Conceptual Analysis of Parallel Corpus Collected From the Web

Kar Wing Li and Christopher C. Yang
Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong, People's Republic of China. E-mail: yang@se.cuhk.edu.hk

As illustrated by the World Wide Web, the volume of information in languages other than English has grown significantly in recent years. This highlights the importance of multilingual corpora. Much effort has been devoted to the compilation of multilingual corpora for the purpose of cross-lingual information retrieval and machine translation. Existing parallel corpora mostly involve European languages, such as English–French and English–Spanish. There is still a lack of parallel corpora between European languages and Asian languages. In the authors' previous work, an alignment method to identify one-to-one Chinese and English title pairs was developed to construct an English–Chinese parallel corpus that works automatically from the World Wide Web, and a 100% precision and 87% recall were obtained. Careful analysis of these results has helped the authors to understand how the alignment method can be improved. A conceptual analysis was conducted, which includes the analysis of conceptual equivalent and conceptual information alternation in the aligned and nonaligned English–Chinese title pairs that are obtained by the alignment method. The result of the analysis not only reflects the characteristics of parallel corpora, but also gives insight into the strengths and weaknesses of the alignment method. In particular, conceptual alternation, such as omission and addition, is found to have a significant impact on the performance of the alignment method.

Introduction

The amount of information that is available in languages other than English on the World Wide Web is increasing significantly, and the demand for searches across language boundaries is growing rapidly. This highlights the importance of multilingual corpora. A corpus is one of the linguistic resources of natural language processing applications and information retrieval, the term *linguistic resources* referring to (usually large) sets of language data and descriptions in machine-readable form that are used for the construction, improvement, or evaluation of natural language and speech algorithms or systems (Godfrey & Zampolli, 1995).

In multilingual applications, dictionary-based approaches are normally applied. However, a general-purpose dictionary is less sensitive in the areas of genre and domain. To manually construct tailor-made bilingual dictionaries or sophisticated multilingual thesauri for large applications is impractical and time consuming. The corpus-based approach can overcome the limitations of dictionaries because it provides a statistical translation model to assist in the crossing of the language boundary. This approach puts the emphasis on statistical analysis rather than linguistic theory; it can be viewed as an automatic thesaurus construction technique. The association of terms is calculated from statistics on word usage across documents (Oard & Dorr, 1996). The advantage of processing a text corpus is that it is possible to obtain context-specific information about syntactic structures and the usage of words in a given language. In the case of parallel corpora, one can obtain context-specific correlations between languages that are usually much less ambiguous than in comparable collections. The resulting data from these corpus analysis processes can be used to develop context-specific tools for translation, and to standardize the usage of structures and word sets for the production of multilingual documents.

Much effort in the past was devoted to the compilation of multilingual corpora for the purpose of cross-lingual information retrieval (Brown et al., 1990; Brown, Lai, & Mercer, 1991) and machine translation (Davis & Dunning, 1995; Nie, Simard, Isabelle, & Durand, 1999; Sheridan & Ballerini, 1996). These corpora were primarily parallel corpora between European languages, such as English–French and English–Spanish. The lack of an Asian–European corpus, and of an English–Chinese corpus in particular, has brought many research projects to a halt. As English and Chinese are the most popular languages in the world, the need for an automatic Chinese–English parallel corpus construction approach is urgent.

Different types of corpora are available for multilingual information retrieval depending on the criteria according to

Accepted March 9, 2005

 ● Published online 1 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20326

which they are designed and the purpose for which they are created. A multilingual corpus that is developed based on parallelism is known as a parallel corpus (Yang & Li, 2003). Several techniques have been developed in the past for the automatic construction of parallel corpora. One prominent system that generates parallel corpora from the World Wide Web is the structural translation recognition for acquiring natural data (STRAND) system, which was developed by Resnik (1998). Another example is the PTMiner, which was developed by Nie (Chen & Nie, 2000; Nie and Cai, 2001; Nie et al., 1999).

As Hong Kong is a bilingual community, many Web sites that are hosted in the territory provide information in both English and Chinese. The Web site of the Hong Kong SAR Government is an example of such a site. This site publishes many government, legal, and financial documents on the Web for public access. These documents are written in both Chinese and English based on covert translation. The bilingual documents are organized on the Web site using the monolingual subtree structure (Figure 1), in which there are no direct links between the English and Chinese documents.

An alignment method to identify the one-to-one Chinese and English title pairs was developed for the automatic construction of English–Chinese parallel corpora that are collected from Web sites with a monolingual subtree structure (Yang & Li, 2003). The method includes alignment at title

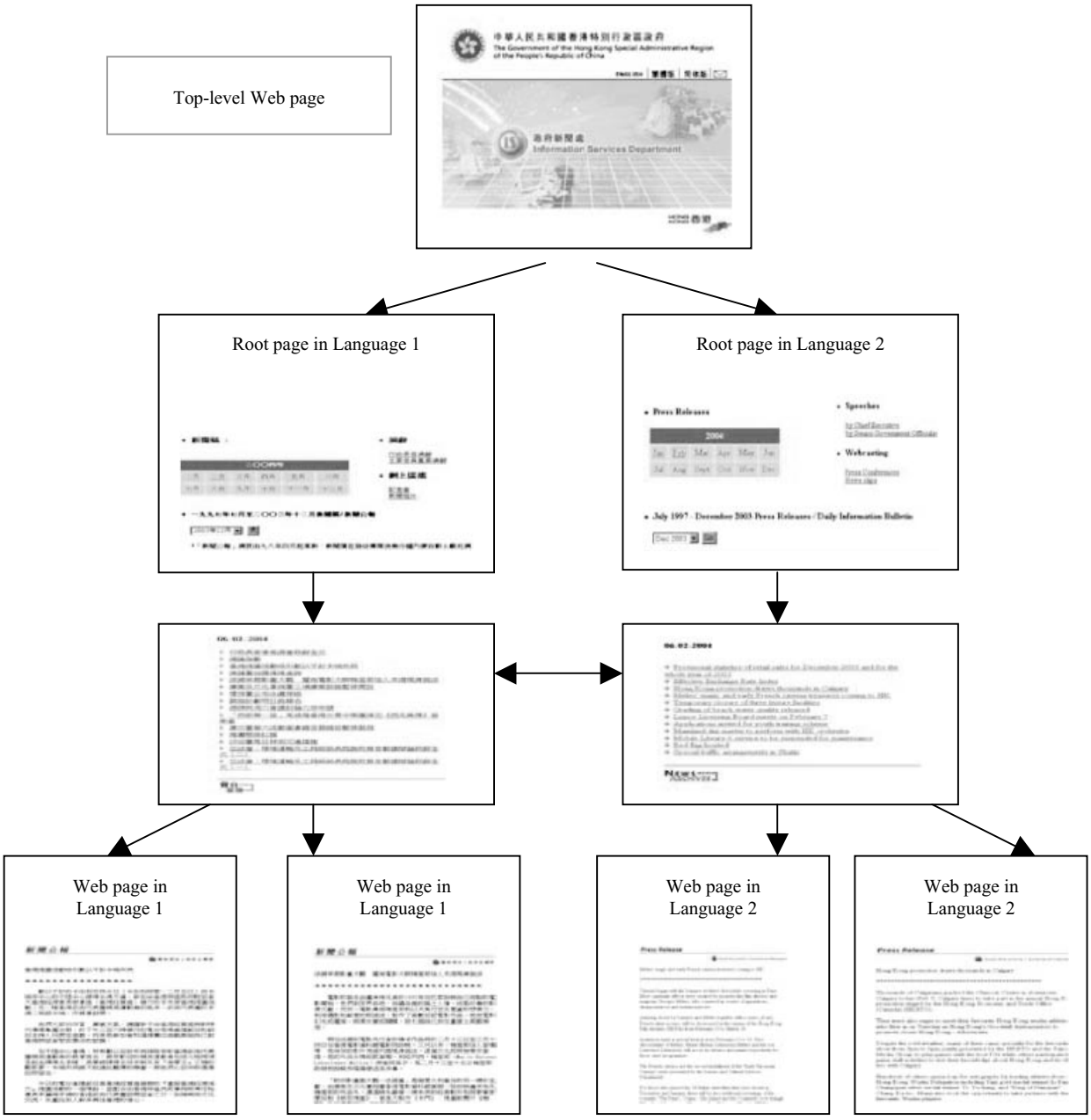

FIG. 1. The organization of the press releases on the Web site of the Hong Kong government is an example of a monolingual subtree Web site structure.

DOI: 10.1002/asi

level, word level, and character level. The longest common subsequence (LCS) is applied to find the most reliable Chinese translation of an English word. Deletion, an edit operation, is used to resolve redundancy, as one word in one of the languages may be translated into two or more words in the other language. A score function was then developed to determine the optimal title pairs.

To evaluate the performance and investigate the characteristics of this automatic generation technique for Chinese–English parallel corpora, we focus on the examination of conceptual equivalent and conceptual information alteration in the Chinese–English parallel corpus from the bidirectional point of view (from English to Chinese and Chinese to English). To justify the reliability between judges, several calculations, such as Cohen's kappa, are applied to measure the reliability between different judges. The analysis between the conceptual pairs shows the corpus characteristics, and also reveals the strength and weakness of the alignment model.

Concept Equivalence and Conceptual Alteration

For a given language, a concept may not only be represented by a word or words but also by a morpheme, by an idiomatic expression, by a tone, or by word order (He, 2000). In translation, several concepts may be represented by a single word in one language, yet may be translated into one word, two words, a phrase, or even a sentence in another language. This is called *concept equivalence*, which is a common occurrence in translation between languages.

In addition to concept equivalence, conceptual alteration may occur in translation. Sager (He, 2000) has listed five major types of conceptual alteration: modification of meaning, inversion of meaning, omission of meaning, addition of meaning, and deviation from meaning.

One of the common conceptual alterations that occurs in translation is known as the *generic-specific* relationship (Larson, 1998). A concept in one language can be a broad concept that encompasses some narrower concepts, and the translation of such a concept may result in an altered concept in another language. In contrast, a concept that is narrow in one language may be translated as a broader concept in another language. Nida (He, 2000) explains that conceptual alteration arises for three major reasons. The first is that no two languages are completely isomorphic, the second is that different languages may have different domain vocabularies, and the third is that some languages are more rhetorical than other languages.

Some researchers have studied translation as information transfer through the English-to-Chinese and Chinese-to-English translation of journal article titles (He, 1998). However, few people have examined the concept equivalence and conceptual information transfer between Chinese and English taking concept as a recognizable unit of meaning in any given language (He, 2000).

Conceptual analysis is important in the effective alignment of two titles. In this work, we analyze the alignment of the titles of the press releases that are published by the Hong Kong government to construct an English–Chinese corpus. According to Sidiropoulou (1995), the titles of news articles in the press are rarely translated intact into the target language, and there are usually some partial or total alterations. This is probably due to the social, cultural, and cognitive constraints on the organizational properties of media messages in different languages, because there is a systematic relationship between news text and context. The type of alteration that is adopted is expected to have a consistent pattern.

Automatic Construction of Parallel Corpora

For any given 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 et al., 1992). Empty segments are allowed, which correspond to the translator's omissions or additions. In other words, alignment is the process of finding relations between a pair of parallel documents. An alignment may also constitute the basis of a deeper automatic analysis of a translation, for example, it can be used to flag possible omissions in a translation, or to signal common translation mistakes such as terminological inconsistencies.

A textual alignment usually signifies a representation of two texts that are mutual translations in such a way that the reader can easily see how certain segments in the two languages correspond with each other (Macklovitch & Hannan, 1996). He (2000, p. 1048) states that the titles of documents present "micro-summaries of texts" that contain "the most important focal information in the whole representation" and are "the most concise statement of the content of a document." Titles may also be used as signals that help the reader to make effective guesses about the most important information in the text (Sidiropoulou, 1995). The reader may also infer the title depending on his beliefs and attitudes (van Dijk, 1985). Titles significantly influence the comprehension and memory of texts because they serve as advance organizers for the information that follows (Sidiropoulou, 1995). To take advantage of titles, the automatic corpus construction method relies on one-to-one Chinese and English title pairs.

Title Alignment Model

As the automatic corpus construction approach includes alignment at title level, word level, and character level, the analysis of translation characteristics begins at title level, and proceeds through word level to character level.

A title can be treated as a short sentence that consists of lexical items (words). Lexical differences in title alignment can be created by the representation of concepts in different languages. A lexical item (word) in a title can be a concept in one language (Larson, 1998), which is a recognizable unit of meaning in any given language (He, 2000).

To overcome conceptual alterations in title alignment, sequence comparison methods are adopted that rely on dynamic programming (Simard, 1999). In the title alignment

DOI: 10.1002/asi

model (Yang & Li, 2003), a title is viewed as a sequence of words, and a word can be treated as a sequence of characters. The longest common subsequence (LCS), which is employed in sequence comparison methods, is utilized to optimize the alignment of English and Chinese titles. The LCS is commonly exploited to maximize the number of matches between characters in two sequences. The title alignment algorithm has three major steps: alignment at word level and character level, the resolution of redundancy, and score function.

Alignment at Word Level and Character Level

An English title, E, is formed by a sequence of English simple words, i.e., $E = e_1\ e_2\ e_3 \ldots e_i \ldots$, where e_i is the i^{th} English word in E. A Chinese title, C, is formed by a sequence of Chinese characters, i.e., $C = char_1\ char_2\ char_3 \ldots char_q \ldots$, where $char_q$ is the q^{th} Chinese character in C. An English word in E, e_i, can be translated to a set of possible Chinese translations, $Translated(e_i)$, by dictionary lookup. $Translated(e_i) = \{T^1_{e_i}, T^2_{e_i}, T^3_{e_i}, \ldots, T^j_{e_i}, \ldots\}$, where $T^j_{e_i}$ is the j^{th} Chinese translation of e_i. A sequence of Chinese characters forms each Chinese translation. The set of the LCSs 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 LCSs 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, then $MatchList(e_i) = \varnothing$ and no reliable translation of e_i can be found in C.

We assume that if the characters of the Chinese translation of an English word appear adjacently in a Chinese sentence, then the Chinese translation would be more reliable than translations in which the characters do not appear adjacently in the Chinese sentence. This hypothesis is thus applied to the algorithm. $Contiguous(e_i)$ is used to determine the most reliable translation based on adjacency.

$$\begin{aligned} Contiguous(e_i) = \{x \mid\ & x \in MatchList(e_i) \\ & \textit{and all the characters of x that} \\ & \textit{appear adjacently in } C\} \end{aligned} \tag{2}$$

The second criterion for the most reliable Chinese translation is the length of the translation. $Reliable(e_i)$ is used to identify the longest sequence in $Contiguous(e_i)$

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

Resolving Redundancy

Redundancy has the primary function of adding cohesion in a language but at the same time, it is a problem for alignment. Because of redundancy, the translations of an English word may overlap partially or completely in Chinese. To deal with redundancy, $Dele(x,y)$ is added as an edit operation to remove the $LCS(x,y)$ from x. *WaitList* is a list that saves all of the sequences that are obtained by removing the overlapping of the elements of $MatchList(e_i)$ and $Reliable(e_i)$.

$$\begin{aligned} WaitList = {} & DELE(WaitList, Reliable(e_i)) \\ & \cup DELE(MatchList(e_i) \backslash \{Reliable(e_i)\}, \\ & Reliable(e_i)), \end{aligned} \tag{4}$$

where $DELE(X, y) = \bigcup_{i=1}^{n} Dele(x_i, y)$ and x_i is the i^{th} element of X.

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. *Remain* is a sequence that is initialized as C, and $Reliable(e_i)$ is removed from *Remain* from the e_1 until the last English word.

$$Matching\_Ratio(E, C) = \frac{|C| - |\text{Re}main|}{|C|}. \tag{5}$$

For any given English title, the Chinese title that has the highest *Matching_Ratio* among all of the Chinese titles is considered to be the counterpart of the English title. If more than one Chinese title has the highest *Matching_Ratio* to the English title E, then the Chinese title with the lowest value of $| Matching\_Ratio(E,C) - Matching\_Ratio^*(E,C) |$ is considered to be the counterpart of E (see Figure 2).

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

where $R(e_i) = \begin{cases} 0 & \textit{if } \text{Re}liable(e_i) = \varepsilon \\ 1 & \textit{otherwise} \end{cases}$.

Measurement of Concepts

The comparison of languages is of great interest from both a theoretical and an applied perspective. It reveals what is general and what is language specific. Therefore, language comparison is important both for the understanding of language in general and for the study of the individual languages that are being compared. Comparative analysis has applications within lexicography, language teaching, and translation studies.

To investigate concepts in translation, we need to understand that translation consists of transferring the meaning of one language into another language. This is done by changing the form of one language to the form of a second language. Another way of looking at form and meaning is to think of them as surface structure (grammatical, lexical, phonological) and deep structure (semantic) (Larson, 1998).

DOI: 10.1002/asi

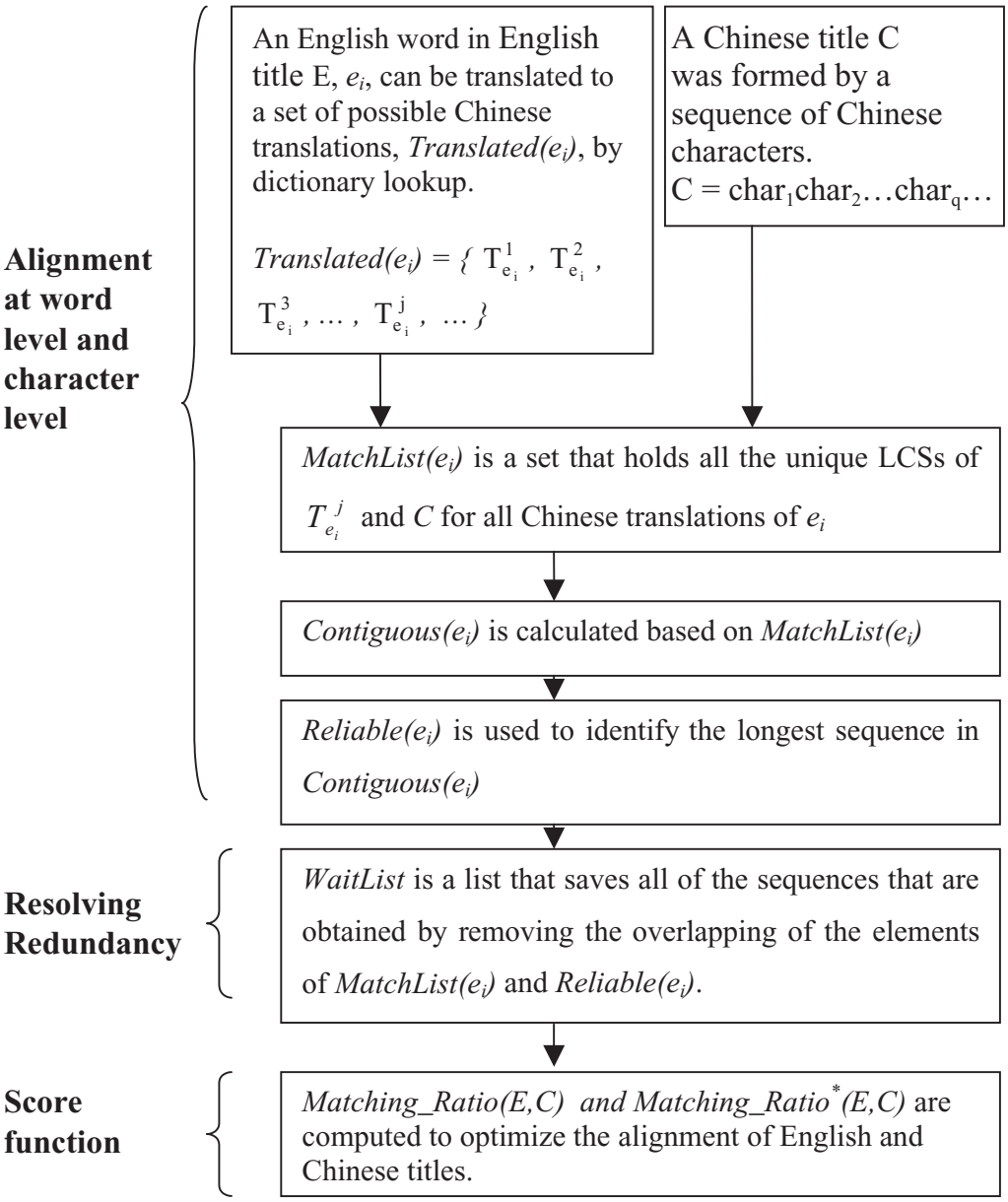

FIG. 2. The flowchart of the title alignment model.

Semantic propositions occur in all languages, and consist of concepts (groupings of meaning components) that are related to one another.

The smallest unit in semantic structure is a *meaning component*. Meaning components are grouped together to form concepts. Units are grouped into increasingly larger units in a hierarchy of semantic structures. The relationship between semantic configurations and grammatical structures is shown in Table 1 (Larson, 1998).

In the analysis of concepts, two issues must be considered (Larson, 1998):

- There will be concepts in the source language that are *known* (shared) in the target language.
- There will be concepts in the source language that are *unknown* in the target language.

TABLE 1. The relationship between semantic configurations and grammatical structures.

Semantic configurations	Grammatical structures
Meaning component	Morpheme (roots and affixes)
Concept	Word
Complex concept (concept cluster)	Phrase
Proposition	Clause
Propositional cluster	Sentence

Concepts that are known in the target language can be divided into synonyms, antonyms, and reciprocal lexical items.

1. The words "policeman" in English and 警察 in Chinese are *synonyms*.

DOI: 10.1002/asi

2. Conceptual alteration may also be found in the use of negating *antonyms*. For example, "no disruption" 沒有混亂 is translated as 正常 ("normal") in Chinese.
3. John offers me the hat and the hat is given by John. The relationship between the concepts "offers" and "is given" is *reciprocal*.

For concepts that are *known* in one language, conceptual analysis also involves *generic* or *specific* terms. For example, the generic concept "food" includes the specific word "bread."

For concepts that are *unknown* in one language, conceptual equivalents involve loan words and cultural substitutes.

1. A *loan word* is a word from another language and is unknown to most of the speakers of the target language. Loan words are commonly used for names of people, places, and geographical areas. For example, the word "Kowloon" (九龍), which is loaned from Chinese, is a place name in Hong Kong.
2. A *cultural substitute* is a word that does not exactly occur in the source language, but occurs in the target language. For example, the ballet "Cinderella's Slipper—Lost and Found" is unknown in Chinese. The cultural substitute is 仙履奇緣 ("a fairy").

In some cases, a single word is translated by several words, which together are known as a *descriptive phrase*. For example, the Chinese abbreviation 法援署 (法律援助署) is known as "Legal Aid Department" in English.

Directionality in Translation

To investigate concepts, we examine a Chinese–English parallel corpus that is aligned by a title-based algorithm. The relationship between a pair of bilingual documents in the corpus is known as a *covert translation* (Rose, 1981). Neither text in the pair is marked as translated text or source text. Therefore, the proposed measurement is bidirectional. The nature of bidirectional measurement is derived from Doorslaer (1995), who suggests that it involves "working simultaneously along the comparative line from source language to target language and target language to source language," and Toury (1986).

When Toury (1986) discusses the segmentation of text in the translation process, he observes, "in any transfer context, a language user activates two of the languages that he has at his disposal simultaneously." Toury states that although translators initially transfer from source language to target language in a unidirectional way, they then turn back to the source language utterance "to resegment a certain segment, or to go on segmenting the utterance." This reverse directionality may be viewed as a mapping of the target language elements onto the source text as part of the translation process.

In this analysis, we start the measurement from English to Chinese and then from Chinese to English.

English to Chinese. We treat English as the source language and Chinese as the target language (Table 2).

TABLE 2. An example measurement from English to Chinese.

Source concept	Target concept	Category for target concept, e.g., L, C, S, A, R, D
LCSD	康樂文化署	D (Descriptive phrase)
extends	擴展	S (Synonym)
use	使用	S (Synonym)
of		**Omission**
Octopus cards	八達通	C (Cultural Substitute)
to	至	S (Synonym)
Kowloon region	九龍區	L (Loan word)
sporting venues	運動場地	S (Synonym)

1. The judges first study the surface structure of a title in the source language, find the concepts of each title, and then write down each concept and its counterpart on the provided sheet. There are 55 Chinese–English title pairs.
2. The judges then classify the concepts and their counterparts into six categories: loan word (L), cultural substitute (C), synonym (S), antonym (A), reciprocal (R), and descriptive phrase (D). The relevant symbol is then marked on the sheet provided.

For example, "LCSD Extends Use of Octopus Cards to Kowloon Region Sporting Venues" is 康樂文化署擴展使用八達通至九龍區運動場地.

Chinese to English. As with the English to Chinese analysis, we treat Chinese as the source language and English as the target language (Table 3).

For example, 康樂文化署擴展使用八達通至九龍區運動場地 is "LCSD Extends Use of Octopus Cards to Kowloon Region Sporting Venues."

Judges and Reliability

The bidirectional measurement was carried out by two judges from the Department of Translation of the Chinese

TABLE 3. An example measurement from Chinese to English.

Source concept	Target concept	Category for target concept, e.g., L, C, S, A, R, D
康樂文化署	LCSD	D (Descriptive phrase)
擴展	extends	S (Synonym)
使用	use	S (Synonym)
	of	**addition**
八達通	Octopus Cards	C (Cultural substitute)
至	to	S (Synonym)
九龍區	Kowloon region	L (Loan word)
運動場地	sporting venues	S (Synonym)

DOI: 10.1002/asi

University of Hong Kong with a good knowledge of English and Chinese. One is a postgraduate research student, and the other is a research assistant. They both have more than 2 years translation and research experience in both Chinese to English translation and English to Chinese translation.

To ensure the reliability of the two judges, they must be independent. Independence in this case relates to the freedom of the judges to make autonomous judgments without input from other judges or the authors of the analysis. Any communication between the judges would invariably influence coding toward a higher agreement, and this lack of independence would be likely to make the data appear more reliable than they are (Krippendorff, 1980). When reliability matters, only the most capable individuals should be employed (Krippendorff, 1980).

Reliability ensures the quality of a measurement, and serves as an index to estimate the dependability of scores. Reliability can be considered as the ratio of the true level of the measure to the entire measure. The true level of a measure can be calculated using the variance of the true score, and the entire measure can be computed using the variance of the measure. In other words, reliability may be thought of as the proportion of truth in a measure (Trochim, 1999).

According to Krippendorff (1980), reliability is expressed as a function of the agreement that is achieved among raters regarding the assignment of units to categories. If agreement among raters is perfect for all units, then reliability is assured. In this study, interrater reliability is applied to evaluate the reliability of the judgments of the judges on conceptual alteration in the titles of the press releases. Two formulations, percentage of agreement and Cohen's kappa, are adopted.

Percentage of Agreement

Percentage of agreement is the simplest reliability indicator, and is the proportion of total agreement among the judges. The index can be written as

$$A = F_o/TOT,$$

where F_o is the number of agreements that the judges have made and *TOT* is the total number of judgments that the judges have made (Perreault & Leigh, 1989).

Cohen's Kappa

Cohen's kappa is the most widely used measure of interrater reliability across the behavioral science literature (Perreault & Leigh, 1989).

Cohen (Cohen, 1960) defined kappa (K) as

$$K = (P_o - P_c)/(1 - P_c),$$

where P_o is the proportion of times that the judges agree and P_c is the proportion of times that the judges are expected to agree by chance.

TABLE 4. The results of interrater reliability.

Interrater reliability method	Result
Percentage of agreement	0.98
Cohen's kappa	0.97

The percentage of agreement and Cohen's kappa of the measurement in this study are given in Table 4.

All of the interrater reliability methods show a high reliability between the two judges, which demonstrates the quality of the measurement.

Lexical Analysis

In this section, we report the result of the lexical analysis of the parallel corpus that was collected from the press releases on the Web site of the Hong Kong SAR government. A similar analysis was conducted by He (1998) on 100 medical article titles. He investigated the occurrence of concept similarity and conceptual information alteration in the translation of medical article titles between English and Chinese. In translation, though the study of large quantities of documents is often the ultimate goal, in most cases it is impossible to achieve (Doorslaer, 1995). When working with a corpus of a few thousand translated pages, the translator does not need to be exhaustive. In fact, the aim of completeness may even have a negative influence on the depth of analysis (Doorslaer, 1995).

In this study, we conduct an analysis of 220 English–Chinese Hong Kong government press release titles. Fifty-five English–Chinese title pairs were randomly selected from the 20,619 press release article title pairs that were aligned by the automatic parallel corpus construction system as described in a previous section. Another 55 title pairs, which could not be aligned by the automatic corpus construction approach, were also analyzed. These pairs were randomly selected from the 3,081 title pairs that could not be aligned by the automatic corpus construction approach. The purpose is to investigate whether there is any significant lexical difference between the title pairs that can and cannot be aligned by our automatic parallel corpus construction system.

English–Chinese Title Pairs Aligned by the Automatic System

In the 55 English–Chinese title pairs that were aligned by the automatic parallel corpus construction system, 1,032 Chinese characters and 487 English words were found. Two hundred fifty-three concept pairs were identified by the judges (Table 5).

TABLE 5. The number of Chinese characters and English words in the 55 aligned title pairs.

55 Aligned title pairs (110 titles)	
Number of concept pairs identified	253
Number of Chinese characters	1032
Number of English words	487

DOI: 10.1002/asi

Many concept pairs feature in our dictionary, and include the following.

1. Synonym. For example, "injection" is translated as 注射劑 in the Chinese title. "Traffic accident" is known as 交通事故. "Professionalism" is translated as 專業水平, and "transparency" as 透明度. "Home-schooling" is known as 家庭教學, and "Housing Department" is translated as 房屋署 in Hong Kong.
2. Loan. For example, "Victoria Harbour" (a harbor in Hong Kong) is translated as 維多利亞港 in Chinese. 旺角, which is a loan word from Chinese, is translated as "Mongkok" in English. 的士 (taxi) is also a loan word from English.
3. Cultural substitute. The category of concepts usually refers to the names of people or places. For example, Martin Lee is a member of the Legislative Council of Hong Kong. His Chinese name is 李柱銘. "Admiralty" is a place in Hong Kong, and is known as 金鐘 (golden bell) in Hong Kong.
4. Reciprocal. For example, "seizure" is translated as 被檢獲 (was seized).
5. Descriptive phrase. The English abbreviation "CFA" (Court of Final Appeal) is translated as 終審法院. The Chinese abbreviation 法援署 (法律援助署) is known as "Legal Aid Department" in English.

Table 6 shows the number of concept pairs in different conceptual categories among the 253 concept pairs.

Among the 253 concept pairs, some concepts and parts of concepts are not found in our dictionary and generate difficulty in the alignment process. Four loan words, two cultural substitutes, and two English abbreviations are not included in our dictionary (Table 7).

TABLE 6. The concept pairs of different categories in the 55 title pairs.

Concept category	Number of concept pairs in 55 title pairs (110 titles)
Synonym	196
Antonym	0
Loan words	19
Cultural Substitute	5
Reciprocal	10
Descriptive phrase	23
Total number of concept pairs that can be maintained from one language to the other language	253

TABLE 7. Number of concept pairs not included in our dictionary.

Concept category	# Concept pairs not in dictionary	Total # concept pairs in each category
Synonym	0	196
Antonym	0	0
Loan words	4	19
Cultural substitute	2	5
Reciprocal	0	10
Descriptive phrase	2	23

One of the two cultural substitutes that is not included in our dictionary is "Queensway Plaza" (金鐘廊), which is a shopping center in Hong Kong, and another is Martin Lee (李柱銘).

The four loan words are formed by nine Chinese characters and six English words. These four loan words are all names, including "Cox" (考克斯), "Taikoo Shing" (太古城, an estate in Hong Kong), "Pat Heung" (八鄉, a place in Hong Kong) and "Daya" (大亞, a place in China). "Daya" is part of the loan word "Daya Bay" (大亞灣), which is a bay in Guangdong province. "Cox" refers to "the Cox Report" (考克斯報告). The Cox Report, which was released by a special committee of the House of Representatives of the United States of America in 1999, made accusations that China had stolen confidential information on military technology from the United States, in which the Hong Kong Special Administrative Region (HKSAR) was also implicated. Among the 19 loan words that were identified, 10 of them originate from English and are unknown in Chinese, and 9 of them originate from Chinese and are unknown in English (as is shown in Table 8). There is no significant difference between the number of loan words that originate from English or from Chinese.

The two English abbreviations that are not included in our dictionary are "HAD" (Home Affairs Department, 民政事務局) and "PRH" (public rental housing, 公共租住屋). The Chinese abbreviation of 公共租住屋 is 公屋 (public housing). In the developed algorithm, an English descriptive phrase is easily aligned with its Chinese abbreviation using the longest common subsequence (LCS) and adjacency if the English descriptive phrase is included in our dictionary. Of the 23 English abbreviations that were identified, 20 are English abbreviations and 3 are Chinese abbreviations (as is shown in Table 9). However, 18 out of the 20 English abbreviations are included in the dictionary and can therefore be handled by our algorithm.

As the usage of language is a creative activity, new terms enter the lexicon all the time. The effort that is involved in creating up-to-date and subject-specific dictionaries is often overwhelming, and a dictionary may quickly become obsolete.

TABLE 8. The total number of loan words, e.g., geographical name, company name.

Type of loan word	Number of loan word
Loan words unknown in Chinese	10
Loan words unknown in English	9

TABLE 9. The number of descriptive phrases.

Type of descriptive phrase	# Descriptive phrases
English abbreviations translated as Chinese descriptive phrases	20
Chinese abbreviations translated as English descriptive phrases	3

DOI: 10.1002/asi

As a result, certain loan words and cultural substitutes may not appear in our dictionary. To improve the performance of the system, our future work will cover the utilization of the automatically collected parallel corpus to extract the translations using a statistical method with which to update the bilingual dictionary.

Complex Concepts: Omission, Addition, and Redundancy

Simple concept pairs can be easily aligned, for example, "clarified" and 澄清. In the case of complex (multiword) concept pairs, the sequence of Chinese words in a complex concept may be different from the sequence of English words. For example, "small and medium enterprises" (SMEs) are known as 中小型企業 in Chinese. The Chinese translation of "medium" is 中型 and of "small" is 小型. Therefore, the sequence of English words in the concept "small and medium enterprises" is different from the sequence of the Chinese translation. In the developed algorithm, each English word is treated individually; therefore, it can easily align the complex concepts.

Omission, addition, and redundancy may occur within a complex concept. For example, the proper noun phrase 深港西部通道 is known as "the Shenzhen Western Corridor" (SWC) in English, and the character 深 is a Chinese abbreviation for "Shenzhen." However, there is no translation for 港 (the equivalent of "Hong Kong") in the English counterpart. In addition, in an example of reciprocity, the English concept "seizure" is translated as 被檢獲 (was seized). The Chinese word 被(to be) is omitted in the English concept.

Conceptual alteration (omission and addition) also occurs outside the scope of concept pairs. For instance, the Chinese counterpart of the title "Specific Fees Under Builders' Lifts and Tower Working Platforms (Safety) Ordinance to Be Revised" is 政府調整建築工地升降機及塔式工作平台(安全) 條例下特定收費. The Chinese concept 政府 (government) is omitted in the English title. In our algorithm, such Chinese addition words stay in the *Remain* list because they cannot be matched by any English words. The *Matching_Ratio*(*E*,*C*) will thus be lowered.

Among the 55 title pairs, our analysis shows that 13 Chinese simple concepts and 3 Chinese complex concepts were added, and 22 Chinese characters were thus added to express these concepts. Of these 22 characters, 6 Chinese characters belonged to the 3 complex concepts and 16 Chinese characters belonged to the 13 simple concepts. Table 10 illustrates the data.

TABLE 10. The number of additions of Chinese concepts in the 55 title pairs.

Conceptual alteration	# Concepts	# Characters
Addition of Chinese simple concepts	13	16
Addition of Chinese complex concepts	3	6

TABLE 11. The number of omissions of English words in the 55 title pairs.

Conceptual alteration	# Concepts
Omission of English simple concepts	6 (one concept is "and")
Omission of English prepositions within complex concepts	8
Omission of English words other than prepositions within complex concepts, e.g., "and"	10

Apart from the Chinese words that were omitted in the English titles, some English words were also omitted in the Chinese complex concepts, especially English prepositions. For example, the Chinese translation of "the University of Hong Kong" is 香港大學. There is no Chinese translation of either the English preposition "the" or "of" in this case. Apart from the prepositions, only a few of the English words do not align with their Chinese counterparts. For example, the title "Hong Kong Monetary Authority Hong Kong Mortgage Corporation Note Issuance Programme Tender Results" is translated as 香港金融管理局香港按揭證券有限公司債券發行投標結果. The concept of "programme" (計劃) is not translated in the Chinese title.

We observe that 24 English words do not have any translation in the 55 Chinese titles. Among these 24 English words, 19 of them are English stop words such as "and," "the," "by," "of," "on." Table 11 shows the distributions of the English words that were omitted. The omission in the Chinese translations only affects the value of *Matching_Ratio**(*E*,*C*), and because the developed algorithm relies heavily on *Matching_Ratio*(*E*,*C*) to determine the optimal title pairs, the omission of English words in the Chinese translation does not significantly affect the performance of the developed algorithm. In addition, the English stop words are removed using a stop word list before alignment is carried out.

Redundancy also occurs in the English abbreviations that are represented by descriptive phrases in Chinese. For example, the English abbreviation for "Central and Western District Council" in Hong Kong is "C&WDC", and its Chinese translation is 中西區區議會. 中西區 stands for "Central and Western District" and 區議會 is translated as "District Council," and hence the second instance of 區 (district) is redundant. However, this redundancy can be solved using the DELE function in the alignment approach.

English–Chinese Title Pairs That Could Not Be Aligned by the System

In the 55 title pairs (nonaligned pairs) that could not be aligned by the automatic corpus construction approach, there were 768 Chinese characters and 389 English words. Two hundred twenty six concept pairs were identified (Table 12).

There are fewer Chinese characters, English words, and concept pairs in the nonaligned pairs than there are in the aligned title pairs. This indicates not just that there are fewer concept pairs in the 55 nonaligned title pairs, but also that

DOI: 10.1002/asi

TABLE 12. The number of Chinese characters and English words in the 55 nonaligned title pairs.

55 Non-aligned title pairs (110 titles)	
# Concept pairs	226
# Chinese characters	768
# English words	389

the length of the nonaligned titles is relatively shorter. Fewer concept pairs may mean that there is not enough information to perform correct alignment.

Table 13 shows the 226 concept pairs divided into different conceptual categories.

Among the 226 concept pairs, 37 concept pairs are not included in our dictionary, which explains why the title pairs that contain these concept pairs cannot be aligned by the developed approach (Table 14). These concepts include the following.

1. Three synonyms. "Over-stayer" (逾期居留人士), "revelers" (醉酒人士), and "ampoule" (注射瓶).
2. Fourteen loan words. The loan words can be divided into the place names, names of events, and names for specific terms (Table 16). For example, "Whampoa" (黃埔) is part of "Whampoa Garden," an estate in Hong Kong. "Che Kung Festival" is known as 車公誕 in Hong Kong and "calcium gluconate" is known as 葡萄糖酸鈣.
3. Ten cultural substitutes. The concepts "psychedelic" (迷幻藥) and "psychotropic" (as in psychotropic drug, 精神科藥物) are not included in our dictionary. The company name "PowerPhone Network" is known as 面面通 (face-to-face communication) in Chinese. The company name "Jardine Matheson" is translated as 怡和. The concept "more initiatives" is translated as an idiom 集思廣益 (collecting more initiatives is better).
4. Ten descriptive phrases. For example, "UBW" (unauthorized building works) is known as 違例僭建物 in Chinese. The concept "TPS" (Tenants Purchase Scheme, 租置計劃) is also not found in our dictionary.

TABLE 13. The number of concept pairs of different categories in the 55 non-aligned title pairs.

Concept category	Number of concept pairs in 55 nonaligned title pairs (110 titles)
Synonym	184
Antonym	0
Loan word	14
Cultural substitute	10
Reciprocal	8
Descriptive phrase	10
Total number of concept pairs that can be maintained from one language to the other language	226

TABLE 14. Number of concept pairs that are not included in our dictionary.

Concept category	# Concept pairs not in dictionary	Total # concept pairs in each category
Synonym	3	184
Antonym	0	0
Loan words	14	14
Cultural Substitute	10	10
Reciprocal	0	8
Descriptive phrase	10	10

TABLE 15. The number of descriptive phrases.

Type of descriptive phrase	# Descriptive phrases
English abbreviations that are translated as Chinese descriptive phrases	10
Chinese abbreviations that are translated as English descriptive phrases	0

TABLE 16. The total number of loan words, e.g., geographical name, company name.

Type of loan word	# loan word
Loan words unknown to Chinese	6
Loan words unknown to English	8

As is illustrated in Table 15, all of the abbreviations are English abbreviations for the names of organizations and other specific terms. The developed algorithm starts the alignment from English to Chinese. Thus, the English abbreviations generate significant problems for the alignment process if they are not included in our dictionary. None of these 10 cases can be found in the dictionary.

The effect of out-of-vocabulary concepts is unlikely to be eliminated because language is a creative activity and new terms enter the lexicon all the time. None of the loan words and cultural substitutes that are shown in Table 14 is found in the dictionary, and the high number of out-of-vocabulary concepts generates a significant limitation to the automatic alignment process, because the number of characters in the *Remain* list will be high, and thus the value of *Matching_Ratio*(*E*,*C*) will be low, which creates uncertainty in the automatic alignment.

The conceptual alterations of omission and addition occur significantly often in the nonaligned title pairs. For example, the counterpart of the English title "Consultation Paper on Local Completed Residential Properties Released" is 法改會發表諮詢文件. There is no translation for 法改會 (Law Reform Commission) in the English title, and there is no translation for the long phrase "Local Completed Residential Properties" in Chinese (本地已建成住宅物業), although the translation can be found within the text of the press release itself. This generates difficulty for the automatic alignment (Table 17).

As with the analysis of the aligned title pairs, omission and addition also occur within the two complex concept pairs. The English translation for 文化康樂及社會事務委

DOI: 10.1002/asi

TABLE 17. The number of addition of Chinese concepts in the 55 nonaligned title pairs.

Conceptual alteration	# Concepts
Addition of Chinese concepts	27
Addition of Chinese characters within complex concepts	2

員會 is "Cultural and Leisure Services Committee," and there is no translation of 社會 (society) in the English concept. "The Agriculture, Fisheries and Conservation Department" (AFCD) is known as 漁農自然護理署, but there is no translation of 自然 (nature) in the English concept, and the word "and" is not translated in the Chinese counterpart. However, the developed algorithm, which relies on the longest common subsequence, can deal with addition and omission within the complex concepts effectively.

Our analysis shows that 56 Chinese characters (or 29 Chinese words) were added in the Chinese titles, including 2 words (or 4 Chinese characters) that were added in the two complex concepts. As the algorithm relies on *Matching_ratio*(*E*,*C*) for alignment, the number of Chinese characters remaining in the *Remain* list is significantly high, which therefore lowers the value of *Matching_ratio*(*E*,*C*) significantly.

In addition, 36 English words have no Chinese counterpart in the Chinese titles as shown in Table 18. Among the 36 English words, 20 are English stop words. Compared with the 19 stop words out of the 24 omitted English words in the aligned titles (as shown in Table 11), there are significantly more omitted English words that are not stop words in the nonaligned titles.

The analysis reveals the weakness of using a dictionary. The dictionary only provides a definite set of conceptual equivalents for a given word, and other words that closely relate to the given term are not presented. For example, the word "student" is strongly related to the word "school," but they are not conceptually equivalent. However, different languages use different vocabulary to express the same ideas, and historical information sometimes lies behind a concept that is new in other languages (Table 17). This is called *implicit information*. During the translation process, the translator may translate the implicit information in one language to explicit information in another language (He, 2000; Larson, 1998).

TABLE 18. The number of omissions of English words in the 55 title pairs.

Conceptual alteration	# Concepts
Omission of English simple concepts	13
Omission of English prepositions within complex concepts	12
Omission of English words other than preposition within complex concept e.g., "and"	11 (8 words are stop words)

In the case of this study, the words of some English titles were not conceptually equivalent to the words in their Chinese counterpart titles, and thus there are some titles that cannot be aligned by the title-based alignment approach. For example, the Chinese article 盧偉聰情繫國際刑警心懸香港 is the counterpart of the English article "From Bak Choy to Baguette . . ." According to our dictionary, the word "baguette" denotes a type of French bread, and the phrase "bak choy" (白菜) refers to a type of Chinese vegetable. The English title implicitly tells of someone's life in Hong Kong and France, even though there is no conceptual equivalent between the Chinese and English titles. The content of the two articles is generated by covert translation.

Comparison of Two Sets of Title Pairs

To compare the characteristics of the title pairs that are aligned with the title pairs that cannot be aligned (nonaligned pairs), we start by examining the number of concept pairs that are not included in our dictionary (Figure 3).

Even though there is no significant difference between the number of concept pairs in either the aligned title pairs or the nonaligned title pairs, the number of concept pairs that are included in our dictionary is significantly different. Only 189 out of the 226 concept pairs in the 55 nonaligned title pairs are included in our dictionary. However, 245 out of the 253 concept pairs in the 55 aligned title pairs are included in our dictionary. As the developed approach is based on the dictionary to search for optimal title pairs, the significant number of out-of-vocabulary concepts can make the alignment result in error.

As has been previously mentioned, to improve the approach we propose to apply these title pairs to update the bilingual dictionary using a statistical method. After storing the non-aligned title pairs for 3 months and calculating the frequency that two words occur in one day using mutual information, some encouraging results were initially obtained.

As is shown in Figure 4, the conceptual alteration (addition and omission) in terms of the Chinese characters is more significant in the title pairs that cannot be aligned by the alignment approach. The Chinese additions generate problems for the alignment of the correct title pairs because the Chinese characters stay in the *Remain* list of

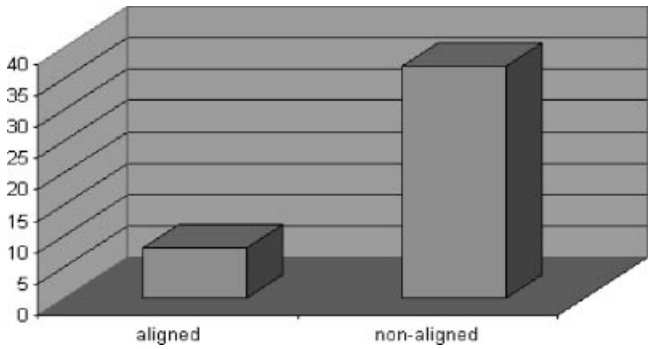

FIG. 3. The number of concept pairs that are not in our dictionary.

DOI: 10.1002/asi

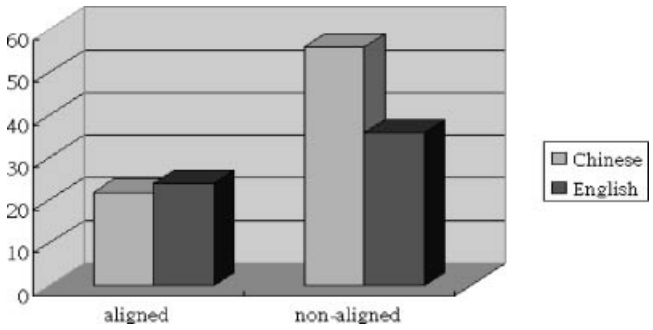

FIG. 4. The number of Chinese characters and English words added in both the aligned and nonaligned pairs.

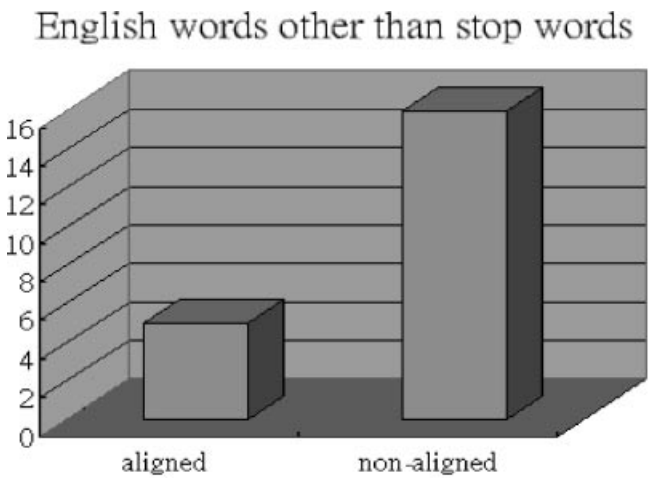

FIG. 5. The number of English omitted words other than stop words in the aligned and nonaligned title pairs.

Matching_ratio(*E*,*C*) and affect the alignment result. As the length of many of the nonaligned title pairs is relatively short and may not provide sufficient lexical information for alignment, the additions and omissions can significantly degrade the performance of the algorithm by providing false information. To ensure greater precision, a threshold can be set to guarantee that a certain number of the English words in an English title have reliable Chinese translations in a Chinese title. This eliminates the uncertain alignments, but the recall is also decreased.

Apart from the stop words, 5 English words had no Chinese counterpart in the aligned title pairs and 16 English words had no Chinese counterpart in the nonaligned title pairs (Figure 5). The English additions in the nonaligned title pairs significantly affect the value of *Matching_ratio**(*E*,*C*), and create uncertainty in the automatic alignment process.

Conclusion

We investigated the phenomena of conceptual equivalent and conceptual information alteration in English–Chinese title pairs that are aligned by the automatic corpus construction approach. This approach includes alignment at title level, word level, and character level. The longest common subsequence was applied to find the most reliable Chinese translation of an English word. As one word may be translated into two or more words, deletion, which is an edit operation, was used to resolve redundancy. In addition, a score function is proposed to find the optimal title pairs.

The analysis shows that the mechanisms of the method can effectively align title pairs from English to Chinese. However, the analysis also reveals the weakness of the method. It is hard to align short titles, because the number of concepts in a short title is relatively smaller than the number of concepts in a long title. Omission and addition are normal occurrences in parallel title pairs that generate uncertainty in the alignment process because they affect the value of *Matching_ratio*(*E*,*C*) and *Matching_ratio**(*E*,*C*). To ensure greater precision, a threshold can be set to ensure that a certain number of the words in an English title have reliable translations in a Chinese title before the alignment process starts. This eliminates the uncertain alignments, but the recall is decreased.

As the titles are aligned based on a dictionary, many technical terms fall outside the scope of the dictionary, and some abbreviations or slang words are also not included. The dictionary thus lacks some terms that are essential for correct alignment. However, as language usage is a creative activity, new dictionaries may quickly become obsolete, and the effort that would be involved in creating an up-to-date and domain-specific dictionary is overwhelming. This gives rise to the need to update the bilingual dictionary using a statistical method that involves the use of term co-occurrence statistics across large title collections.

The analysis also reveals other weaknesses that relate to the use of a dictionary. A dictionary only provides a definite set of conceptual equivalents for a given word, and other words that are closely related to the given term are not presented. However, different languages use different vocabulary to express the same ideas. To overcome the problem, we propose the creation of a concept space from the term co-occurrence statistics across a large title collection that can be used as a supplement to the dictionary. A concept space is a semantic network that consists of concepts (noun phrases in the textual domain) and related concepts, and is computed based on co-occurrence relationships.

Acknowledgments

This project was supported by the Direct Research Grant of the Chinese University of Hong Kong, 2050268, and the Earmarked Grant for Research from the Hong Kong Research Grant Council, 4335/02E. We would like to thank Ms. Jennifer Eagleton and Ms. Hui Tung Eos Cheng for their effort in the evaluation of the parallel corpus and many valuable comments.

References

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., et al. (1990). A statistical approach to machine translation. Computational Linguistics, 16(2), 79–85.

DOI: 10.1002/asi

Brown, P., Lai, J., & Mercer, R. (1991). Aligning sentences in parallel corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (pp. 169–176). Morristown, NJ: Association for Computational Linguistics.

Chen, J., & Nie, J.Y. (2000, May). Automatic construction of parallel Chinese-English corpus for cross-language information retrieval. Proceedings of the sixth conference on Applied Natural Language Processing, Seattle, WA (pp. 21–28). San Francisco: Morgan Kaufmann Publishers.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 27–46.

Davis, M., & Dunning, T. (1995, March). Query translation using evolutionary programming for multilingual information retrieval. In J.R. McDonnell, R.G. Reynolds, & D.B. Fogel (Eds.), Evolutionary Programming IV: Proceedings of the Fourth Annual Conference on Evolutionary Programming (pp. 175–185). Cambridge, MA: A Bradford Book, MIT Press.

Doorslaer, L.V. (1995). Quantitative and qualitative aspects of corpus selection in translation studies. Target, 7(2), 246–260.

Godfrey, J.J., & Zampolli, A. (1995). Language resources. In Ron Cole (Ed.), Survey of the state of the art in human language technology (pp. 441–474). Cambridge/New York: Cambridge University Press.

He, S. (1998). Concept similarity and conceptual information alteration via English-to-Chinese and Chinese-to-English translation of medical article titles. Journal of the American Society for Information Science, 49(2), 169–175.

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.

Kolbe, R.H. & Burnett, M.S. (1991). Content-analysis research: an examination of applications with directives for improving research reliability and objectivity. Journal of Consumer Research, 18(2), 243–250.

Krippendorff, K. (1980). Content analysis. Beverly Hills, CA: Sage.

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

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

Nida, E.A. (1985). Translating means translating meaning. In H. Buhler (Ed.), Translators and their position in society (pp. 119–125). Vienna: Wilhelm Braumuller.

Nie, J.Y., & Cai, J. (2001). Filtering noisy parallel corpora of web pages. In Proceedings of IEEE Symposium on NLP and Knowledge Engineering (pp. 453–458). Piscataway, NJ: IEEE.

Nie, J.Y., Simard, M., Isabelle, P., & Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web (ACM SIGIR'99). New York: ACM.

Oard, D.W., & Dorr, B.J. (1996). A survey of multilingual text retrieval (UMIACS-TR-96-19). College Park, MD: University of Maryland.

Perreault, W.D., & Leigh, L.E. (1989). Reliability of nominal data based on qualitative judgments. Journal of Marketing Research, 2, 135–148.

Resnik, P. (1998). Parallel strands: A preliminary investigation into mining the web for bilingual text. In D. Farwell, L. Gerber, & E. Hovy (Eds.), Machine translation and the information soup: Third Conference of the Association for Machine Translation in the Americas (AMTA-98). Lecture Notes in Artificial Intelligence 1529. Heidelberg/Berlin/New York: Springer Verlag.

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.

Sager, J.C. (1989). Quality and standards: The evaluation of translation. In C. Picken (Ed.), The translator's handbook (pp. 91–102). London: Aslib.

Sager, J.C. (1994). Language engineering and translation. Amsterdam: John Benjamins.

Sheridan, P., & Ballerini, J.P. (1996). Experiments in multilingual information retrieval using the SPIDER system. In Proceedings of the 19th ACM SIGIR Conference (pp. 58–65). New York: ACM.

Sidiropoulou, M. (1995). Headlining in translation: English vs Greek press. Target, 7(2), 285–304.

Simard, M. (1999, June). Text-translation alignment: Three languages are better than two. Paper presented at Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (pp. 2–11).

Simard, M., Foster, G., & Isabelle P. (1992). 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.

Toury, G. (1986). Monitoring discourse transfer: A test-case for a developmental model of translation. In J. House & S. Blum-Kulka (Eds.), Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies (pp. 79–94). Tübingen: Gunter Narr.

Trochim, W. (1999). The research methods knowledge base (1st ed.). Cincinnati, OH: Atomic Dog Publishing.

Van Dijk, T.A. (1985). Structures of news in the press. In van Dijk (Ed.), Discourse and communication (pp. 69–93). Berlin: De Gruyter.

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

DOI: 10.1002/asi