ELSEVIER

Information Processing and Management xxx (2004) xxx–xxx

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Building parallel corpora by automatic title alignment using length-based and text-based approaches

Christopher C. Yang *, Kar Wing Li

Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Ho Sin Hang Engineering Building, Shatin, NT, Hong Kong

Received 27 February 2003; accepted 13 November 2003

Abstract

Cross-lingual semantic interoperability has drawn significant attention in recent digital library and World Wide Web research as the information in languages other than English has grown exponentially. Cross-lingual information retrieval (CLIR) across different European languages, such as English, Spanish, and French, has been widely explored; however, CLIR across European languages and Oriental languages is still in the initial stage. To cross language boundary, corpus-based approach is promising to overcome the limitation of the knowledge-based and controlled vocabulary approaches but collecting parallel corpora between European language and Oriental language is not an easy task. Length-based and text-based approaches are two major approaches to align parallel documents. In this paper, we investigate several techniques using these approaches and compare their performances in aligning English and Chinese titles of parallel documents available on the Web.

Keywords: Cross-lingual information retrieval; Parallel corpus; Sentence alignment; Covert translation

1. Introduction

Many parallel text alignment techniques have been developed in the past. These techniques attempt to map various textual units to their translation and have been proven useful for a wide range of applications and tools, e.g. cross-lingual information retrieval (Oard, 1997), bilingual lexicography, automatic translation verification and the automatic acquisition of knowledge about translation (Simard, 1999). Translation alignment technique has been used in automatic corpus construction to align two documents (Ma & Liberman, 1999).

* Corresponding author. Tel.: +852-2609-8239; fax: +852-2603-5505.
E-mail address: yang@se.cuhk.edu.hk (C.C. Yang).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2003.11.002

Given a text and its translation, an alignment is a segmentation of two texts such that the *n*th segment of one text is the translation of the *n*th segment of the other (Simard, Foster, & Isabelle, 1992). Empty segments are allowed which can be corresponding either to translator's omissions or to 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 deeper automatic analyses of translations. For example, it could be used to flag possible omissions in a translation, or to signal common translation mistakes, such as terminological inconsistencies.

There are three major structures of parallel documents on the World Wide Web, *parent page structure*, *sibling page structure*, and *monolingual sub-tree structure*. Resnik (1999) noticed that if a Web page has been written in many languages, the parent page of the Web page may contain the links to different versions of the Web page. For example, in a Web page, there are two anchor texts such as A_1 and A_2. A_1 is linked to Language 1 version and A_2 is linked to Language 2 version as shown in Fig. 1(a). The sibling page structure refers to the cases where the page in one language contains a link directly to the translated pages in the other language. The third structure contains a completely separate monolingual sub-tree for each language, with only the single top-level Web page pointing off to the root page of single-language version of the site (Resnik, 1999). Parallel corpus can be generated using *overt translation* or *covert translation*. The overt translation (Rose, 1981) posses a directional relationship between the pair of texts in two languages, which means texts in language A (source text) are translated into texts in language B (translated text) (Zanettin, 1998). The covert translation (Rose, 1981) is non-directional. Parallel corpora generated by overt translation usually use the parent page structure and sibling page structure. To collect parallel corpora based on parent page and sibling structures, link analysis is sufficient. However, parallel

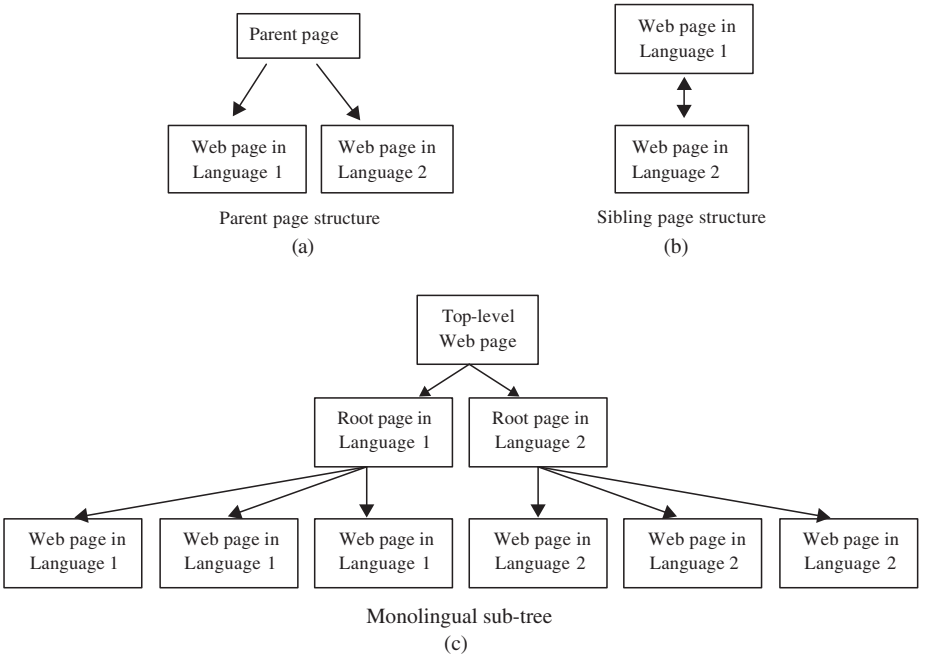

Fig. 1. Structures of parallel documents on the World Wide Web.

corpora generated by covert translation use monolingual sub-tree structure. Each sub-tree is generated independently. (The press release of the HKSAR government is a typical example.) Collecting parallel corpora based on monolingual sub-tree structure, techniques that are more advanced than link analysis are required since direct links or links through a parent page are not available between the pair of parallel documents. Length-based and text-based approaches are two typical approaches to align such parallel corpora (Fig. 1).

Given a set of parallel texts, the alignment that maximizes the probability over all possible alignments is retrieved (Gale & Church, 1991).

$$\arg\max_A Pr(A|T_1, T_2) \approx \arg\max_A \prod_{(L_1 \Longleftrightarrow L_2) \in A} Pr(L_1 \Longleftrightarrow L_2|L_1, L_2) \tag{1}$$

where A is an alignment, T_1 and T_2 are the English and Chinese texts, respectively L_1 and L_2 are the passages of two languages, $L_1 \Longleftrightarrow L_2$ is an individual aligned pairs, an alignment A is a set consisting of $L_1 \Longleftrightarrow L_2$ pairs.

1.1. Sentence alignment

Aligning a sentence with its translation is not fundamentally different from retrieving a sentence on the same topic as the source sentence in the target corpus using the source sentence as a query (Fluhr, Bisson, & Elkateb, 2000). These two processes are based on the semantic proximity of two sentences in different languages. The major difference is that retrieving sentence only needs to insure that the sentence retrieved contains most of the information available in the query, whereas sentence alignment requires the parts that are not common to both sentences in different languages to be as little as possible. As a result, in information retrieval, the proximity value refers the semantic overlap between the reference text (query) and the texts in the database. The larger the similarity between the retrieved text and the query, the more relevant the retrieved text is. In case that a sentence in one language aligns to a sentence in another language (1-1 mapping), a proximity value can be calculated to evaluate whether or not the two sentences can be considered to be translations of each other.

Sentence alignments can be viewed as mathematical relations between linguistic entities (Simard, 1999).

Given two texts, A and B, as sets of linguistic units: a_i and b_i, a binary alignment X_{AB} is defined as a relation on $A \cup B$:

$$X_{AB} = \{(a_1, b_1), (a_2, b_2), (a_3, b_3), \ldots\} \tag{2}$$

The interpretation of X_{AB} is: (a, b) belongs to X_{AB} if and only if some translation equivalence exists between a and b, total or partial.

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

Length-based algorithms assume that the sentences, which are mutual translations in the parallel corpus, are similar in length (Gale & Church, 1991). The sentence alignment algorithm developed by Brown, Lai, and Mercer (1991) is based on the number of words in each sentence.

Gale and Church (1991) developed a similar algorithm except that alignment is based on the number of characters in sentences. These approaches based exclusively on sentence lengths work quite well with a clean input, such as the Canadian Hansards and have been widely used by other researchers e.g. Resnik (1998), Chen, Kishida, Jiang, Liang, and Gey (1999), and Wu (1994). However, for cases where sentence boundaries are not clearly marked, such as OCR input (Church, 1993), or where the languages are originated from different family of languages, such as Asian–European language pairs (Fung, 1995; Wu, 1994), these algorithms do not perform well.

Text-based algorithms use lexical information across the language boundary to align sentences. Warwick-Armstrong and Russell (Warwick-Armstrong & Russell, 1990) used a bilingual dictionary to select word pairs in sentences from a parallel corpus and then aligned the sentences based on the word correspondence information. Another type of lexical information, which is helpful in alignment of European language pairs, is called cognate (Simard et al., 1992). Cognates are pairs of tokens of different languages, which share obvious phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations. The pair of *generation*/*génération* constitute a typical example for English and French. Simard et al. (1992) illustrated that cognate provides a reliable source of linguistic knowledge.

1.2. Title alignment

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

Titles of two texts can be treated as the representations of two texts. Referring to He (He, 2000), the titles present "micro-summaries of texts" that contain "the most important focal information in the whole representation" and as "the most concise statement of the content of a document". In other words, titles function as the condensed summaries of the information and content of the articles. The HKSAR government press releases have the titles of the Chinese articles and English articles listed on two separate Web pages. Aligning the parallel press release articles requires alignment of the Chinese titles with the English titles.

In order to align titles effectively, the characteristic of title translation pattern should be first analyzed carefully. Similar to Gale and Church (1991), three characteristics of translation pattern has been identified at the sentence (title) level:

(1) one title in Language A translates into one title in Language B;
(2) title is not translated at all, document is available in one language only, e.g. English only or Chinese only articles;
(3) title in Language A has no equivalent title translation in Language B.

In the second and third cases, there is either no alignment or impossible to find an alignment based on the titles. For example, an English title, "From bak choy to baguette . . .," appears in the English version of the Hong Kong government press release Web site. However, the corresponding Chinese title in the Chinese version of the Web site is "盧偉聰情繫國際刑警　心懸香港"

(Lo Wai-chung loves Interpol and Hong Kong). Although the corresponding English and Chinese documents possess the relationship of covert translation, the Chinese title is not equivalent to the English title.

The characteristics of translation pattern at word level can also be identified as follows:

(1) one word in Language A is translated one word in Language B;
(2) many words in Language A is translated into one word in Language B;
(3) some words are not translated;
(4) a word in Language A is not always translated in the same way in Language B;
(5) a word in Language A is translated into morphological or syntactic phenomena rather than a word in Language B.

In this paper, we have investigated seven alignment techniques (three length-based approaches and four text-based approaches) and compared their performances in aligning the Chinese and English titles of the Web documents to build the English/Chinese parallel corpora. The techniques that we have investigated are:

(1) Gale and Church's length-based approach;
(2) Wu's length-based approach with lexical cues;
(3) Sun et al.'s length-based approach with lexicon checks;
(4) Utsuro et al.'s dictionary-based approach;
(5) Melamed's geometric sentence alignment;
(6) Ma and Liberman's Bilingual Internet Text Search;
(7) Our proposed text-based approach using longest common subsequence.

2. Length-based approach

Length-based alignment method (Gale & Church, 1991) is developed based on the following approximation to Eq. (1):

$$Pr(L_1 \Longleftrightarrow L_2|L_1, L_2) \approx Pr(L_1 \Longleftrightarrow L_2|\ell_1, \ell_2) \tag{3}$$

where $\ell_1 = \text{length}(L_1)$ and $\ell_2 = \text{length}(L_2)$ measured in the number of characters.

The length-based alignment model assumes that each character in L_1 is responsible for generating some number of characters in L_2. This leads to a further approximation that encapsulates the dependence to a single parameter δ. δ is function of ℓ_1 and ℓ_2.

$$Pr(L_1 \Longleftrightarrow L_2|L_1, L_2) \approx Pr(L_1 \Longleftrightarrow L_2|\delta(\ell_1, \ell_2)) \tag{4}$$

Based on the Bayesian Rule,

$$Pr(L_1 \Longleftrightarrow L_2|\delta) = \frac{Pr(\delta|L_1 \Longleftrightarrow L_2)Pr(L_1 \Longleftrightarrow L_2)}{Pr(\delta)} \tag{5}$$

Although it has been suggested that length-based methods are language-independent (Gale & Church, 1991), they may in fact rely on some extent on length correlations arising from the historical relationships of the languages being aligned. If translated sentences share cognates, then

the character lengths of those cognates are correlated. Grammatical similarities between related languages may also produce correlations in sentence lengths. However, Chinese and English have no history of common development.

An experiment has been conducted to test the correlation between the English and Chinese titles. Since Chinese texts do not contain obvious word boundaries but consists of a linear sequence of non-spaced or equally spaced ideographic characters (Wu & Tseng, 1993). Wu's byte count approach (Wu, 1994) is used to count each Chinese character as a length of 2 and each English or punctuation character as a length of 1. Fig. 2 shows the plot of the length of the Chinese titles against the English titles. The mean number of Chinese characters generated by each English character is $c = E(\ell_2/\ell_1) = 0.7316$, with a standard deviation $\sigma = 0.19767$. The correlation is 0.7033. The results show that the data points are substantially scatter in the plot and many data points are deviated from the regression line. The correlation between the length of Chinese titles and the length of English titles is not high. As a result, purely rely on the length for aligning Chinese and English titles may not be reliable.

Wu (1994) assumed that $\ell_2 - \ell_1 c$ is normally distributed and it can be transformed into a new Gaussian variable of standard form (i.e. with the mean 0 and variance 1) by the appropriate normalization:

$$\delta(\ell_1, \ell_2) = \frac{\ell_2 - \ell_1 c}{\sqrt{\ell_1 \sigma^2}} \tag{6}$$

Fig. 3 plots the distribution of δ for 150 aligned Chinese and English titles. The distribution deviates from a gaussian distribution substantially. The result is worse than what Gale and Church (1991) has reported in their experiment for French/German/English alignment. It raises further doubts about the potential performance of pure length-based alignment.

According to Gale and Church (1991), the prior of 6 classes of alignment are used to estimate $Pr(L_1 \Longleftrightarrow L_2)$. The six classes include a sentence in one language matching exactly one sentence in the other language(1-1) and several additional possibilities (1-0, 0-1, 2-1, 1-2, 2-2). Table 1 shows the values of $Pr(L_1 \Longleftrightarrow L_2)$ for each of the six classes. For title alignment, only three classes, 1-1,

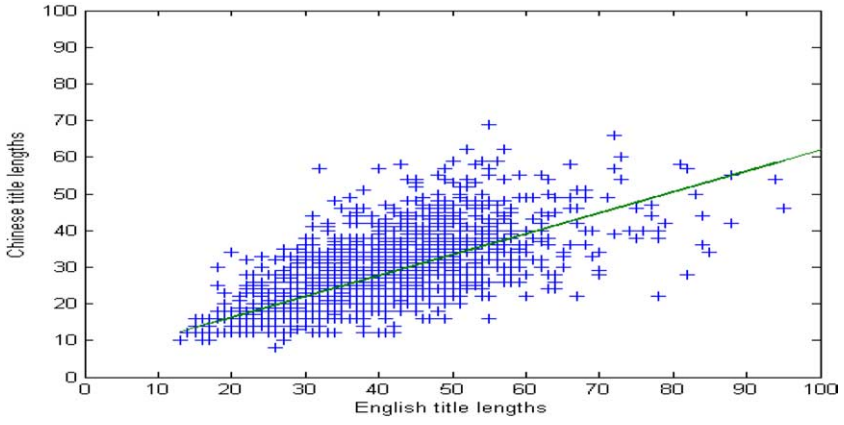

Fig. 2. A plot of the length of the Chinese titles against the length of the English titles for 150 aligned title pairs retrieved from the HKSAR press releases. The regression line is shown in the plot.

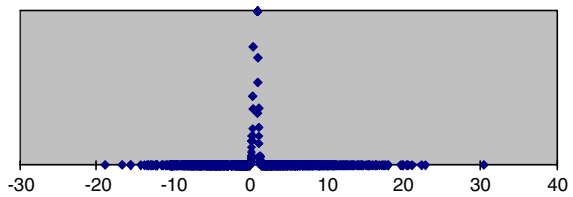

Fig. 3. Empirical density of δ for 150 aligned title pairs.

Table 1
$Pr(L_1 \Longleftrightarrow L_2)$

Class	$Pr(L_1 \Longleftrightarrow L_2)$ used in (Gale & Church, 1991)	$Pr(L_1 \Longleftrightarrow L_2)$ for title alignment
1-1	0.89	0.878
1-0 (or 0-1)	0.00099	0.122
2-1 (or 1-2)	0.089	
2-2	0.011	

1-0 and 0-1, are utilized due to the characteristics of translation pattern at title level as described in Section 1.2.

The dynamic programming algorithm is then applied to determine the minimum distance $D(i,j)$ between the sentences $s_1,\ldots,s_i$ and their translations $t_1,\ldots,t_j$, under the maximum likelihood alignment. $D(i,j)$ is computed by minimizing over the six classes.

$$D(i,j) = \min \begin{cases} D(i,j-1) + d(0,t_j;0,0) \\ D(i-1,j) + d(s_i,0;0,0) \\ D(i-1,j-1) + d(s_i,t_j;0,0) \\ D(i-1,j-2) + d(s_i,t_j;0,t_{j-1}) \\ D(i-2,j-1) + d(s_i,t_j;s_{i-1},0) \\ D(i-2,j-2) + d(s_i,t_j;s_{i-1},t_{j-1}) \end{cases} \tag{7}$$

2.1. Wu's length-based approach with lexical cues

To improve the purely length-based alignment, Wu (1994) incorporated lexical criteria without giving up the statistical approach. Eq. (4) is modified as follows:

$$Pr(L_1 \Longleftrightarrow L_2|L_1,L_2) \approx Pr(L_1 \Longleftrightarrow L_2|\ell_1,\ell_2,v_1,\omega_2,\ldots,v_n,\omega_n) \tag{8}$$

where $v_i =$ no. of occurrences of English cue$_i$ in L_1, $w_i =$ no. of occurrences of Chinese cue$_i$ in L_2.

Consequently, Eq. (5) is modified as follows:

$$Pr(L_1 \Longleftrightarrow L_2|L_1,L_2) \approx Pr(L_1 \Longleftrightarrow L_2|\delta_0(\ell_1,\ell_2),\delta_1(v_1,\omega_2),\ldots,\delta_n(v_n,\omega_n)) \tag{9}$$

2.2. Sun et al.'s length-based approach with lexicon check

Sun, Du, Sun, and Jin (1999) utilized an English/Chinese lexicon to check the result of the alignments obtained from the length-based approach. A score S_A is computed for every aligned

sentence pair. Aligned sentence pairs that score above a threshold, $t1$, are judged as correct alignments. After removing the correct alignments, the rest are aligned by the length-based approach again. The second result is checked by lexicon again and the alignments whose score is above a threshold, $t2$, are considered as correct alignments.

$$S_A = \frac{\mathrm{No}_{\mathrm{correct}} \times 2}{\mathrm{No}_{\mathrm{English}} + \mathrm{No}_{\mathrm{Chinese}}} \tag{10}$$

where $\mathrm{No}_{\mathrm{correct}}$ corresponds to the number correct alignment of English and Chinese words identified by the lexicon, $\mathrm{No}_{\mathrm{English}}$ corresponds to the number of words in English sentence, $\mathrm{No}_{\mathrm{Chinese}}$ corresponds to the number of words in Chinese sentence.

3. Text-based approach

3.1. Utsuro et al.'s dictionary-based approach

Utsuro, Ikeda, Yamane, Matsumoto, and Nagao (1994) have proposed a sentence alignment approach based on word pairs available in the bilingual dictionary and the statistical information of word pairs. A dictionary was first used to align parallel sentences. Word pairs that are not available in the dictionary will then be evaluated based on their frequencies.

n sentences in language A and m sentences in language B are grouped into $m \times n$ pairs, P. ($P = p_1, p_2, \ldots, p_k$, where p_k is a sentence pair.) Words are first extracted from each sentence and their correspondences are identified using dictionary. Based on the word pairs that are identified, the score $h(p)$ of the sentence pair p_k is computed as follows:

$$h(p) = \frac{n_{\mathrm{st}}(p)}{n_s(a,x) + n_t(b,y)} \tag{11}$$

where $n_{\mathrm{st}}(p)$ the number of word pairs in the sentence pairs p, $n_s(a,x)$ is the number of words in the sequences of sentences $S_{a-x+1}, \ldots, S_a$ in language S, $n_t(b,y)$ is the number of words in the sequences of sentences $T_{b-y+1}, \ldots, T_b$ in language T.

The score function follows the recursion equation below:

$$H(P_i) = H(P_{i-1}) + h(p_i) \tag{12}$$

where P_i is the sequence of sentence pairs from the beginning of the bilingual text to the pair p_i.

The maximum score of $H(P_i)$ will be the optimal solution to the alignment problem.

3.2. Melamed's geometric sentence alignment (GSA)

Melamed (1996) extended the smooth injective map recognizer (SIMR) to develop an algorithm called geometric sentence alignment (GSA) for sentence alignment. The smooth injective map recognizer (SIMR) is based on a greedy algorithm for mapping bitext correspondence. A bitext comprises two versions of a text, such as a text in two different languages. Translators create a

bitext each time they translate a text. Each bitext defines a rectangular bitext space. The lower left corner of the rectangle is the origin of the bitext space and it represents the beginning of two texts. The upper right corner is the terminus and it represents the end of two texts. The line between the origin and the terminus is the main diagonal. The width and height of the rectangle are the lengths of the two component texts, in characters.

Each bitext space contains a number of true points of correspondence (TPCs), other than the origin and the terminus. For example, if a token at position p on the x-axis and a token at position q on the y-axis are translations of each other, then the coordinate (p, q) in the bitext space is a TPC. Since distances in the bitext space are measured in characters, the position of a token is defined as the mean position of its characters. TPCs exist at the corresponding boundaries of text units such as sentences. Groups of TPCs with a roughly linear arrangement in the bitext space are called chains. For each bitext, the true bitext map (TPM) is shortest bitext map that runs through all the TPCs. SIMR considers only chains that are roughly parallel to the main diagonal. Since Chinese and English languages do not share an alphabet, the Chinese/English matching predicate deemed two tokens to match if they constituted an entry in the translation lexicon (Melamed & Marcus, 1998).

3.3. Ma and Liberman's Bilingual Internet Text Search (BITS)

Ma and Liberman (1999) have developed a system called Bilingual Internet Text Search (BITS). To determine if two texts are mutual translation of each other, corresponding regions of one text and its translation will contain word token pairs that are mutual translations.

Given text A in language $L1$ and text B in language $L2$, text A and text B are tokenized. The similarity between A and B is computed as follows:

$$\mathrm{sim}(A, B) = \text{Number of translation token pairs/Number of tokens in text } A \tag{13}$$

If text B is most similar to text A, and $\mathrm{sim}(A, B)$ is greater than a threshold, t, then text A and text B are treated as a translation pairs. The following is their algorithm:

```
For each text A in language L1
   Tokenize A
   max_sim = 0
   For each text B in language L2
      Tokenize B
      s = sim(A, B)
      If s > max_sim Then
       max_sim = s
       most_sim = B
      Endif
   Endfor
   If max_sim > t Then
      Output(A, B)
   Endif
Endfor
```

3.4. Proposed text-based approach

In our proposed text-based approach, the longest common subsequence is utilized to optimize the alignment of English and Chinese titles. The longest common subsequence (LCS) is commonly exploited to maximize the number of matches between characters of two sequences. Our alignment algorithm has three major steps: (1) alignment at word level and character level, (2) reducing redundancy, (3) score function.

3.4.1. Alignment at word level and character level

An English title, E, is formed by a sequence of English simple words, i.e., $E = e_1e_2e_3 \cdots e_i \cdots$, where e_i is the ith English word in E. A Chinese title, C, is formed by a sequence of Chinese characters, i.e., $C = \text{char}_1\text{char}_2\text{char}_3 \cdots \text{char}_q \cdots$, where char_q is a Chinese character in C.

An English word in E, e_i, can be translated to a set of possible Chinese translations, Translated(e_i), by dictionary lookup. Translated$(e_i) = \{T_{e_i}^1, T_{e_i}^2, T_{e_i}^3, \ldots, T_{e_i}^j, \ldots\}$ where $T_{e_i}^j$ is the jth Chinese translation of e_i. Each Chinese translation is formed by a sequence of Chinese characters. The set of the longest-common-subsequence (LCS) of a Chinese translation $T_{e_i}^j$ and C is LCS$(T_{e_i}^j, C)$. MatchList(e_i) is a set that holds all the unique longest common subsequences of $T_{e_i}^j$ and C for all Chinese translations of e_i.

$$\text{MatchList}(e_i) = \bigcup_j \text{LCS}(T_{e_i}^j, C) \tag{14}$$

If there is no common subsequence of $T_{e_i}^j$ and C, MatchList$(e_i) = \emptyset$ and no reliable translation of e_i can be found in C. If there is at least one common subsequence of $T_{e_i}^j$ and C, we determine the most reliable translation based on the adjacency and length of Chinese translations found in C.

Based on the hypothesis that if the characters of the Chinese translation of an English word appears adjacently in a Chinese sentence, such Chinese translation is more reliable than other translations that their characters do not appear adjacently in the Chinese sentence. For example, the English word "propose" can be translated as "建議" in Chinese. The translation "建議" can be aligning with "就建築條例的動議辯論" (on the "Construction Bill" motion debate) using LCS, which is not correct in this case. Contiguous(e_i) is used to determine the most reliable translation based on adjacency.

$$\text{Contiguous}(e_i) = \{x | x \in \text{MatchList}(e_i) \text{ and all the characters of } x \text{ appear adjacently in } C\} \tag{15}$$

The second criteria of the most reliable Chinese translations, is the length of the translations. Reliable(e_i) is used to identify the longest sequence in Contiguous(e_i).

$$\text{Reliable}(e_i) = \begin{cases} \arg\max_{x \in \text{Contiguous}(e_i)} |x| & \text{if Contiguous}(e_i) \neq \emptyset \\ \arg\max_{x \in \text{MatchList}(e_i)} |x| & \text{Otherwise} \end{cases} \tag{16}$$

3.4.2. Resolving redundancy

Due to redundancy, the translations of an English word may be repeated completely or partially in Chinese. For example, given $E =$ red color and $C =$ 赤紅色, Translated(red) = {紅, 紅色,

紅色的, 赤}} and Translated(color) = {色, 顏色, 色彩}. MatchList(red) = {紅, 紅色, 赤} and MatchList(color) = {色}. Reliable(red) = 紅色 and Reliable(color) = 色. To deal with redundancy, Dele(x, y) is an edit operation to remove the LCS(x, y) from x. WaitList is a list to save all the sequences obtained by removing the overlapping of the elements of MatchList(e_i) and Reliable(e_i). MatchList(e_i) is initialized to $\emptyset$ and Reliable(e_i) is initialized to ε.

$$\text{WaitList} = \text{DELE}(\text{WaitList}, \text{Reliable}(e_i)) \cup \text{DELE}(\text{MatchList}(e_i)\{\text{Reliable}(e_i)\}, \text{Reliable}(e_i)) \tag{17}$$

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

Remain is a sequence that is initialized as C, and Reliable(e_i) are removed from Remain starting from the e_1 until the last English word. WaitList will also be updated for each e_i. When all Reliable(e_i) are removed from Remain, the elements in WaitList will also be removed from Remain in order to remove the redundancy.

3.4.3. 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.

$$\text{Matching\_Ratio}(E, C) = \frac{|C| - |\text{Remain}|}{|C|} \tag{18}$$

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

$$\text{Matching\_Ratio}^*(E, C) = \frac{\sum_i R(e_i)}{|E|} \tag{19}$$

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

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

4. Experiment

An experiment is conducted to measure the precision and recall of the aligned parallel Chinese/English documents from the HKSAR government press releases between 1998 and 2001 using the length-based approaches and the text-based approaches as described in Sections 2 and 3. Press release articles of the HKSAR government are usually distributed through the World Wide Web in English and/or Chinese based on the covert translation. However, it is not necessary for all articles to be published in both languages. In some cases, only the English version is available or

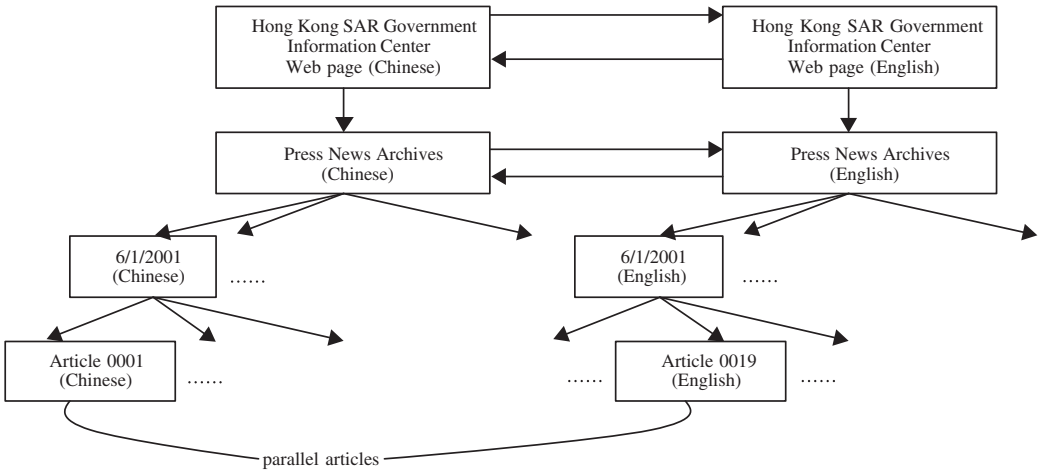

Fig. 4. Organization of Hong Kong SAR Government's press release articles in the Hong Kong SAR Government Information Center Web site.

only the Chinese version is available. There are approximately 40 articles published in each language every day by the government. Fig. 4 shows the organization of the Hong Kong SAR Government News Archives. The arcs in the figure represent the link between the Web pages. There are 31,567 Chinese articles, 30,810 English articles, and 23,701 pairs of English/Chinese parallel articles in the HKSAR government press release corpus.

4.1. Evaluation Metric

In the previous work of alignment techniques, accuracy has been widely used for their evaluation (Wu, 1994; Utsuro et al., 1994; Sun et al., 1999; Melamed, 1996). Accuracy measures the proportion of correct aligned pairs extracted by the system. Recently, Simard and Plamondon (1998) have defined another metric, alignment precision and alignment recall, using the terminology of information retrieval. Their definitions are as follow:

$$\text{Alignment precision} = \frac{|A_R \cap A|}{|A|}$$

$$\text{Alignment recall} = \frac{|A_R \cap A|}{|A_R|}$$

where A is the set of aligned pairs extracted by the system, A_R is the set of correct aligned pairs in the corpus.

The alignment precision measures how much of the alignment extracted by the system is correct. The alignment precision is exactly the same as the accuracy adopted in the evaluation metric of many other papers. The alignment recall measures how much of correct alignment is found by the system. The precision and recall have also been used by Ma and Liberman (1999), Nie and Cai (2001), and Langlais, Simard, and Veronis (1998) to evaluate the performance of their proposed alignment techniques.

Table 2
Experimental results

	Alignment precision	Alignment recall
Gale and Church's length-based approach	0.10	0.06
Wu's length-based approach with lexical cues	0.62	0.61
Sun et al.'s length-based approach with lexicon checks	0.76	0.05
Utsuro et al.'s dictionary-based approach	0.91	0.82
Melamed's GSA (text-based approach)	0.73	0.65
Ma and Liberman's BITS (text-based approach)	0.93	0.86
Our proposed text-based approach using LCS	1.00	0.87

In this work, we shall adopt the alignment precision and alignment recall as our evaluation metric. This metric includes a measure that is equivalent to the traditional accuracy measure and an additional measure of alignment recall. Simply using accuracy (or alignment precision) is not complete in evaluating the alignment performance. One alignment technique can be high in accuracy but it may only find a small percentage of the alignment in the corpus. We should also consider how much correct alignment it can find.

4.2. Experimental results

Experimental results are shown in Table 2.

Experimental result shows that the text-based approaches out-perform the length-based approaches and our proposed text-based approach using LCS has the best performance. Chinese text contains fewer characters; character length is a less discriminating feature, varying over a range of fewer possible discrete values than the corresponding English is. As a result, the length-based approach is not as reliable as the text-based approach in the title alignment. Lexical knowledge can effectively improve both alignment precision and alignment recall in the title alignment. Since our proposed approach has adopted the longest common sequence to consider those Chinese translations that do not appear as adjacent characters in the Chinese sentence and the problem of redundancy, it produces the best performance.

4.3. Discussion

In this section, we shall discuss and compare the performance of the seven alignment techniques.

Gale and Church's length-based approach is developed based on the probability (Eq. (5)) of an alignment given a parameter δ where δ is a function of the lengths of the Chinese and English titles. Based on our experimental results, the average and standard deviation of the probability of the aligned pairs using Eq. (5) are 0.41 and 0.36. It is found that the average probability is low and the standard deviation is high. It infers that the correlation between the lengths of Chinese and English titles is not high. Purely utilizing the information of length cannot produce a reliable alignment. This is why the alignment precision and alignment recall are only 0.10 and 0.06, respectively.

Wu has improved Gale and Church's approach by incorporating lexical criteria using Eq. (9). Our experimental results show that the average and standard deviation of the probability of the aligned pairs using Eq. (9) are 0.61 and 0.21. Comparing to Gale and Church's probability function (Eq. (5)), the average has been increased and the standard deviation has been decreased. As a result, the alignment precision has been significantly increased from 0.10 to 0.62 and the alignment recall has been significantly increased from 0.06 to 0.61. It shows that the performance of alignment can be increased by incorporating the lexical criteria with the pure length-based approach. However, the alignment precision and alignment recall of Wu's approach are still considered low because his approach only considers co-occurrence of Chinese and English cues.

Sun's length-based approach utilizes the score function in Eq. (10) for lexicon checks in addition to the lengths of Chinese and English titles. Our experimental results show that average and standard deviation of the score computed by Eq. (10) for the aligned pairs are 0.80 and 0.09. The average score of the aligned pairs using Eq. (10) is higher than the average probability of the aligned pairs using Eq. (9). The standard deviation using Eq. (10) is lower than that of using Eq. (9). As a whole, the alignment precision is improved by Eq. (10). However, the alignment recall is significantly decreased to 0.05. A high threshold of the score can ensure a higher alignment precision but it significantly decreases the number of extracted alignments. That means the alignment recall is extremely poor.

The three length-based approaches by Gale and Church, Wu and Sun are not very satisfactory. Although Sun's approach has higher alignment precision than Wu's approach, Sun's approach has extremely poor alignment recall. Wu's length-based approach with lexical cues is the best among the three approaches. In general, we find that the lexical cues or lexical checks incorporating with the pure length-based approach can significantly improved the performance in comparing with Gale and Church's pure length-based approach.

Utsuro's text-based approach utilizes the score function in Eq. (12). Eq. (12) is similar to the score function (Eq. (10)) utilized by Sun except that Utsuro's score function does not have a multiple of two and considers the sequence of sentences in two languages. Sun used Eq. (10) as a filtering criterion in addition to the length. Utsuro used Eq. (12) to identify the optimized mapping without any information of length. Our experiment results show that the average and standard deviation of the score computed by Eq. (12) for the aligned pairs are 0.45 and 0.10. The alignment precision and alignment recall are increased to 0.91 and 0.82, respectively. It shows that simply using the score function to identify the best mapping is more effective than using the score function and length for filtering.

Melamed uses the SIMR to identify the true bitext map based on the true points of correspondence obtained from the tokens in Chinese and English text. No probability or score functions are utilized. However, the alignment precision and alignment recall are 0.73 and 0.65 that are lower than those achieved by Utsuro's approach.

Ma and Liberman's BITS uses the similarity function in Eq. (13) for alignment. Our experimental results show that the average and standard deviation of the similarity values for the aligned pairs are 0.93 and 0.09. The similarity of the aligned pairs is high and the standard deviation is low. The alignment precision and alignment recall are 0.93 and 0.86. The performance is higher than the performance achieved by Utsuro's approach but not significantly. Utsuro's score function (Eq. (12)) and Ma and Liberman's similarity functions (Eq. (13)) are similar. Utsuro's score function considers both the number of words in English text and Chinese

text in its denominator. Ma and Liberman's similarity function only considers the number of tokens in one text in its denominator. Therefore, the average value of the aligned pairs obtained by Utsuro's score function is approximately half of that by Ma and Liberman's similarity function. The overall performance of Utsuro's approach and Ma and Liberman's BITS are very close.

Utsuro's approach and Ma and Liberman's BITS achieve significantly better performance than the length-based approaches. However, in their score function or similarity function for alignment, they only consider the word pairs or token pairs from the English and Chinese text that are obtained in lexicon.

In our text-based approach, we use the longest common subsequence to identify the reliable translation of word pairs in English and Chinese text that may or may not be simply obtained in the lexicon. Our approach also considers the problem of redundancy. The score functions in Eqs. (18) and (19) are then used to identify the alignment. Our experimental results show that the averages and standard deviations of the scores using Eq. (18) for the aligned pairs are 0.98 and 0.05. Given an English title, in case there is more than one Chinese title obtains the same score using Eq. (18), Eq. (19) will be utilized. For those cases, the average and standard deviation of the scores using Eq. 19 are 0.96 and 0.06. It shows that the average scores are high using Eqs. (18) and (19) and the standard deviations are low. As a result, the alignment precision of our approach is perfect, which is not achievable by any other investigated approaches. It is also significantly higher than all the other approaches. The alignment recall of our approach is the highest among all approaches but it is close to the alignment recall of Utsuro's and Ma and Liberman's approaches. It proves that the additional effort of using longest common subsequences to identify the most possible translation of word pairs that do not appear in lexicon and resolving the redundancy problem can significantly improve the alignment precision. The perfect alignment precision ensures that all the aligned pairs obtained by our approach are correct.

5. Conclusion

Cross-lingual information retrieval has drawn significant attention recently. Parallel corpora are important linguistic resources that provide a statistical translation model to cross the language boundary. However, constructing English/Chinese parallel corpora is not an easy task due to the significant difference between two languages. In this paper, we have investigated seven English/Chinese sentence (or title) alignment techniques. Three of them are length-based approaches and four of them are text-based approaches. Experimental result shows that the text-based approaches out-performed the length-based approaches. In particular, our proposed text-based approach using LCS and score function based on matching ratios produces the best performance with 100% alignment precision and 87% alignment recall.

Acknowledgements

This project was supported by the Direct Research Grant of the Chinese University of Hong Kong, 2050268.

References

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*, Berkeley, CA, USA.

Chen, A., Kishida, K., Jiang, H., Liang, Q., & Gey, F. (1999). Automatic construction of a Japanese–English lexicon and its application in cross-language information retrieval. In *Proceedings of the multilingual information discovery and access workshop of the ACM SIGIR'99 conference*, August 14.

Church, K. W. (1993). Char_align: a program for aligning parallel texts at the character level. In *Proceedings of ACL-93*, Columbus, OH.

Fluhr, C., Bisson, F., & Elkateb, F. (2000). Parallel text alignment using crosslingual information retrieval techniques. In J. Veronis (Ed.), *Parallel text processing: alignment and use of translation corpora* (pp. 187–200).

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

Fung, P. (1995). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd annual meeting of the association for computational linguistics*, Boston, MA.

Gale, W. A., & Church, K. W. (1991). Identifying word correspondences in parallel texts. In *Proceedings of the fourth DARPA workshop on speech and natural language*, Asilomar, California.

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.

Langlais, P., Simard, M., & Veronis, J. (1998). Methods and practical issues in evaluating alignment techniques. In *Proceedings of 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistic*, Montreal, Canada.

Ma, X., & Liberman, M. (1999). BITS: a method for bilingual text search over the Web. In *Machine translation summit VII, September 13th, 1999*, Kent Ridge Digital Labs, National University of Singapore.

Macklovitch, E., & Hannan, M.-L. (1996). 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.

Melamed, I. D., & Marcus, M. P. (1998). *Automatic construction of Chinese–English translation lexicons*, IRCS Technical Report #98-28.

Melamed, I. D. (1996). A geometric approach to mapping bitext correspondence. In *Proceedings of the first conference on empirical methods in natural language processing (EMNLP'96)*, Philadelphia, PA.

Nie, J., & Cai, J. (2001). Filtering noisy parallel corpora of Web pages. In *Proceedings of IEEE symposium on NLP and knowledge engineering*, Tucson AZ, October (pp. 453–458).

Oard, D. W. (1997). Alternative approaches for cross-language text retrieval. In D. Hull, & D. Oard (Eds.), *AAAI symposium in cross-language text and speech retrieval*. American Association for Artificial Intelligence, March, 1997.

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)*, Langhorne, PA, lecture notes in artificial intelligence 1529, Springer, October.

Resnik, P. (1999). Mining the Web for bilingual text. In *37th annual meeting of the association for computational linguistics (ACL'99)*, College Park, Maryland, June.

Rose, M. G. (1981). Translation types and conventions. In M. G. Rose (Ed.), *Translation Spectrum: Essays in Theory and Practice* (pp. 31–33). New York: State University Press.

Simard, M., & Plamondon, P. (1998). Bilingual sentence alignment: balancing robustness and accuracy. *Machine Translation, 13*(1), 59–80.

Simard, M. (1999). Text-translation alignment: three languages are better than two. In *Proceedings of EMNLP/VLC-99*, College Park, MD.

Simard, M., Foster, G., & Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. In *Fourth international conference on theoretical and methodological issues in machine translation (TMI-92)*, Montreal, Canada.

Sun, L., Du, L., Sun, Y., & Jin, Y. (1999). Sentence alignment of English–Chinese complex bilingual corpora. In *Proceeding of the 5th natural language processing Pacific Rim symposium*, Beijing, China.

Utsuro, T., Ikeda, H., Yamane, M., Matsumoto, Y., & Nagao, M. (1994). Bilingual text matching using bilingual dictionary and statistics. In *Proceeding of 15th international conference on computational linguistics*, Kyoto.

Warwick-Armstrong, S., & Russell, G. (1990). *Bilingual concordancing and bilingual lexicography, Euralex*.

Wu, D. (1994). Aligning a parallel English–Chinese corpus statistically with lexical criteria. In *32nd annual conference of the association for computational linguistics*, Las Cruces, New Mexico (pp. 80–87).

Wu, Z., & Tseng, G. (1993). Chinese text segmentation for text retrieval: achievements and problems. *Journal of The American Society for Information Science, 44*(9), 532–542.

Zanettin, F. (1998). Bilingual comparable corpora and the training of translatorsS. Laviosa (Ed.), *The corpus-based approach: a new paradigm in translation studies [special issue]. META, 43*(4), 616–630.