Mining English/Chinese Parallel Documents from the World Wide Web

Christopher C. Yang
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, Hong Kong
yang@se.cuhk.edu.hk

Kar Wing Li
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, Hong Kong

ABSTRACT

The information available in languages other than English on the World Wide Web is increasing significantly. To cross language boundaries between different languages, dictionaries are the most typical tools. However, the general-purpose dictionary is less sensitive in genre and domain and it is impractical to manually construct tailored bilingual dictionaries or sophisticated multilingual thesauri for large applications. Corpus-based approaches, which do not have the limitation of dictionaries, provide a statistical translation model to cross the language boundary. The objective of this research work is to mine English/Chinese parallel documents automatically from the World Wide Web. In this paper, we present an alignment method based on dynamic programming to identify the one-to-one Chinese and English title pairs. The longest common subsequence (LCS) is applied to find the most reliable Chinese translation of an English word. A score function is then proposed to determine the optimal title pairs. Experiments have been conducted to investigate the performance of the proposed method. The precision of the result is 0.995 while the recall is 0.8096.

Keywords

Parallel Documents, Covert Translation, Cross-lingual Information Retrieval

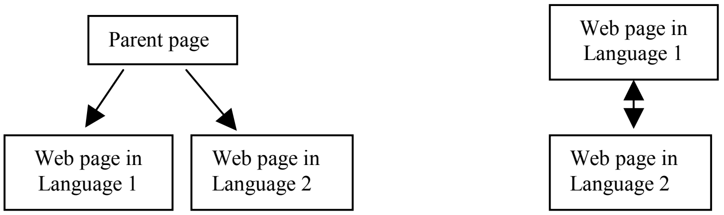
1. INTRODUCTION

In multilingual applications, dictionary-based approach is common; however, it lacks of sensitivity in genre and domain. Corpus-based approach overcomes such problem. Corpus-based approach can be viewed as automatic thesaurus construction techniques where the relationship between terms is obtained from statistics of term usage across different languages [6]. In this paper, we focus on the mining of the multilingual (or bilingual) documents from the World Wide Web. Document pairs aligned based on "parallelism" is known as parallel documents, which are generated using either overt translation or covert translation. The overt translation [9] possess a directional relationship between the pair of texts in two languages, which means texts in language A (source text) is translated into texts in language B (translated text) [10]. In this case, links are usually available on the Web page in language A to the Web page in language B and vice versa. The covert translation [4] is non-directional. Multilingual documents expressing the same content in different languages are generated by the same source [2]. Links are usually not available since the parallel documents are generated independently.

In this paper, we focus on automatic construction of English/Chinese parallel documents collected from the Web sites with monolingual sub-tree structure. The English/Chinese alignment model at the character level, word level and title level is presented. Longest common subsequence is employed to maximize the number of marches between the titles of the English and Chinese documents. As one word in one language may translate into two or more words in another language, deletion, an edit operation, is used to resolve redundancy in the alignment at word level. Score functions are developed to align English and Chinese titles.

2. Automatic Parallel Corpus Construction

Several techniques have been developed to construct parallel documents automatically. The most prominent system that generates parallel documents from the World Wide Web is Structural Translation Recognition for Acquiring Natural Data (STRAND), developed by Resnik [7]. Resnik [8] noticed that if a web page has been written in many languages, the parent page of the Web page might 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 Figure 1(a). Another phenomenon is "sibling" pages. This refers to the cases where the page in one language contains a link directly to the translated pages in the other language as shown in Figure 1(b). Parallel corpus based on overt translation usually has these structures.



(a) (b)

Figure 1. Web site structure for both sibling and parent pages

Some Web sites with bilingual text are arranged according to a third characteristic. They contain 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 [8]. Parallel corpus based on covert translation usually has this structure, which is also the focus of this paper.

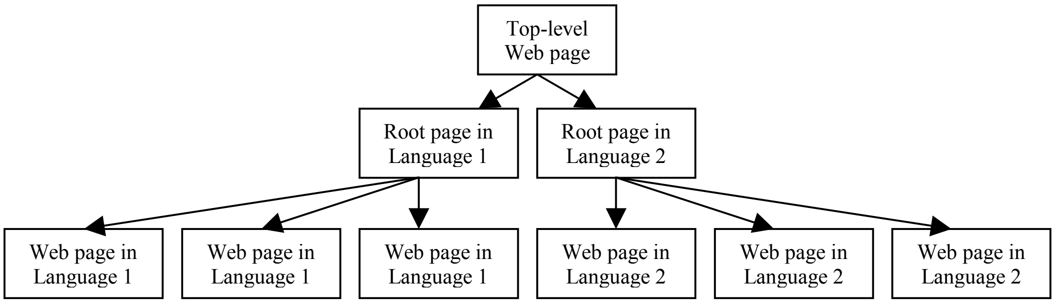


Figure 2. Web site structure for monolingual sub-tree

3. English/Chinese Alignment Model

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 [5]. According to He [3], the titles of document 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. To take the advantage of importance of titles, we use English and Chinese titles to construct the bilingual parallel documents automatically. In the title alignment model, 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), employed in sequence comparison methods, is utilized to optimize the alignment of English and Chinese titles.

3.1 Longest common subsequence (*LCS*)

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 ith English word in *E*. A Chinese title, *C*, is formed by a sequence of Chinese characters, i.e., $C = char_1\, char_2\, char_3 \ldots char_q \ldots$, where $char_q$ is a Chinese character in C. An English word in E, e_i, can be translated to a set of possible Chinese translations, Translated(e_i), by dictionary lookup. $Translated(e_i) = \{ T^1_{e_i}, T^2_{e_i}, T^3_{e_i}, \ldots, T^j_{e_i}, \ldots \}$ where $T^j_{e_i}$ 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 $T^j_{e_i}$ and *C* is $LCS(T^j_{e_i}, C)$. $MatchList(e_i)$ is a set that holds all the unique longest common subsequences of $T^j_{e_i}$ and *C* for all Chinese translations of e_i.

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

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.

Contiguous(e_i) is used to determine the most reliable translation based on adjacency.

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

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

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

The translations of two or more English words in an English title can be overlapped in a Chinese title and the translation of an English phrase may cause repetition in Chinese. To resolve the redundancy, an edit operation, *deletion* [1]**,** is used to remove the overlapping of two sequences of Chinese characters (Chinese translations) from the former sequence of Chinese characters, where the former sequence is any *LCS* of the Chinese translations of English word and *C* and the later sequence is the most reliable translation identified. The characters in the former sequence that do not match with the later sequence are remained and these sequences of characters will be saved in a waiting list. The sequences in such waiting list may match with a sequence in *C* that does not match with the reliable translation of other words since these sequences are redundant in translation.

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)*, which is initialized to $\varnothing$.

$$WaitList = DELE(WaitList, Reliable(e_i)) \cup DELE(MatchList(e_i)\backslash\{Reliable(e_i)\}, Reliable(e_i)) \qquad (4)$$

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

x_i is the i^{th} element of X

If the elements of *WaitList* match with the Chinese title after all the reliable translations of English words in English title are removed, redundancy may occur.

3.2 Ratio of Matching

Given *E* and *C*, *Reliable(e_i)* for each e_i are found. *Remain* is a sequence that is initialized as *C*, and *Reliable(e_i)* are removed from *Remain* starting from the e_1 until the last English word. *WaitList* will also be updated for each e_i. When all *Reliable(e_i)* are removed from *Remain*, the elements in *WaitList* will also be removed from *Remain* in order to remove the redundancy. Given *E* and *C*, the ratio of matching is determined by the portion of *C* that matches with the reliable translations of English words in *E*.

$$Matching\_Ratio(E,C) = \frac{|C| - |\mathrm{Re}main|}{|C|} \qquad (5)$$

Given an English title, the Chinese title that has the highest *Matching_Ratio* among all the Chinese titles is considered as the counterpart of the English title. However, it is possible that more than one Chinese title 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.

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

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

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

4. Experiments

In this paper, we have conducted an experiment to measure the performance of the mining of English/Chinese parallel documents. Press release articles and monthly reports of he 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 only the Chinese version is available. There are approximately 40 articles published in each language every day by the government. Figure 3 shows the organization of the Hong Kong SAR Government News Archives. The arcs in the figure represent the link between the Web pages.

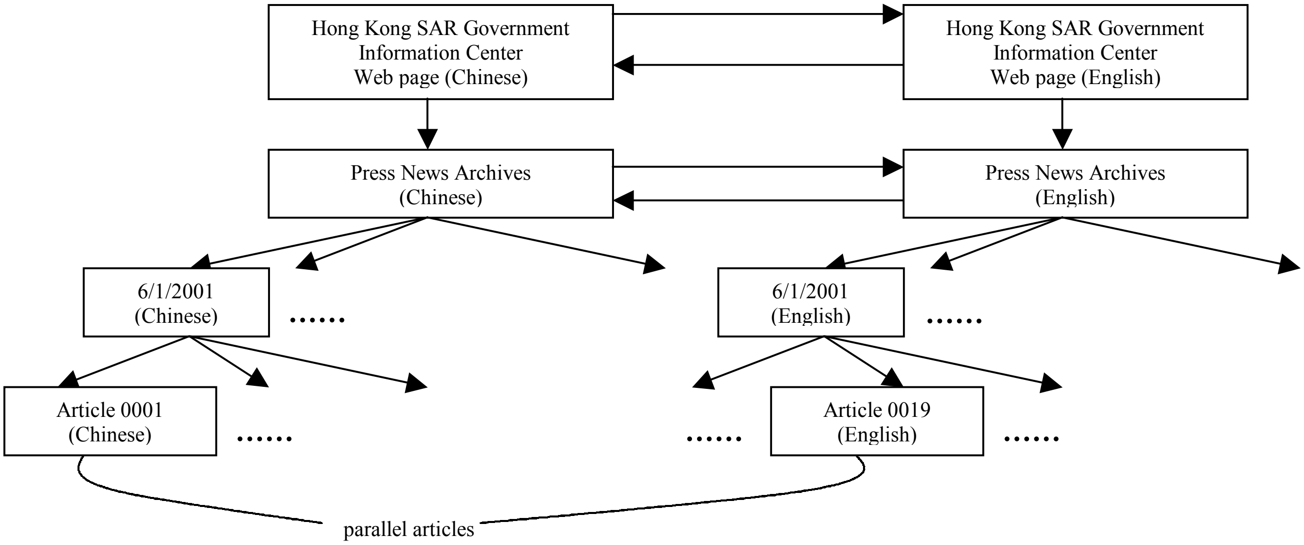

Hong Kong SAR Government Information Center Web page (Chinese)
Hong Kong SAR Government Information Center Web page (English)
Press News Archives (Chinese)
Press News Archives (English)
6/1/2001 (Chinese)
6/1/2001 (English)
Article 0001 (Chinese)
Article 0019 (English)
parallel articles

Figure 3. Organization of Hong Kong SAR Government's press release articles in the Hong Kong SAR Government Information Center Web site.

The result is as shown in Table 1.

Table 1. Precision, Recall and F-measure of the automatically generated parallel documents

Sources	**Precision**	**Recall**	**F-measure**
Hong Kong SAR Government Press Release articles	0.995	0.8096	0.8928

5. Conclusion

In this paper, we have developed an automatic English/Chinese parallel documents construction system to align parallel documents that are organized as separate monolingual sub-tree structure on the Web. Such parallel documents are desired because the increasing demands of cross-lingual information retrieval and the unsatisfactory performance of general-purpose dictionary. Longest common subsequences, edit operations, and score functions based on matching ratio are adopted in our algorithms. Experimental result shows that high precision and acceptable recalls are obtained. A domain specific dictionary can increase the recall. The high precision makes the automatically generated parallel documents a promising tool for other English/Chinese cross-lingual information retrieval applications.

6. Acknowledgement

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

References

1. Allison, L., “Information-Theoretic Sequence Alignment,” Technical Report 98/14 School of Computer Science and Software Engineering, Monash University, 1998.
2. Ebeling, J., “Contrastive Linguistics, Translation, and Parallel Corpora,” *Meta*, Vol 43, Issue 4, 1998, pp.602-615.
3. He, S., “Translingual Alteration of Conceptual Information in Medical Translation: A Cross-Language Analysis between English and Chinese,” *Journal of the American Society for Information Science*, Vol. 51, No. 11, 2000, pp.1047-1060.
4. Leonardi, V., “Equivalence in Translation: Between Myth and Reality,” *Translation Journal*, Vol. 4, No.4., 2000.
5. Macklovitch, E., & Hannan, M., “Line'Em Up: Advances In Alignment Technology And Their Impact on Translation Support Tools,” *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96)*, Montréal, Québec, 1996.
6. Oard, D. W., & Dorr, B. J., “*A Survey of Multilingual Text Retrieval*,” UMIACS-TR-96-19 CS-TR-3815, 1996.
7. Resnik, P., “Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text,” Farwell D., Gerber L., and Hovy E. (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, 1998.
8. Resnik, P., “Mining the Web for Bilingual Text,” *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, Maryland, 1999.
9. Rose, M. G., “Translation Types and Conventions,” *Translation Spectrum: Essays in Theory and Practice*, Marilyn Gaddis Rose, Ed., State University of New York Press, 1981, pp.31-33.
10. Zanettin, F., “Bilingual comparable corpora and the training of translators,” Laviosa, Sara. (ed.) *META, 43:4, Special Issue. The corpus-based approach: a new paradigm in translation studies*: 616-630, 1998.