地" (Yau Ma Tei, a place in Hong Kong) associated with concepts such as cellular phone theft, drug trafficking, traffic accident, 淫褻物品 (obscene articles), 刑事毀壞 (criminal damage). The information indicated the criminal rate in this area was significantly high and special attention was needed to prevent different criminal activities. Another example of a concept, "盜版光碟" (pirated VCDs), also associated with more than 14 concepts such as obscene materials, money laundering, triad offences, narcotic-related crimes. The information identified the fact that the recent trend of selling pirated VCDs may help to investigate complex organized crime and serious triad offences.

Figures 4 and 5 present the precision of the Chinese and English associated concepts for each category and the overall precision of the associated concepts for each category, respectively. It is shown that the precision decreased as the number of the automatically generated associated concepts increased. However, the precision of the Chinese associated concepts decreased more than the precision of the English associated concepts as the number of associated concepts increased. T-tests were conducted to measure the differences between the precisions of the associated Chinese and English concepts in different categories. In category 1, the difference

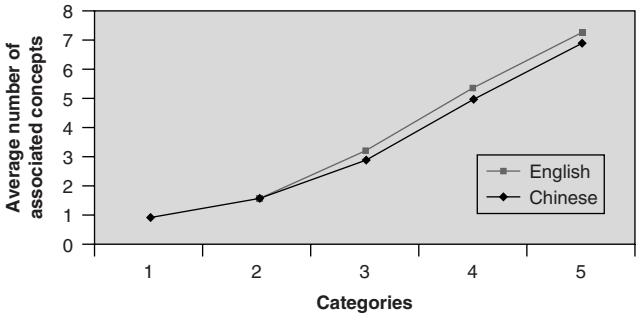

FIG. 3. The average number of Chinese and English associated concepts for each category.

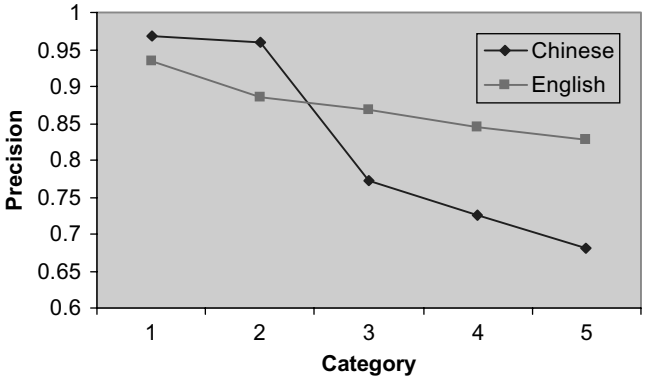

FIG. 4. Precision of the Chinese and English associated concepts for each category.

was insignificant. However, the differences were significant at $p < 0.02$. The differences may due to the significant grammatical and lexical differences between English and Chinese languages. The conceptual alteration in translation causes the lexical difference between two languages. A lexical item (word) may be a concept in one language, where *concept* is a recognizable unit of meaning in any given language. 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. In some cases, a concept may not only be represented by a word or words, but may also be represented by a morpheme, by an idiomatic expression, by tone, or by word order (He, 2000). The grammatical features such as gender, tense, voice and redundancy also enlarged the difference between English and Chinese concepts in a different category. Both the conceptual alteration and grammatical features affect the term frequencies and document frequencies of the English and Chinese concepts in the parallel corpus. The synaptic weights in the Hopfield Network are therefore different from Chinese to English.

Figures 6 and 7 presents the recall of the Chinese and English associated concepts for each category and the overall recall of the associated concepts for each category, respectively. It is shown that the recall increased as the number of the automatically generated associated concepts increased. T-tests were conducted to measure the differences between the recalls of the associated Chinese and English concepts in different categories. It was found that the differences were insignificant in all categories.

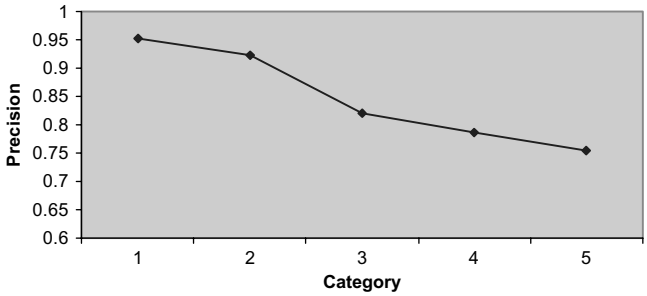

FIG. 5. Overall precision of all associated concepts for each category.

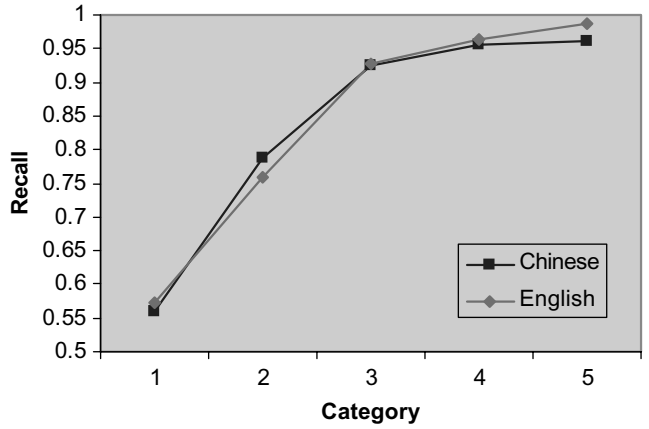

FIG. 6. Recall of the Chinese and English associated concepts for each category.

Among these 9222 test descriptors, 87.7% of them obtained their direct translations from the associated concepts. (The direct translation of the test descriptors were first identified before generating the automatically associated concepts by the Hopfield network.) It shows that the concept space generated through the Hopfield network can effectively recognize the translations of a concept in a parallel corpus and identify the highly associated concepts.

Discussion

We observed that some of the associated concepts were judged as irrelevant by the subjects because the associated concepts did not show a clear association with their test descriptor. For example, one of associated concepts for the test descriptor "走私活動" (smuggling) was "Mr. Mark Steeple" (施德博), because the Chief Inspector of the Anti-Smuggling Task Force in Hong Kong was Mr. Mark Steeple at the time. Another associated concept was "Mirs Bay" (大鵬灣) because of the recent trend of smuggling by small craft in the Mirs Bay area. However, all subjects did not have a prior knowledge of these relationships and judged them as irrelevant. Since the corpus was a dynamic resource, it was not surprising that the subjects did not have prior knowledge. For a criminal analyst, the information was important for identifying the recent trend of smuggling by small craft in the Mirs Bay area. It was thus proven that the automatic system was able to identify the associations of terms generated due to new events.

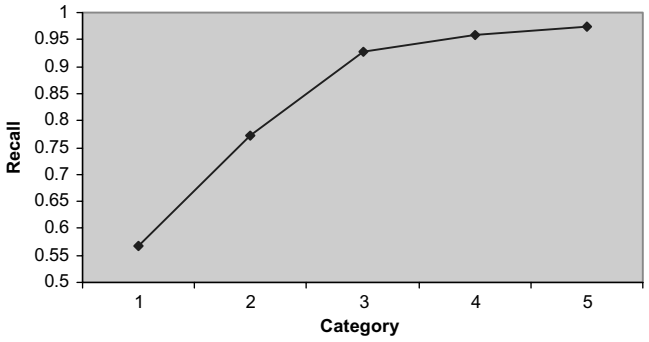

FIG. 7. Overall recall of all associated concepts for each category.

In addition, one of the associated concepts for "Golden Bauhinia Square" (金紫荊廣場) was "警察" (Police). A 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 was comprised of the Senior Inspector of Police and four flag raisers. Without knowing this, the subjects only read the concept space and judged that there was no clear association between "Police" and "Golden Bauhinia Square." This shows that the clustering process using the 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 (Larson, 1998), where *concept* is a recognizable unit of meaning in any given language (He, 2000). 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 (He, 2000). 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 a relationship is known as a *generic-specific* relationship (Larson, 1998). 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 the dictionary, but it refers to "停止服務" (stop service) in some cases (deviation). Therefore, *conceptual alternation* may occur in translation. This also caused the judges to judge some associated concepts to be irrelevant.

Nida (He, 2000) explains that conceptual alteration is caused by three major reasons: (a) no two languages are completely isomorphic; (b) different languages might have different domain vocabulary; and (c) some languages are more rhetorical than other languages.

Courtial and Pomian (1987) 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 (1969) found little overlap between term relationships generated through term associations and those presented in existing thesauri. Such 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 noncriminal activities, and

communicate through bulletin boards, e-mail, phone calls, letters, word-of-mouth, etc.—encrypted or not. Ekmekcioglu, Robertson, and Willet (1992) 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 (Chen et al., 1997).

Conclusion

The tragic event of September 11th has prompted the rapid growth of attention on 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, *crosslingual 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 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 manmade fog of deliberate deception militates against normal pattern learning from databases and causes much crucial information and any underlying knowledge 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.

Crosslingual 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 crosslingual information retrieval.

In this paper, we have presented a bilingual concept space approach using the 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 such information can help the analyst to identify important anomalies, i.e., traffic accidents frequently happen at a particular location.

Since the press release collection was dynamically generated, the subjects may not have had full prior knowledge. However, experimental results show the precision and recall for the bilingual concept space are over 78% in all cases.

Among 9222 test descriptors, 87.7% of them obtained their translations from the associated concepts. It shows that the concept space generated through the Hopfield network can effectively recognize the translations of a concept in a parallel corpus.

Acknowledgment

This project is supported by RGC Earmarked Grant for research 4335/026.

References

Bates, M.J. (1986). Subject access in online catalogs: A design model. Journal of the American Society for Information Science, 37, 357–376.

Chen, H., & Lynch, K.J. (1992). Automatic construction of networks of concepts characterizing document database. IEEE Transactions on Systems, Man and Cybernetics, 22(5) 885–902.

Chen, H., Ng, T., Martinez, J., & Schatz, B. (1997). A concept space appraoch to addressing the vocabulary problem in scientific information retrieval: An experiment on the Worm community system. Journal of The American Society for Information Science, 48(1), 17–31.

Chen, H., Schatz, B., Ng, T., Martinez, J., Kirchhoff, A., & Lin, C. (1996). 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, 18(8), 771–782.

Chien, L.F. (1997). PAT-tree-based keyword extraction for Chinese information retrieval (pp. 50–58). In Proceedings of ACM SIGIR. New York: ACM.

Courtial, J.P. & Pomian, J. (1987). A system based on associational logic for the interrogation of databases. Journal of Information Science, 13, 91–97.

Ebeling, J. (1998). Contrastive linguistics, translation, and parallel corpora [Special issue]. Meta, 43(4), 602–615.

Ekmekcioglu, F.C., Robertson, A.M., & Willett, P. (1992). Effectiveness of query expansion in ranked-output document retrieval systems. Journal of Information Science, 18, 139–147.

Fung, P., & McKeown, K. (1997) A technical word- and term-translation aid using noisy parallel corpora across language groups. Machine Translation, 12, 53–87.

Fung, P. (1995, June). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. Paper presented at the 33rd Annual Meeting of the Association for Computational Linguistics, Boston, MA.

He, S. (2000). Translingual alteration of conceptual information in medical translation: A cross-language analysis between English and Chinese. Journal of the American Society for Information Science, 51(11), 1047–1060.

Larson, M.L. (1998). Meaning-based translation: A guide to cross-language equivalence. Lanham, MD: University Press of America.

Leonardi, V. (2000). Equivalence in translation: Between myth and reality. Translation Journal, 4(4).

Lesk, M.E. (1969). Word-word associations in document retrieval systems. American Documentation, 20(1), 27–38.

Lin, C.H., & Chen, H. (1996). An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents. IEEE Transactions on Systems, Man and Cybernetics, 26(1), 75–88.

Ma, X., & Liberman, M. (1999, September). BITS: A method for bilingual text search over the web. Paper presented at the Machine Translation Summit VII, Kent Ridge Digital Labs, National University of Singapore.

Macklovitch, E., & Hannan, M.-L. (1996). Line 'em up: Advances in alignment technology and their impact on translation support tools. Paper presented at the Second Conference of the Association for Machine Translation in the Americas (AMTA-96), Montréal, Québec.

Oard, D.W. (1997). Alternative approaches for cross-language text retrieval. In D. Hull & D. Oard (Eds.), Proceedings of the AAAI Symposium in Cross-Language Text and Speech Retrieval (pp. 131–139). Menlo Park, CA: AAAI Press.

Resnik P. (1999, June). Mining the web for bilingual text. Paper presented at the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), College Park, MD.

Rose, M.G. (1981). Translation types and conventions. In M.G. Rose (Ed.), Translation spectrum: Essays in theory and practice (pp. 31–33). Albany, NY: State University of New York Press.

Salton, G. (1989) Automatic text processing. Reading, MA: Addison-Wesley.

Simard, M. (1999, June). Text-translation alignment: Three languages are better than two. Paper presented at the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, MD.

Simard, M., Foster, G., & Isabelle P. (1992, June). Using cognates to align sentences in bilingual corpora. Paper presented at the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92), Montreal, Canada.

Wu, D. (1994, June). Aligning a parallel English-Chinese corpus statistically with lexical criteria. Paper presented at the 32nd Annual Conference of the Association for Computational Linguistics, Las Cruces, New Mexico.

Yang, C.C., & Li, K.W. (2003a). Automatic construction of English/Chinese parallel corpora. Journal of the American Society for Information Science and Technology, 54(8), 730–742.

Yang, C.C., & Li, K.W. (2003b, May). Generating cross-lingual concept space from parallel corpora on the web. Paper presented at the International World Wide Web Conference, Budapest, Hungary.

Yang, C.C., & Li, K.W. (2003c, December). Segmenting Chinese unknown words by heuristic method. Paper presented at the International Conference on Asia Digital Libraries, Malaysia.

Yang, C.C., Luk, J., Yung, S., & Yen, J. (2000). Combination and boundary detection approach for Chinese indexing [Special issue]. Journal of the American Society for Information Science, 51(4), 340–351.

Zanettin, F. (1998). Bilingual comparable corpora and the training of translators. The corpus-based approach: A new paradigm in translation studies [Special issue]. META, 43(4), 616–630.