

Automatic Generation of English/Chinese Thesaurus Based on a Parallel Corpus in Laws

Christopher C. Yang

Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, E-mail: yang@se.cuhk.edu.hk

Johnny Luk

Department of Computer Science and Information Systems, The University of Hong Kong, Hong Kong.

The information available in languages other than English in the World Wide Web is increasing significantly. According to a report from Computer Economics in 1999, 54% of Internet users are English speakers ("English Will Dominate Web for Only Three More Years," *Computer Economics*, July 9, 1999, <http://www.computereconomics.com/new4/pr/pr990610.html>). However, it is predicted that there will be only 60% increase in Internet users among English speakers versus a 150% growth among non-English speakers for the next five years. By 2005, 57% of Internet users will be non-English speakers. A report by CNN.com in 2000 showed that the number of Internet users in China had been increased from 8.9 million to 16.9 million from January to June in 2000 ("Report: China Internet users double to 17 million," *CNN.com*, July, 2000, <http://cnn.org/2000/TECH/computing/07/27/china.internet.reut/index.html>). According to Nielsen/NetRatings, there was a dramatic leap from 22.5 millions to 56.6 millions Internet users from 2001 to 2002. China had become the second largest global at-home Internet population in 2002 (US's Internet population was 166 millions) (Robyn Greenspan, "China Pulls Ahead of Japan," *Internet.com*, April 22, 2002, http://cyberatlas.internet.com/big_picture/geographics/article/0,,5911_1013841,00.html). All of the evidences reveal the importance of cross-lingual research to satisfy the needs in the near future.

Digital library research has been focusing in structural and semantic interoperability in the past. Searching and retrieving objects across variations in protocols, formats and disciplines are widely explored (Schatz, B., & Chen, H. (1999). *Digital libraries: technological advances and social impacts*. IEEE Computer, Special Issue on Digital Libraries, February, 32(2), 45–50.; Chen, H., Yen, J., & Yang, C.C. (1999). *International activities: development of Asian digital libraries*. IEEE Computer, Special Issue on Digital Libraries, 32(2), 48–49.). However, research in crossing language boundaries, especially across European languages and Oriental languages, is still in the initial stage. In this proposal, we put our focus on *cross-lingual semantic interoperability* by

developing automatic generation of a cross-lingual thesaurus based on English/Chinese parallel corpus. When the searchers encounter retrieval problems, professional librarians usually consult the thesaurus to identify other relevant vocabularies. In the problem of searching across language boundaries, a cross-lingual thesaurus, which is generated by co-occurrence analysis and Hopfield network, can be used to generate additional semantically relevant terms that cannot be obtained from dictionary. In particular, the automatically generated cross-lingual thesaurus is able to capture the unknown words that do not exist in a dictionary, such as names of persons, organizations, and events. Due to Hong Kong's unique history background, both English and Chinese are used as official languages in all legal documents. Therefore, English/Chinese cross-lingual information retrieval is critical for applications in courts and the government. In this paper, we develop an automatic thesaurus by the Hopfield network based on a parallel corpus collected from the Web site of the Department of Justice of the Hong Kong Special Administrative Region (HKSAR) Government. Experiments are conducted to measure the precision and recall of the automatic generated English/Chinese thesaurus. The result shows that such thesaurus is a promising tool to retrieve relevant terms, especially in the language that is not the same as the input term. The direct translation of the input term can also be retrieved in most of the cases.

1. Introduction

As the Web-based information systems in languages other than English are growing exponentially, the demand for searching across language boundaries is obvious. We are expecting that the number of Internet users who are non-English speakers will increase to more than half of the population. The Digital Library Initiative funded by NSF/DARPA/NASA in the U.S. has laid the foundation for searching and retrieving objects across variations in protocols and formats, *structural interoperability*, and across different disciplines, *semantic interoperability* (Chen,

1998). However, the work done on *linguistic interoperability* is relatively less. A few cross-lingual information retrieval works have been done, mostly focused on European languages, such as English, Spanish, German, French, and Italian. Relatively less cross-lingual information retrieval works have been done on Oriental language or across European and Oriental languages. The difficulties of cross-lingual information retrieval between European and Oriental languages are comparatively higher than the difficulties among European languages. These are due to several reasons: 1) machine translation techniques between European and Oriental languages are rather less mature, 2) parallel/comparable corpus in European and Oriental languages are comparatively rare, 3) the grammar of European and Oriental languages are significantly different. As English and Chinese are the most popular languages in the world, the desire on an efficient and effective Chinese-English cross-lingual information retrieval system is significant although there are still many research problems to be solved.

1.1 Cross-Lingual Information Retrieval

Cross-lingual information retrieval (CLIR) refers to the ability to process a query for information in one language, search a collection of objects, including text, images, audio files, etc. and return the most relevant objects, translated into the user's language if necessary (Klavans et al., 1999; Oard & Dorr, 1996). For users who are able to read more than one language, CLIR systems allow them to submit one query and search for documents in more than one language instead of submitting multiple queries in different languages. For users who are able to read more than one language but may not be able to write fluently in other non-native languages, the CLIR systems help them to search across the language boundaries without them expressing their queries in the language that they are not familiar with. Even for users who only read a single language, the CLIR systems help them to retrieve a small portion of relevant documents from the collection of documents in other languages, and therefore fewer documents are examined by users and being translated. Current research in cross-lingual information retrieval can be divided into two major approaches: (1) *controlled vocabulary* and (2) *free text*.

In the controlled vocabulary approach, documents are manually indexed using a predetermined vocabulary and queries from users are using terms drawn from the same vocabulary (Oard & Dorr, 1996). Systems exploiting such approach bring queries and documents into a representation space by use of a multilingual thesaurus to relate the selected terms from each language to a common set of language-independent concept identifier, where the document selection is based on the concept identifier matching (Davis & Dunning, 1995; Fluhr, 1995; Flur & Radwan, 1993; Radwan & Fluhr, 1995). However, it imposes the limitation of the user-employed vocabulary and the selection of thesaurus highly affects the performance of the retrieval. Al-

though such approach is widely used in commercial and government applications, it is not practical as the number of concepts increases and becomes less manageable. The performance becomes worse as the collection of documents increases, such as the World Wide Web, because the documents generated from diverse sources cannot be standardized easily. Besides, the training of users to select search terms from the controlled vocabulary is difficult. General users are usually unable to identify the appropriate term to represent their information needs.

The free text approach does not limit the usage of vocabulary but using the words that appear in the documents. Queries or documents are translated and then retrieval is performed. The free text approach can be further divided into two approaches: (1) *knowledge-based approach* and (2) *corpus-based approach*. Knowledge-based approach employs *ontology* or *dictionary* (Ballesteros & Crosft, 1997; Hull & Grefenstette, 1996; Oard, 1997; Radwan & Fluhr, 1995). *Ontology* is an inventory of concepts, which is organized under some internal structuring principle. Multilingual thesaurus is an example of ontology to organize terminology from more than one language. Complex thesauri, which encode syntactic and semantic information about terms, are used as concept index in automatic text retrieval systems (Oard & Dorr, 1996). *Dictionary* replaces each term in the query with an appropriate term or set of terms in the target language (Salton, 1970). However, the knowledge-based approach encounters several problems. Terms may be ambiguous and may have multiple meanings. The thesaurus or dictionary may not have terms that are essential for a correct interpretation of the query. For examples, technical terms, abbreviation, names of persons, organization or events may not be included in the thesaurus or dictionary. There are also new terms entering the lexicon all the time since language usage is a creative activity.

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. 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 across large document collections to construct a statistical translation model for cross-lingual information retrieval. Several works have been done based on this approach. Landauer and Littman use Latent Semantic Indexing (LSI) to construct a multi-lingual semantic space [Dumais et al, 1997; Landauer & Littman, 1990]. Experiments are conducted on a French-English collection. Davis and Dunning (1995a, b) apply evolutionary programming on a Spanish-English collection. Sheridan and Ballerini (1996) apply thesaurus-based query expansion techniques on a comparable Italian-English collection. However, there is a lack of corpus-based approach being applied between European and Oriental languages, in particular between English and Chinese, mainly due to their significant difference in grammar and low availability of such corpus.

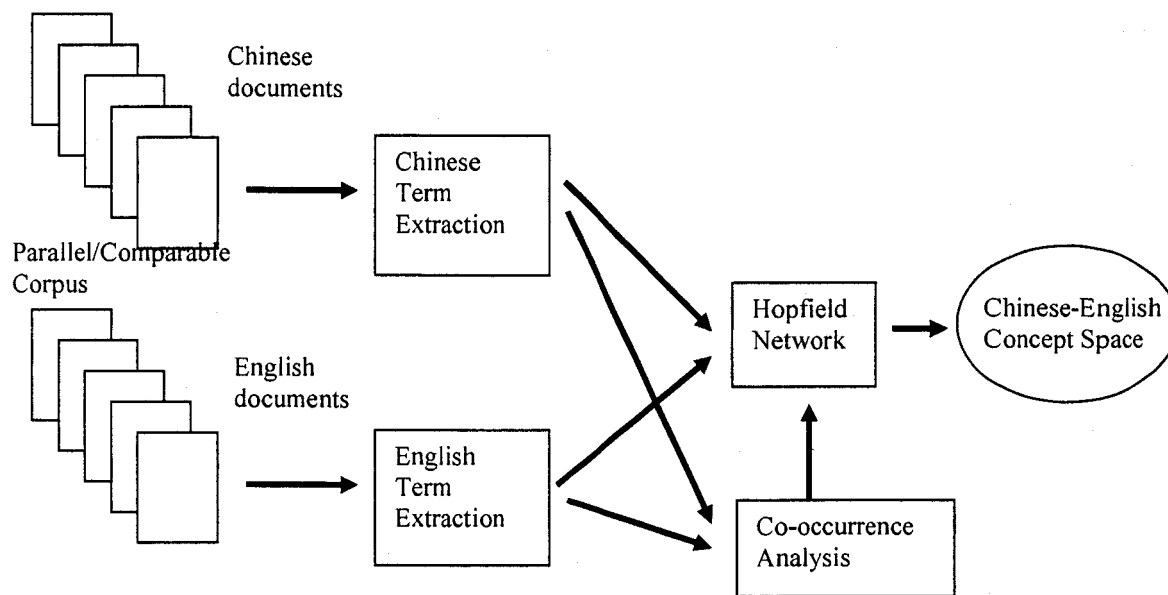


FIG. 1. Automatic thesaurus generated from a parallel/comparable corpus.

2. Multilingual Corpus

Multilingual corpus is a collection of text in electronic form (written language corpus) where texts in different languages are put together either based on parallelism or comparability. Multilingual corpus constructed based on parallelism and comparability are known as parallel corpus and comparable corpus, respectively. Parallel corpus can be developed using overt translation or covert translation. The overt translation 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) (Rose, 1981). The covert translation is non-directional. Multilingual documents expressing the same content in different languages are generated by the same source (Leonardi, 2000). Therefore, none of the text in each pair of such parallel corpus is marked as translated text or source text.

Comparable corpus consists of texts in multiple languages composed independently, which shared similar criteria of content, domain, and communicative function. Criteria for creating comparable corpora depends on the homogeneity of texts, both with or across languages, in terms of features such as subject domain, author-reader relationship, text origin and constitution, factuality, technicality and intended outcome (Zanettin, 1998).

Several techniques have been developed to construct parallel corpus automatically. The most prominent system that generate parallel corpus from the World Wide Web is Structural Translation Recognition for Acquiring Natural Data (STRAND), developed by Resnik (1998, 1999). STRAND consists of three modules, candidate generation module, candidate evaluation module and candidate pair filtering module. A similar system has also been developed by Nie (Nie et al., 1999). However, these systems focus on generating parallel corpus in European languages only.

2.1 Automatic Cross-lingual Thesaurus

In this paper, we develop a English/Chinese cross-lingual thesaurus based on a English/Chinese parallel corpus collected from the Bilingual Legal Information System (BLIS) of the HKSAR Government. Part of the BLIS parallel corpus is based on the overt translation and part of it is based on covert translation depending on the date of the document released. Using the co-occurrence analysis of English and Chinese terms extracted in the parallel corpus, the synaptic weights of the Hopfield network are determined. The cross-lingual thesaurus is generated automatically by training in the Hopfield network. Such cross-lingual thesaurus supports cross-lingual information retrieval by suggesting semantically relevant terms in different languages so that human users who are not familiar with the languages are able to obtain better retrieval performance.

3. Automatic Generation of English/Chinese Thesaurus

Unlike the controlled vocabulary approach and knowledge-based approach, the cross-lingual thesaurus is generated from the parallel corpus automatically without human intervention in the corpus-based approach. Knowledge representations are created from the corpus without pre-defined language representation and linguistic information by human experts. The automatic Chinese-English thesaurus generation system basically consists of five components: i) the parallel/comparable Chinese-English corpus, ii) English term extraction, iii) Chinese term extraction, iv) co-occurrence analysis, and v) Hopfield network as shown in Figure 1. Important terms (or concept descriptors) are first extracted from the parallel or comparable corpus automatically using the English term extraction and Chinese term

extraction modules. To be more precise, a *concept* is a recognizable unit of meaning in any given language (He, 2000). A term is a word or expression that has a precise meaning in peculiar to a subject. Lin (Lin & Chen, 1996) refers terms extracting from a corpus as concept descriptors. The co-occurrence analysis module measures the similarities between the extracted terms. The extracted terms and their similarities are then modelled as the nodes and the weights between the nodes of the Hopfield network. The Hopfield network resembles an associate network of the extracted English and Chinese terms from the parallel corpus and the classifying behaviour of the Hopfield network converge the strongly associated neighbors to generate the concept space. A concept space is a cluster of related concept descriptors to represent the subject matter of a corpus (Lin & Chen, 1996).

3.1 Extraction of English and Chinese Terms

Term extraction or indexing has been one of the most important research issues in information science. Without proper indexes, effective representation and retrieval of documents are impossible. A traditional human indexer recognizes and selects the essence and then represents them. However, human indexing is time consuming and expensive. An automatic term extraction (or automatic indexing) is important for extracting important terms from a large collection of documents.

Different languages exhibit different linguistic and grammatical characteristics. The difference is particularly significant between European languages and Oriental languages, such as English and Chinese. These characteristics strongly affect how English and Chinese documents are segmented and indexed. English is a phonographic language. Every word has one or multiple meanings. On the contrary, Chinese is a pictographic language. Every word or character has a unique meaning although ancient Chinese writing is more concise with a single character conveying several meanings. Chinese words usually consist of more than one character to convey precise meaning in modern Chinese writing. Therefore, the techniques to extract English terms and Chinese terms are different.

3.1.1 English term extraction. The term extraction of unstructured English documents includes four major steps: (1) lexical analysis, (2) stopwording, (3) stemming, and (4) term-phrase formation. The purpose of *lexical analysis* is treating digits, hyphens, punctuation marks, and the case of letters. It is a process of converting the text of the document representing as a stream of characters into the candidate words representing as a stream of words. The objective of *stopwording* is filtering words without any specific meaning. Articles, prepositions, and conjunctions are natural candidates for a list of stopwords. Some verbs, adverbs and adjectives are also potential candidates. Such stopwords occur frequently in documents but they are not good dis-

criminators. Eliminating the stopwords also reduces the size of indexed terms by 40% or more. *Stemming* removes affixes of the words to identify the root forms of the words. Plurals, gerund forms, and past tense suffixes are typical examples of syntactical variations of the stem words. For example, “extract” is the stem word of “extracting,” “extracts,” and “extracted.” *Term-phrase formation* groups adjacent words extracted from a phrase segment to form term phrases. Due to the phonographic characteristics of English, term phrases with more words carry more precise meaning than individual terms. For example, “cross-lingual information retrieval” has three words and is able to form three term phrases, “cross-lingual information retrieval,” “cross-lingual information,” and “information retrieval.” Each of these term phrases has a unique semantic.

3.1.2 Chinese term extraction. Unlike English language, there are not any natural delimiters in Chinese language to mark word boundaries. Written Chinese consists of strings of characters (or ideographs) separated by punctuation. The smallest indexing units in Chinese documents are words, while the smallest units in a Chinese sentence are characters. A character can perform as a word with meaning(s) or function as an alphabet forming a short word with one or more adjacent characters and having specific meaning (Ikehara, 1995). In English or other languages using Roman or Greek-based orthographies, spacing often reliably indicates word boundaries. In Chinese, a number of characters are placed together without any delimiters indicating the boundaries between consecutive characters. The first step of word-based indexing is segmentation. Segmentation is the process of breaking a string of characters into words. In another word, segmentation is determining the boundaries of single or multi-character words in a string. Word segmentation in Chinese is known to be a difficult task, especially for unknown words, such as names, locations, translated terms, technical terms, abbreviations, etc.

Previous work on Chinese word segmentation can be divided into three categories, lexical rule-based approach (Nie et al., 1994; Wu & Tseng, 1995), statistical approach (Gan et al., 1996; Lua, 1990; Sproat and Shih, 1990), and hybrid approach that is based on statistical and lexical information (Leung & Kan, 1996; Nie et al., 1994). The *lexical rule-based approach* that deals with mentation ambiguities is also known as the dictionary-based approach and the most popular method is the maximum matching method. Starting from the beginning or end of a sentence, the maximum matching method groups the longest initial sequence of characters that matches a dictionary entry as a word and they are called forward maximum matching and backward maximum matching. The idea is to minimize the number of words segmented. A variation of the maximum matching is the minimum matching or shortest matching, which treats a word as the shortest initial string of characters that match a dictionary entry. However, the major concerns of lexical rule-based approach are (i) how to deal with

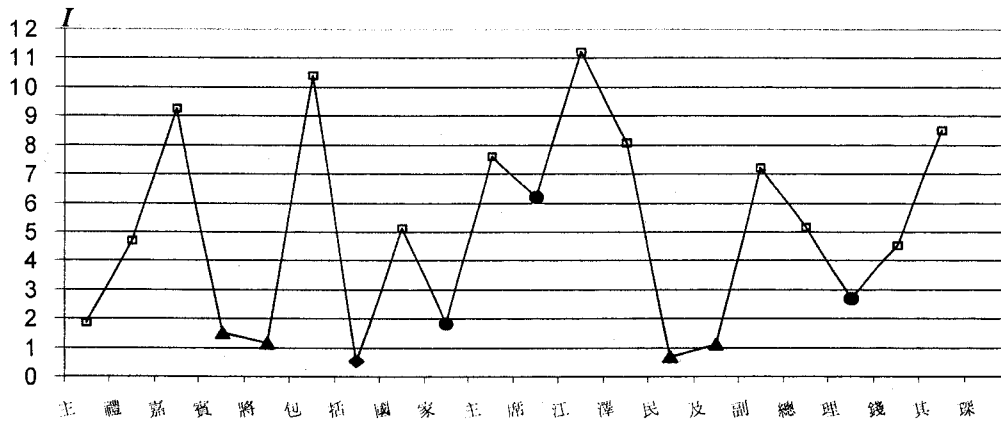


FIG. 2. Mutual information, $I(c_j, c_{j+1})$, of each pair of adjacent characters.

ambiguities in segmentation, and (ii) how to extend the lexicon beyond the dictionary entries. The *statistical approach* extracts the segments based on the frequencies of n-grams in the corpus. Sproat uses mutual information as a statistical measurement of any two characters (Sproat & Shih, 1990). Two adjacent characters with highest value of mutual information are extracted first. Bi-grams are recursively extracted until no more pairs can be extracted. However, n-grams longer than two characters are not segmented. Chien (1997) developed a PAT-tree-based approach for keyword extraction. All of the lexical patterns without limitation of pattern length are first extracted. A mutual-information-based filtering algorithm is then applied to filter out the character strings in the PAT tree. A refined method based on a common-word lexicon, a general domain corpus and a keyword determination strategy are utilized finally. The performance is good but building a PAT tree is time consuming and large space overhead is required because all of the lexical patterns are investigated.

In this work, we adopt our previously developed technique, boundary detection (Yang et al., 2000a), to extract the Chinese terms. The boundary detection is a statistical approach using the mutual information. Mutual information, $I(c_1, c_2)$, measures association between two consecutive characters, c_1 and c_2 , in a sentence. Characters that are highly associated are grouped together to form words.

$$I(c_j, c_{j+1}) = \log_2 \left(\frac{f(c_j, c_{j+1})/N}{[f(c_j)/N][f(c_{j+1})/N]} \right) = \log_2 \left(\frac{Nf(c_j, c_{j+1})}{f(c_j)f(c_{j+1})} \right) \quad (1)$$

where c_i and c_j are Chinese characters, $f(c_i, c_j)$ is the frequency of $c_i c_j$, $f(c_i)$ is the frequency of c_i , $f(c_j)$ is the frequency of c_j and N is the total number of documents.

The boundary detection algorithm makes use of the mutual information and the analogy of the edge detection technique on images. If the mutual information value between adjacent characters is less than a threshold, it means the two characters are independent and therefore it is considered as a segmentation point. Similar to gradient operators in image edge detection, changes in mutual information

value are used to identify the points of valley and bowl shaped curves. These abrupt changes in mutual information values are also employed to detect the segmentation points. Figure 2 illustrates the changes of mutual information values for the sentence, 主禮嘉賓將包括國家主席江澤民及副總理錢其琛 (The guests of ceremony include the country chairman, Jiang Zemin, and vice president, Qian Qichen). Using boundary detection, the sentence is segmented to [主 禮 嘉 賓][將][包 括][國 家][主 席][江 澤 民][及][副 總 理][錢 其 琛] ([The guest of ceremony] [will] [include] [country] [chairman] [Jiang Zemin] [and] [vice president] [Qian Qichen]). Experimental results have shown that the boundary detection has over 92% accuracy and is able to identify most of the unknown words.

3.1.3 Automatic term selection. After extracting English and Chinese terms from the English and Chinese parallel corpus, only the most significant terms will be employed to form the concept space. The significant terms are selected based on the term weights, d_{ij} , computed by the term frequencies, inverse document frequencies and the length of terms. The term weight, d_{ij} , represents the relevance weight of term j in document i .

Given the English/Chinese parallel corpus, N pairs of English documents and Chinese documents, E_i and C_i ($i = 1, 2, \dots, N$), are aligned. For each pair of English and Chinese documents, doc_pair_i , the term weight for each extracted English term, $term_j$, and each extracted Chinese term, $term_{j^*}$, are computed as follows:

$$d_{ij} = tf_{ij} \times \log \left(\frac{N}{df_j} \times w_j \right) \quad (2)$$

$$d_{ij^*} = tf_{ij^*} \times \log \left(\frac{N}{df_{j^*}} \times w_{j^*} \right) \quad (3)$$

where tf_{ij} is the term frequency of English term, $term_j$, in document pair, doc_pair_i , df_j is the document frequency of

English term, $term_j$, w_j is the length of English term, $term_j$, tf_{ij}^* is the term frequency of Chinese term, $term_i^*$, in document pair, doc_pair_i , d_{ij}^* is the document frequency of Chinese term, $term_j^*$, and w_j^* is the length of Chinese term, $term_j^*$.

The term that occurs more frequently indicates itself as a good descriptor of the document. On the other hand, the term that occurs frequently on many documents implies itself as a general term that does not have any specific meaning. Therefore, a term, which has a high tf_{ij} and low d_{ij} , corresponds to a good keyword of the documents. Besides, multiple-word terms convey more precise semantic meanings; therefore, they are also assigned with heavier weights.

For each document pair, doc_pair_i , five English terms and five Chinese terms that have the highest d_{ij} and d_{ij}^* , respectively, are selected as inputs to the co-occurrence analysis and Hopfield network to generate the cross-lingual thesaurus.

3.2 Co-occurrence Analysis

The co-occurrence analysis measures the similarities between the extracted terms in either English or Chinese from the English/Chinese parallel corpus. Based on the idea that relevant terms often co-occur in the same document pair, the co-importance weights and relevance weights are computed using the statistics of term frequencies and inverse document frequencies.

The co-importance weight, d_{ijk} , between term j and term k in document pair, doc_pair_i , are computed as follows, where term j and term k can be English terms ($term_j$ and $term_k$) or Chinese terms ($term_j^*$ and $term_k^*$):

$$d_{ijk} = tf_{ijk} \times \log N/df_{jk} \quad (4)$$

where tf_{ijk} is the minimum of tf_{ij} and tf_{ik} in document pair, doc_pair_i , and df_{jk} is the document frequency of both term j and term k

The relevance weights between term j and term k is then computed as follows:

$$W_{jk} = \frac{\sum_{i=1}^N d_{ijk}}{\sum_{i=1}^N d_{ij}} \quad \text{and} \quad W_{kj} = \frac{\sum_{i=1}^N d_{ijk}}{\sum_{i=1}^N d_{ik}} \quad (5)$$

where N is the number of document pairs.

The relevance weights between term j and term k are asymmetric. If term j is more significant than term k in the parallel corpus ($\sum_{i=1}^N d_{ij} > \sum_{i=1}^N d_{ik}$), then W_{jk} is less than W_{kj} . That means, the more significant term will have less influence to the less significant term. For example, given the term “Tung Chee Hwa” (the Chief Executive of Hong Kong Special Administrative Region), we expect to include the

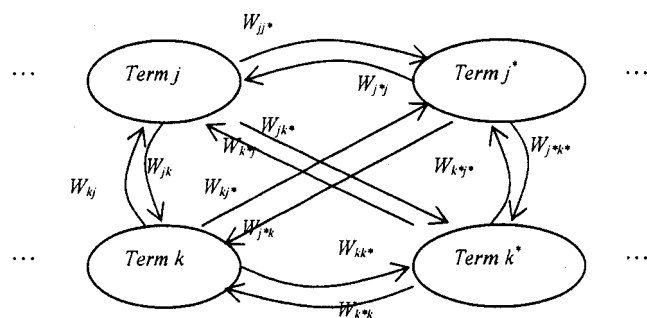


FIG. 3. Associate network of extracted terms from the parallel corpus.

relevant and significant term such as “Chief Executive” into the concept space of “Tung Chee Hwa”. However, we do not expect to include the relevant but much less significant term such as “Hong Kong” into the concept space of “Tung Chee Hwa”. If the less significant terms are also included to the concept space, the training of the Hopfield network may not converge because the concept space may continuously expand until it includes a large collection of relevant and irrelevant terms.

Given the relevance weights between the terms extracted from the parallel corpus, their associations can be modelled as an association network. The nodes of the network represent the extracted terms in either Chinese or English and the bi-directional arcs represent the relevance weights between the corresponding terms. Figure 3 illustrates such an association network.

3.3 Hopfield Network

The Hopfield network can be considered as a nonlinear associate memory or content-addressable memory. The objective of Hopfield network is to retrieve a pattern (concept space) stored in memory in response to the presentation of an incomplete or noisy version of that pattern (Haykin, 1994). The content-addressable memory is error-correcting that overrides inconsistent information in the cues presented to it.

The Hopfield network, as an associate memory, maps a fundamental memory, ξ_μ , onto a stable point, s_μ , of a dynamic system. When a pattern containing partial but sufficient information about one of the fundamental memories is presented to the network, the pattern is represented as a starting point in the phase space. The system will then evolve with time and eventually converge onto the memory state. The operation of the Hopfield network has two phases, the storage phase and the retrieval phase. In the storage phase, the synaptic weight from neuron j to neuron k , W_{jk} , are generated. In the retrieval phase, activation of neurons will be applied until convergence.

To generate the cross-lingual thesaurus, the Hopfield network models the associate network as illustrated in Figure 3 and transforms a noisy pattern into a stable state representation. The synaptic weights in the storage phase

are generated by the co-occurrence analysis. In the canonical Hopfield Networks, if two nodes behave similarly in a sample pattern, the weight between these nodes is usually adjusted with a higher value. Similarly, the relevance weights that computed by Equation (5) are assigned as the synaptic weights since the relevance weights correspond to how these nodes are strongly associated. The higher the relevance weights between two terms, the stronger the corresponding nodes are associated. In the retrieval phase, a searcher starts with an English term. The Hopfield network spreading activation process will identify other relevant English term and gradually converge toward heavily linked Chinese term through association (or vice versa).

The Hopfield network algorithm is briefly described as follows:

1. Assigning Connection Weights

Each neuron is assigned a Chinese or English term and the relevance weights are considered as synaptic weight.

2. Initialization:

$$u_i(0) = x_i, 0 \leq i \leq n - 1$$

$u_i(t)$ is the output of neuron i at time t . x_i indicates a value of neuron i where x_i has a value between 0 and 1. $x_i = 1$ if x_i represents the input term; otherwise, $x_i = 0$. n is the number of terms in the Hopfield network

3. Activation and Iteration

$$u_j(t + 1) = f_s \left(\sum_{i=0}^{n-1} W_{ij} u_i(t) \right), 0 \leq j \leq n - 1$$

where

$$f_s(x) = \frac{1}{1 + \exp \left[\frac{-(x - \theta_j)}{\theta_0} \right]}$$

W_{ij} is the synaptic weight from node i to node j n is the number of terms in the Hopfield network θ_j is a threshold θ_0 is a variable to adjust the shape of the sigmoid function, $f_s(x)$

4. Convergence

Repeat Process 3 until the change in terms of output in the output layer between two iterations is less than ϵ .

$$\sum_{j=0}^{n-1} (\mu_j(t + 1) - \mu_j(t))^2 < \epsilon.$$

4. Experiments

In this paper, we have conducted an experiment to measure the performance of the automatic generation of cross-lingual thesaurus based on a parallel corpus in law. There are two major focuses in the experiment:

- (1) Evaluation of the general performance of the concept space clustering in terms of precision and recall.
- (2) Investigation of the precision/recall of the English and Chinese terms in the concept space provided that the input terms are in different languages.

4.1 Experiment Testbed

Both English and Chinese were established as the official languages of Hong Kong for the purposes of communication between the Government and the public with the passage of the Official Languages Ordinance since 1974. However, before 1987, the statute law is enacted in the English language only under the British Hong Kong Government. The need for the production of an authentic Chinese version of Hong Kong's written law was officially brought about by the Sino-British Joint Declaration on the Question of Hong Kong signed in 1984. The Joint Declaration provided that "in addition to Chinese, English may also be used in organs of government and in the courts in the Hong Kong Special Administrative Region (HKSAR)". In 1986, the Hong Kong Royal Instructions were amended to allow laws to be enacted in English or Chinese. In 1990, the Basic Law of the HKSAR promulgated that, in addition to the Chinese language, English may also be used as an official language by the executive authorities, legislature and judiciary of the HKSAR. The Law Drafting Division of the Department of Justice is responsible for preparing the two language texts of all ordinances and subsidiary legislation introduced by the Government. Based on this history background, legal documents are usually provided in both English and Chinese nowadays in Hong Kong. However, information about Hong Kong legal and justice on the World Wide Web may not be available in both English and Chinese. Therefore, searching across the language boundary is desired and automatic cross-lingual thesaurus is helpful for such tasks.

In the experiment, a parallel corpus was collected from Bilingual Laws Information System (BLIS) of Department of Justice of HKSAR (<http://www.justice.gov.hk>). BLIS is a Web-based system of Chinese and English version of the Laws of Hong Kong. The parallel English-Chinese documents originated before 1987 are based on overt translation, where the English documents are the source and the Chinese documents are translated. However, the parallel English/Chinese documents originated after 1987 are based on covert translation. Totally, 1241 document pairs were collected. The size of the parallel corpus is 11.6MB with 7.6MB of Chinese documents and 4MB of English documents.

Term extraction as described in Section 3.1 was performed on the parallel corpus. There are 154 extracted

TABLE 1. Statistics of Bilingual Laws Information System (BLIS).

Number of English/Chinese document pairs	1241
Size of parallel corpus	11.6 MB
Size of English documents	4 MB
Size of Chinese documents	7.6 MB

segments in each document pair on average. The most significant five English terms and five Chinese terms were selected from each document. There are totally 5171 unique terms selected from the parallel corpus with 3172 English terms and 1999 Chinese terms.

Co-occurrence analysis and the Hopfield Network-based concept space clustering of 5171 terms was performed after term extraction. A machine with 2 processors, 256 MB of RAM, MS SQL Server 6.5 running as database backend and JDK1.1.7 was used to process the concept clustering. The whole process, including term extraction, co-occurrence analysis, and Hopfield network, took about 100 hours to complete.

4.2 Size of Concept Space Retrieved by Hopfield Network

Table 3 shows the distribution of the size of cross-lingual concept space generated from 5171 terms. More than half of the terms generated a concept space of less than 10 terms. However, there are over 16% of terms generated a concept space of over 14 terms and some of them even have 40 terms in the concept space.

Since the synaptic weights are asymmetric in the Hopfield Networks, the convergence is not guaranteed. Among the terms (16%) that generated a concept space of over 14 terms, about one third of them do not converge. Since convergence does not occur, the size of the concept space continuous to increase, but we stop the iterations when the size of the concept space reaches 40. Any terms that are added to the concept space, after 40 terms are retrieved, are mostly irrelevant to the input term. Careful evaluation of the characteristics of the input terms, where convergence does not occur, shows that general input terms are usually difficult to converge. For instance, "Hong Kong" and "Ordinance" are general terms in the parallel corpus. Many terms are added to the concept space in a few iterations and the size of the concept space continuous to grow.

For each input term, there are Chinese and English terms retrieved by the Hopfield Network. Since there are significant differences between Chinese and English language, it is desired to identify if there is any significant difference in the

TABLE 2. Statistics of extracted English and Chinese terms from the parallel corpus.

Number of English and Chinese terms extracted totally	5171
Number of English terms	3172
Number of Chinese terms	1999

TABLE 3. Distribution of size of cross-lingual concept space.

Size of concept space	Distribution	Cumulative distribution
2	17%	17%
3-5	22%	39%
6-9	23%	62%
10-13	22%	84%
14 or over	16%	100%

retrieval of number of English terms and Chinese terms by the Hopfield Network when the input terms are of different languages. Figure 4 shows that the average number of Chinese and English members of the retrieved terms. T-tests have been conducted to measure their significance of difference as shown in Table 4. It shows that if the size of the retrieved term is less than or equal to thirteen, the number of retrieved Chinese terms and retrieved English terms is approximately the same. However, if the size of the retrieved term is over fourteen and the input term is Chinese, the number of retrieved Chinese terms is significantly (significant level of 10%) larger than the number of the retrieved English terms.

Chinese and English are two languages with significant differences in their structure and grammar. English is a phonographic language where each English word has one or more independent meanings. For instance, "bank" may refer to a place of business for the custody, loan, exchange, or issue of money or the rising ground bordering of a lake, river, or sea. However, Chinese is based on pictographs where each Chinese character (ideograph) can perform as a word with unique meaning or function as an alphabet forming a short word with specific meaning. Spacings are reliable word boundaries for English; however, segmentation of Chinese sentences relies on dictionary and word frequencies in the corpus. Therefore, as shown in Table 1, the number of unique terms extracted in English is significantly higher than that in Chinese. However, as shown in Table 4, there is no significant difference in the number of English and Chinese members obtained for different size of concept space retrieved no matter the input terms are English or Chinese except when the size of concept space is over 14 and the input term is Chinese. As a result, the number of available English and Chinese terms in the Hopfield Network does not affect the number of retrieved English and Chinese terms in the concept space. When the size of concept space is over 14, the significance difference in the number of English and Chinese members obtained may due to high percentage of training without convergence.

4.3 Experiment Methodology

Ten students of the Department of Law at the University of Hong Kong, who were studying Postgraduate Certificate in Laws (PCLL), were taken as the subjects of the experiment. They evaluated the performance of concept space

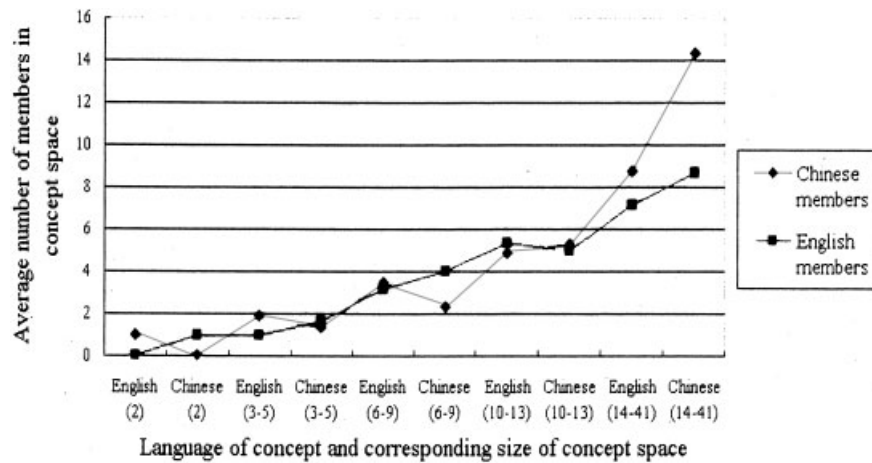


FIG. 4. Average number of members in English and Chinese for different concept space size and language of input terms.

clustering based on their professional knowledge in the laws of Hong Kong. Since the parallel corpus we study here is a specific domain in law, we select the subjects with adequate knowledge in the usage of English and Chinese terms in the laws of Hong Kong SAR without much demographical difference. There are two sessions of evaluation. In the first session, subjects are asked to judge the relevance of the terms returned by the Hopfield Network to the randomly selected input terms. In the second session, subjects are asked to make suggestions of relevant terms given the randomly selected terms. The randomly selected terms in these two sessions are two distinct sets of terms. Therefore, the subjects did not see the returned terms from the Hopfield network before they made any suggestion of relevant terms for the randomly selected terms nor did they suggest any relevant terms before they judged the relevance of the returned terms from the Hopfield Network for any randomly selected terms.

In the first session, 50 terms were randomly selected as the inputs of the Hopfield Network. Each term together with the retrieved terms by the Hopfield Network was listed one by one to our subjects. A small portion (10% of total number of retrieved terms) of noise terms was added to the concept space in order to reduce bias of the subjects to the computer-generated result. The subjects were asked to use

their professional knowledge to determine whether the retrieved terms are relevant to the input term. The subjects were also asked to mark the retrieved term if it is the direct translation of the input term. The average precision of the terms was calculated and recorded.

$$Precision = \frac{\text{Number of relevant retrieved terms}}{\text{Number of retrieved terms}}$$

In the second session, 50 terms were randomly selected and presented to the subjects. The subjects were asked to suggest the relevant terms according to their experience in the laws of Hong Kong. The concept spaces generated by the Hopfield Network are compared with the subjects suggested terms and the average recall of the concept space was calculated and recorded.

$$Recall = \frac{\text{Number of relevant retrieved term}}{\text{Number of term suggested by subjects}}$$

4.4 Experimental Results

4.4.1 Measurement of the reliability of interjudge agreement. In this study, inter-rater reliability is applied to evaluate the reliability of the subjects' agreement on precision and recall of the concept space clustering by the Hopfield Network. Cohen's Kappa (K), with a range between 0 and 1 to represent the strength of agreement, is utilized as the measurement (Cohen, 1960).

$$K = \frac{P_o - P_c}{1 - P_c}$$

where

P_o is the observed percentage of agreement
 P_c is the expected percentage of agreement

TABLE 4. T-tests between the number of members in English and Chinese for different size of concept space and language of input term.

Size of concept space	Input term	p-value
3-5	English	0.323
3-5	Chinese	0.576
6-9	English	0.767
6-9	Chinese	0.209
10-13	English	0.719
10-13	Chinese	0.812
14-41	English	0.126
14-41	Chinese	0.053

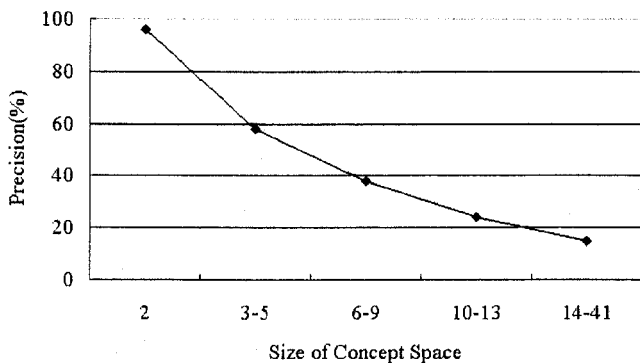


FIG. 5. Precision of concept space with different sizes.

K equals to 0.896 and 0.823 in the first session and the second session, respectively. It shows a substantial reliability of agreement among the subjects. In particular, 97% of the noise terms are rejected by the subjects in the first session and 82% of the terms suggested by the subjects in the second session are the same.

4.4.2 Precision and recall. Figure 5 shows the result of precision against the size of the concept space. As the number of the retrieved terms increases, the precision decreases. The precision is over 90% given that the number of retrieved term is less than two. If the number of retrieved terms is less than six, the precision is up to 80% in average. As the size of concept space continues to increase, the precision may not be acceptable any more.

Due to the significant grammatical and lexical difference between English and Chinese languages, the term frequencies and document frequencies of the English and Chinese terms in the parallel corpus may affect the synaptic weights

TABLE 5. T-tests between the precision of the retrieved English terms and Chinese terms for different sizes of concept space and languages of input term.

Size of concept space	Input term	p-value
3-5	English	0.220
3-5	Chinese	0.582
6-9	English	0.069
6-9	Chinese	0.052
10-13	English	0.128
10-13	Chinese	0.024
14-41	English	0.495
14-41	Chinese	0.591

in the Hopfield Network and hence affect the precision and recall of the retrieved concept space. Several situations may occur in translation between English words and Chinese words due to lexical differences. One word in English may be translated into one word in Chinese, or several words in English may be translated into one word in Chinese or vice versa. In some cases, some words are not translated from English to Chinese at all, or a word in English may not be always translated in the same way in Chinese or vice versa. In other cases, a word in English may be translated into morphological or syntactic phenomena rather than a word in Chinese or vice versa. Grammatical differences, such as word order, redundancy, tense, and voice, may also affect the frequencies of English terms and Chinese terms appearing in the parallel corpus. As a result, it is desired to determine if such grammatical and lexical differences make a significant difference in the precision and recall of the English and Chinese terms in the concept space retrieved by the Hopfield Network.

Figure 6 shows the precision of the retrieved English and Chinese terms. If the input term is English, the precision of

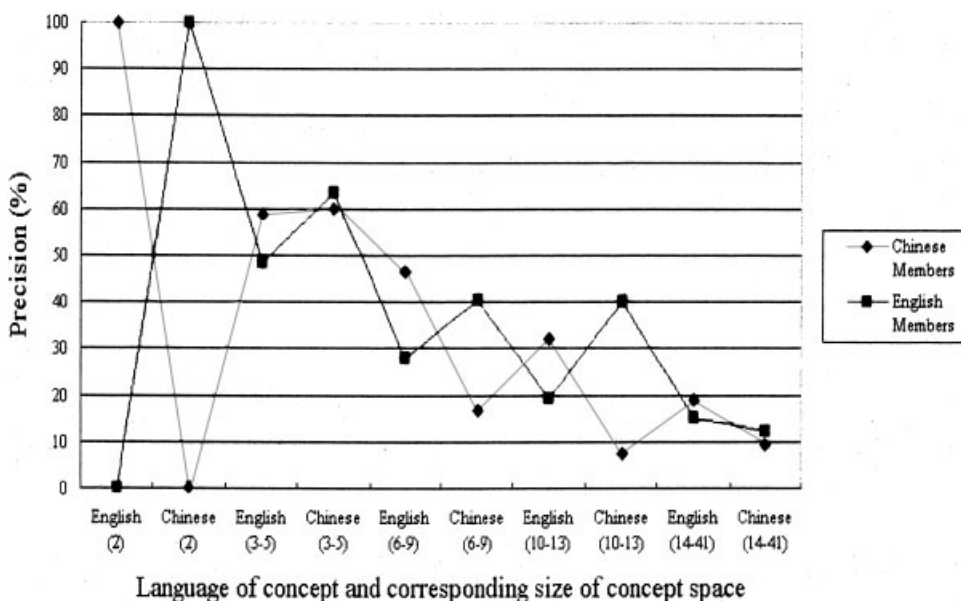


FIG. 6. Precision of concept space with different size and language of input terms.

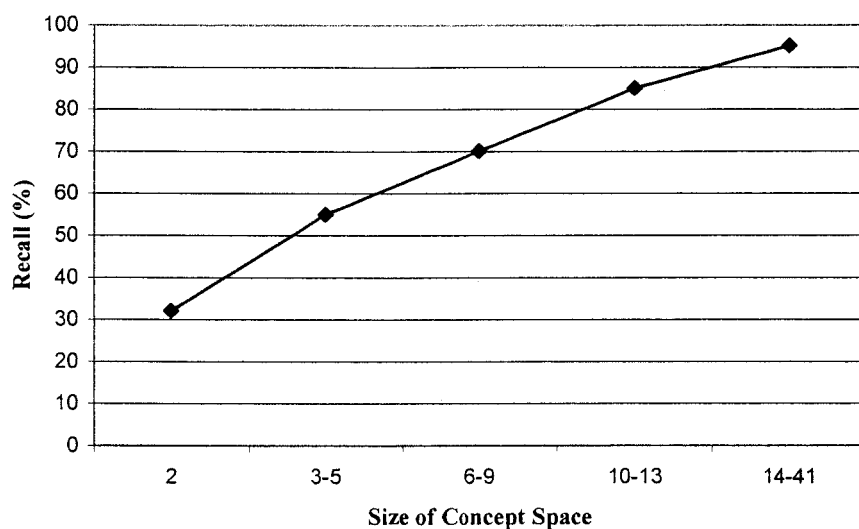


FIG. 7. Recall of concept space with different sizes.

the retrieved Chinese terms is always higher than the number of the retrieved English term. Similar results are observed if the input term is Chinese. That means the precision of the retrieved terms in the opposite language is always higher than the precision of the retrieved terms in the same language. T-tests have been conducted to measure their significance of difference as shown in Table 5. It shows that if the size of the retrieved term is less than or equal to five but greater than two or greater than or equal to 14, the precision of retrieved Chinese terms and retrieved English terms has no significant difference. However, if the size of the retrieved terms is between six and thirteen and the input term is Chinese, the precision of retrieved Chinese terms is significantly (significant level of 10%) higher than the precision of the retrieved English terms. If the size of the retrieved term is between six and nine and the input term is English, the precision of the retrieved English term is significantly (significant level of 10%) higher the precision of the retrieved Chinese term. In addition, for the 50 randomly selected input terms, over 80% of them obtain the direct translation of the retrieved terms. The translation ability of the concept space generated by the Hopfield network is promising.

Figure 7 shows the recall of the retrieved concept space. As the number of concept space increases, the recall increases from 22% to 92%. Similar to the result of precision as shown in Figure 6, the recall of the opposite language of the input term is always higher than the recall of the same language.

Based on the results of the experiment, we find that the automatic cross-lingual English/Chinese thesaurus built by the Hopfield network has good translation ability. For most of the unknown terms that do not appear in the dictionary (such as Tung Chee Hwa, and Basic Law), it is able to retrieve the direct translation no matter whether the input term is English or Chinese. Besides, the capability of re-

trieving the relevant terms in the opposite language of the input term is always higher than retrieving the relevant terms in the same language of the input term.

5. Conclusion

Cross-lingual information retrieval systems are demanded as the multilingual information increases exponentially on the World Wide Web. Such demands are significant especially in an international city like Hong Kong, where both English and Chinese are official languages in the government, legislation, and many other business areas. Unfortunately, the amount of research work in cross-lingual information retrieval across the European and Oriental languages is significantly less than that across different European languages. The traditional dictionary-based approach has been widely used in cross-lingual information retrieval research; however, such approach limits the uses of vocabulary. Automatic construction of cross-lingual concept space relaxes such limitation. In this paper, we present the automatic construction of an English/Chinese thesaurus using the Hopfield network. The parallel corpus collected from the BLIS of the Department of Justice of the HKSAR Government is used as the test bed. The result of the experiment proves that the performance of automatic construction of a cross-lingual thesaurus by the Hopfield network is promising. Given the input term, if the number of retrieved terms is less than six, it can achieve up to 80% precision on average. Over 80% are able to retrieve the direct translation of the input term. It is also interesting to find that the precision and recall of the terms retrieved in the opposite language is always higher than the terms retrieved in the same language of the input term. Such an automatically generated cross-lingual thesaurus is a promising tool for cross-lingual information retrieval.

References

- English Will Dominate Web for Only Three More Years. *Computer Economics*, July 9, 1999, PRIVATE HREF="http://www.computereconomics.com/new4/pr/pr990610.html" MACROBUTTON HtmlResAnchor http://www.computereconomics.com/new4/pr/pr990610.html
- "Report: China Internet users double to 17 million," CNN.com, July, 2000, PRIVATE HREF="http://cnn.org/2000/TECH/computing/07/27/china.internet.reut/index.html" MACROBUTTON HtmlResAnchor http://cnn.org/2000/TECH/computing/07/27/china.internet.reut/index.html
- Greenspan R. (2002). China pulls ahead of Japan. *Internet.com*. April 22, 2002, http://cyberatlas.internet.com/big_picture/geographics/article/0,,5911_1013841,00.html
- Ballesteros L. & Crosft, B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. *Proceedings of the ACM SIGIR*, 1997, p. 84–91.
- Chen, H. (1998). Artificial intelligence techniques for emerging information systems applications: trailblazing path to semantic interoperability, *Journal of the American Society for Information Science*, 49(7), 579–581.
- Chien, L. F. (1997). PAT-tree-based keyword extraction for Chinese information retrieval. *Proceedings of ACM SIGIR*, (pp. 50–58). Philadelphia, PA, 1997.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Dai, Y., Khoo, C., & Loh, T. (1999). A new statistical formula for Chinese text segmentation incorporating contextual information. *Proceedings of the ACM SIGIR* (pp. 82–89). Berkeley, CA. August, 1999.
- Davis, M., & Dunning, T. (1995). A TREC evaluation of query translation methods for multi-lingual text retrieval. *Proceeding of the Fourth Text retrieval Conference (TREC-4)*, (pp. 175–185). NIST, November, 1995.
- Davis, M., & Dunning, T. (1995). Query translation using evolutionary programming for multi-lingual informaiton retrieval. *Proceedings of the Fourth Annual Conference on Evolutionary Programming*, San Diego, CA.
- Dumais, S.T., Letsche, T.A., Littman, M.L., & Landauer, T.K. (1997). Automatic cross-language retrieval using latent semantic indexing. *Proceedings of AAAI Symposium on Cross-Language Text and Speech Retrieval* (pp. 15–21). March, 1997.
- Fan, C.K., & Tsai, W.H. (1988). Automatic word identification in Chinese sentences by the relaxation technique. *Computer Processing of Chinese & Oriental Languages*, 4(1), 33–56.
- Fluhr, C. (1995). Multilingual information retrieval: survey of the state of the art in human language technology. Center for Spoken Language Understanding, Oregon Graduate Institute, PRIVATE HREF="http://www.cse.ogi.edu/CSLU/HLTsurvey/" MACROBUTTON HtmlResAnchor http://www.cse.ogi.edu/CSLU/HLTsurvey/ch8node7.html.
- Fluhr, C., & Radwan, K. (1993). Fulltext database as lexical semantic knowledge for multilingual interrogation and machine translation. *Proceedings of the East-West Conference on Artificial Intelligence* (pp. 124–128). Moscow, September, 1993.
- Gan, K.W., Palmer, M., & Lua, K.T. (1996). A statistically emergent approach for language processing: application to modeling context effects in ambiguous Chinese word boundary perception. *Computational Linguistics*, p. 531–553.
- Haykin, S. (1994). *Neural networks: a comprehensive foundation*. IEEE Press, New York.
- 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.
- Hull, D.A., & Grefenstette, G. 1996. Querying across languages: a dictionary-based approach to multilingual information retrieval. *Proceedings of the ACM SIGIR*, (pp. 49–57).
- Ikehara, S., Shirai, T.S., & Kawaoka, T. (1995). Automatic extraction of uninterrupted and interrupted collocations from very large Japanese corpora using N-gram statistics. *Transactions of the Information Processing Society of Japan*, 36(11), November, 1995, 2584–2596.
- Klavans, J., Hovy, E., Fluhr, C., Frederking, R.E., Oard, D., Okumura, A., Ishikawa, K., and Satoh, K. (1999). Multilingual (or cross-lingual information retrieval). *Multilingual information management: current levels and future abilities*, Pisa, Italy.
- Landauer, T.K., & Littman, M.L. (1990). Fully automatic cross-language document retrieval using latent semantic indexing. *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, (pp. 31–38). Waterloo Ontario, October, 1990.
- Leonardi, V. (2000). Equivalence in translation: between myth and reality. *Translation Journal*, 4(4).
- Lin, C., & 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*, 16(1), 1–14.
- Lua, K.T. (1990). From character to word—an application of information theory. *Computer Processing of Chinese & Oriental Languages*, 4(4), 304–313.
- Leung, C.H., & Kan, W.K. (1996). A statistical learning approach to improving the accuracy of Chinese word segmentation. *Literary and Linguistic Computing*, (11), 87–92.
- Nie, J.Y., Jin, W., & Hannaan, M.L. (1994). A hybrid approach to unknown word detection and segmentation of Chinese. *Proceedings of International Conference on Chinese Computing*, (pp. 326–335). Singapore, 1994.
- Nie, J.Y., Simard, M., Isabelle, P., & Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel text from the web. *Proceedings of the ACM SIGIR* (pp. 74–81). Berkeley, CA, 1999.
- Oard, D.W., & Dorr, B.J. (1996). A survey of multilingual text retrieval. UMIACS-TR96-19 C-TR-3815.
- Oard, D.W. (1997). Alternative approaches for cross-language text retrieval. *Proceedings of the 1997 AAAI Symposium in Cross-Language Text and Speech Retrieval* (pp. 154–162). March, 1997.
- Resnik, P. (1998). Parallel STRANDS: a preliminary investigation into mining the web for bilingual text. *Proceedings of the Third Conference of the Association for Machine Translation in the America: Machine Translation and the Information Soup*, (pp. 72–82). Langhorne, PA, October, 1998.
- Resnik, P. (1999). Mining the web for bilingual text. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, (pp. 527–534), College Park, Maryland, June, 1999.
- Rose, M.G. (1981). *Translation types and conventions. Translation spectrum: essays in theory and practice*. State University of New York Press, p. 31–33, Albany, New York.
- Radwan, K., & Fluhr, C. (1995). Textual database lexicon used as a filter to resolve semantic ambiguity application on multilingual information retrieval. *Proceedings of Fourth Annual Symposium on Document Analysis and Information Retrieval* (pp. 121–136). April, 1995.
- Salton, G. (1970). Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21(3), 187–194.
- Schatz, B., & Chen, H. (1999). Digital libraries: technological advances and social impacts. *IEEE Computer, Special Issue on Digital Libraries*, 32(2), 45–50.
- Sproat, R., & Shih, C. (1990). A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4, 336–351.
- Wu, Z, & Tseng, G. (1995). ACTS: An automatic chinese text segmentation system for full text retrieval. *Journal of the American Society for Information Science*, 46, 83–96.
- Chen, H., Yen, J., & Yang, C.C. 1999. International activities: development of Asian digital libraries. *IEEE Computer, Special Issue on Digital Libraries*, 32(2), 48–49.
- Yang, C.C., Luk, J., & Yung, S. (2000a). Combination and boundary detection approach for chinese indexing. *Journal of the American Society for Information Science, Special Issue on Digital Libraries*, 51(4), 340–351.
- Zanettin, F. (1998). Bilingual comparable corpora and the training of translators. *META, Special Issue on The Corpus-Based Approach: A New Paradigm in Translation Studies*, 43(4), 616–630.