Discovering Parallel Text from the World Wide Web

Jisong CHEN[1], Rowena CHAU[2], Chung-Hsing YEH[3]

School of Business Systems, Faculty of Information Technology

Monash University, Clayton, Victoria 3800, Australia

[1]Jisong.Chen@infotech.monash.edu.au, [2]Rowena.Chau@infotech.monash.edu.au,
[3]ChungHsing.Yeh@infotech.monash.edu.au

Abstract

Parallel corpus is a rich linguistic resource for various multilingual text management tasks, including cross-lingual text retrieval, multilingual computational linguistics and multilingual text mining. Constructing a parallel corpus requires effective alignment of parallel documents. In this paper, we develop a *parallel page identification system* for identifying and aligning parallel documents from the World Wide Web. The system crawls the Web to fetch potentially parallel multilingual Web documents using a Web spider. To determine the parallelism between potential document pairs, two modules are developed. First, a filename comparison module is used to check filename resemblance. Second, a content analysis module is used to measure the semantic similarity. The experiment conducted to a multilingual Web site shows the effectiveness of the system.

1 Introduction

Parallel document is a rich linguistic resource for various multilingual text management tasks, including cross-lingual text retrieval, multilingual computational linguistics and multilingual text mining. Corpora of parallel documents have been proven to be a crucial lexical basis for constructing robust multilingual linguistic knowledge bases, such as translation model [1,2] and multilingual thesaurus [3]. With the rapid growth of the World Wide Web, electronically accessible information is now available in an ever-increasing number of languages. A recent survey predicts that by 2005, more than 50% of Web contents will be in languages other than English. With such a vast linguistic content, the Web encompasses a large amount of multilingual documents thus forming a rich text archive for building parallel corpora.

Multilingual documents can be collected by crawling the Web using a Web spider. Once a set of multilingual Web documents has been gathered, the first task is to correctly identify pairs of parallel documents from the set. It is in this very first task in which the objective of this paper lies. The result delivered by this paper is therefore playing an essential role in obtaining a corpus of parallel Web documents.

In the context of this paper, a parallel corpus represents a collection of documents together with their translated version in another language. Parallel Web pages, also known as bitexts [4], are translated versions of identical text. A Web page, as well as a Web document, refers to as a single page of information posted on the Web, or a single page within a larger site [5]. A Web spider is a software tool that traverses the Web to gather Web pages by following hyperlinks.Identifying valid parallel pages is a crucial task to ensure the effectiveness of the parallel corpus. To deal with the task, a *Parallel Text Identification System* (PTI) is developed. To illustrate this process, we apply our system to the Hong Kong government Web site, http://www.info.gov.hk, which is a multilingual Web site containing documents in both English and Chinese. The system is readily applied to exploit multiple Web sites and can be extended to the entire Web.

2 The Parallel Text Identification System

The Parallel Text Identification System (PTI) is developed to facilitate the construction of parallel corpus by aligning pairs of parallel document from a multilingual document collection. The architecture of the PTI system is depicted as in Figure 1.

The PTI system crawls the Web to fetch parallel multilingual Web documents using a Web spider. To determine the parallelism between potential bilingual document pairs, two different modules are developed. A filename comparison module is used to check filename resemblance. A content analysis module is used to measure the degree of semantic similarity. It incorporates a novel content-based similarity scoring method for measuring the degree of parallelism for every potential document pair based on their semantic content using a bilingual wordlist.

Multilingual Web documents retrieved by the Web spider are first passed to the filename comparison module to undergo a filename comparison process. Any two Web pages in two different languages with their corresponding filenames resemble each other are picked up and aligned to form a parallel document pair. Multilingual Web documents that remain unaligned after

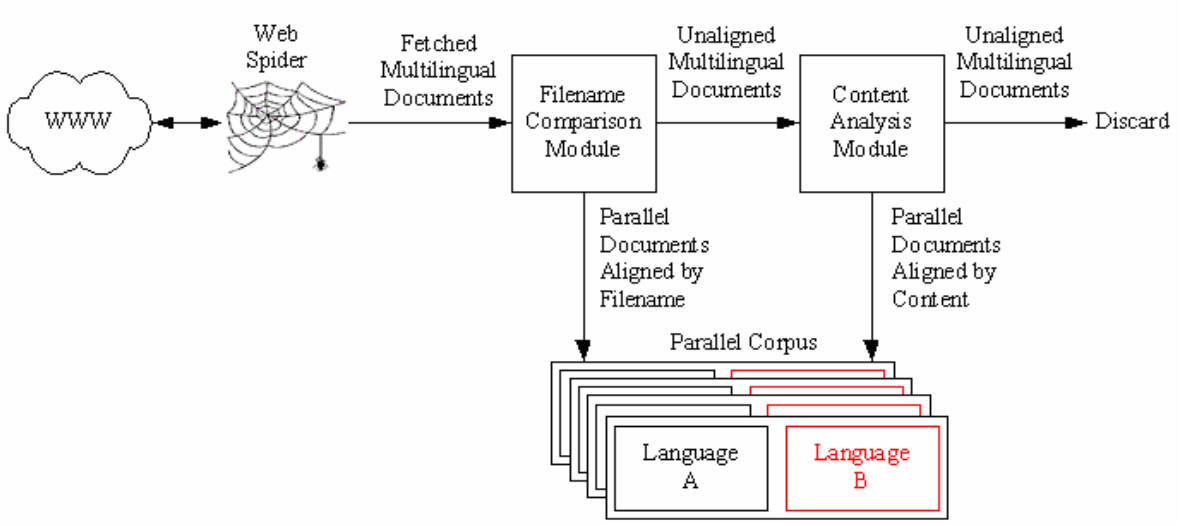

Figure 1. Architecture of the Parallel Text Identification System (PTI)

the filename comparison process will be passed to the content analysis module to carry out semantic content analysis. By calculating a content similarity score for each potential bilingual document pair, parallel documents with identical content will be identified by their high similarity score. Based on this similarity score, parallel documents unable to be detected by their filenames can now be aligned based on their semantic content. Finally, parallel document pairs aligned by either the filename comparison module or the content analysis module are then archived to form a parallel corpus. Such a parallel corpus will provide a valuable linguistic resource for exploiting in-depth multilingual linguistic knowledge to facilitate various multilingual text management tasks. Documents failed to demonstrate any evidence of parallelism after passing through all modules of the system will be discarded.

3 The Filename Comparison Module

Parallel Web documents usually reside in the same Web site with similar filenames [6]. This is a common practice by the Webmaster of a multilingual Web site to keep track of the files by languages. These filenames usually comprise a common sub-string indicating the parallelism of two Web pages, together with another sub-string acting as a language flag indicating the language version of each particular document. Such sub-strings representing the languages are commonly appended to the front, in the middle, or at the end of the common sub-string, which are shared by a pair of parallel documents. Moreover, these language flags are also frequently joined to the common sub-string with a connector such as an hyphen '-' or an underscore '_'. For example, when an English document called "document-eng.htm" is created, its translated version in Chinese will highly likely be named as "document-chi.htm" to make their parallelism explicit for easy Web site management. In other situations, the language flag is only appended to filenames of documents in one particular language. For example, if a document in English called "document.htm" was first created, its Chinese translation generated sometimes later may be named "document-c.htm" just to differentiate their language versions. All these observations suggest that parallel Web documents being maintained following such a naming convention can be effectively traced with the filename comparison.

In the PTI system, the filename comparison module is developed to detect parallel Web documents based on their filename resemblance. To do so, a set of parallel language-indicating sub-strings, such as e↔c, _e↔_c, en-↔ch-, en↔" ", eng↔chi are first generated. These sub-strings are then used to check the existence of similar filenames with matching the corresponding language flags. When the filename of a Web page in one language is encountered, the filename comparison module will generate a list of possible filenames for the translation of this document by appending or removing a prefix, a suffix, or an infix, according to the mapping of the language flags defined previously. A document in the target language with a matching filename is then considered to be parallel. These two documents are thus aligned to form a parallel document pair, which will be archived as a constituent part of a parallel corpus.

For parallel documents being given informative filenames, the filename comparison can be used to identify them effectively. However, parallel documents are not necessarily named consistently with a strict naming convention. Very often, parallel documents are associated to unrelated filenames. To detect this kind of parallel Web documents, their semantic content need to be examined. This task is handled by the content analysis module of the PTI system.

4 The Content Analysis Module

To align parallel documents unable to be detected by the filename comparison, we propose a approach for identifying/aligning parallel text based on their semantic contents using a bilingual wordlist. This is based on the intuition that parallel documents, which are the

translation of each other, should share identical semantic content. Two major steps are involved in this process: multilingual indexing and content similarity scoring.

4.1 Multilingual Indexing

To detect pairs of parallel documents, automatic indexing techniques from information retrieval will be applied. To index documents, a set of content-bearing terms must be first extracted from the text. Index terms can be extracted from a document by either referring to a domain relevant dictionary or by some methods of automatic indexing. In English, automatic indexing is achieved by the elimination of stop words, stemming, etc. For Asian languages, such as Chinese, N-gram based segmentation approach is commonly applied [7]. Although an automatic indexing method based on such term extraction techniques tends to be more exhaustive, the results are noisy. For the discovery of parallel documents that aims at recognising the meaningful semantic correspondence among legitimate content-bearing multilingual terms, the dictionary-based approach to multilingual document indexing is preferred. To this end, a bilingual wordlist is used.

One may argue that when a bilingual wordlist is available for a pair of languages, the need for constructing a parallel corpus for these two languages is no longer justified. The point is that a term-term correspondence between two languages as encoded in a bilingual wordlist is too shallow and is insufficient for many multilingual text management tasks. For instance, information about the multilingual broader/narrower/related terms that are required for query refinement in a cross-lingual text retrieval application is not provided by a bilingual wordlist. A deeper multilingual linguistic knowledge can only be exploited by analysing lexical statistics of well-aligned parallel corpus. The application of parallel corpora is not limited to building a bilingual lexicon. The use of parallel corpora as a linguistic resource for discovering various in-depth multilingual linguistic knowledge is far beyond this. Moreover, the bilingual wordlist in electronic format are commonly available due to the advance of data storage technology nowadays. It is reasonable to use such a handy bilingual wordlist to facilitate the construction of the far more resourceful parallel corpora.

Based on the bilingual wordlist, terms in each document are extracted and their frequencies are counted. Then each document will be represented by a document vector. In this document vector, terms which are the translation of each other will be regarded as the same feature and therefore be assigned the same feature ID (i.e. term ID). Therefore, parallel documents, which are often indexed by the corresponding sets of translated index terms, will be represented by identical document vectors. By computing a similarity score between all document vectors, the document pair exhibiting the highest similarity score should be considered a pair of parallel documents. To achieve this, an appropriate similarity measure is required.

4.2 Content Similarity Scoring

Given M pairs of translated terms and N documents, they form an $M \times N$ term-document matrix where each row corresponds to the term vector shared by a pair of translated terms and each column is a document vector. The feature value of a document vector in the mth row corresponds to the frequency/weight of the mth term in that document. Given this term-document matrix, a similarity score between each pair of documents can be computed.

For any two objects drawn from the same feature space, their similarity can be measured by a distance coefficient or an association coefficient defined on pairs of objects. A variety of distance and association coefficients for comparing textual data are discussed by Salton [8].

To compare the similarity between two objects, $X=(x_1, x_2, \ldots, x_p)$ and $Y=(y_1, y_2, \ldots, y_p)$ drawn from a p-dimensional feature space, the most popular distance coefficient is the Euclidean distance.

$$d(X,Y)=\|X-Y\|=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2} \tag{1}$$

The Euclidean distance has an intuitive appeal as it is useful for evaluating the similarity of objects in a multi-dimensional space. However, a major limitation of the Euclidean distance in the text-based similarity measurement is that it can lead to two objects being regarded as highly similar to each other, despite the fact that they share no features at all in common [9].

There are four commonly used association coefficients suitable for measuring the textual data similarity, namely, the inner product, the cosine coefficient, the Dice coefficient and the Jaccard coefficient.

$$inner\ product=\sum_{i=1}^{p} x_i \cdot y_i \tag{2}$$

$$\cos ine\ coefficient=\frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} x_i^2+\sum_{i=1}^{p} y_i^2}} \tag{3}$$

$$Dice\ coefficient=\frac{2\sum_{i=1}^{p} x_i y_i}{\sum_{i=1}^{p} x_i^2+\sum_{i=1}^{p} y_i^2} \tag{4}$$

$$Jaccard\ coefficient=\frac{\sum_{i=1}^{p} x_i y_i}{\sum_{i=1}^{p} x_i^2+\sum_{i=1}^{p} y_i^2-\sum_{i=1}^{p} x_i y_i} \tag{5}$$

The inner product is the sum of the products of the weights for all the corresponding features. For this reason, it is in principle unbound. As such, when the normalised association coefficient whose values vary between 0 and 1 is required, the other three are preferred.

The cosine coefficient measures the similarity between two objects by calculating the cosine of the angle between their corresponding vectors in the vector space. It is equivalent to the inner product between the vectors if the vectors are normalised to the unit length. The cosine coefficient is sensitive to the relative importance of features, such as which is the strongest feature, and which is the weakest. This characteristic has made the cosine coefficient popular for comparing textual objects in terms of their contextual similarity. The Euclidean distance and the cosine coefficient often disagree as to which object pairs are the most similar and the least similar, since vectors may be very distant from each other (owing, for example, to vastly different lengths) and yet still have a very small angle between them.

Both the Dice coefficient and the Jaccard coefficient are association measures that ignore joint absences of a feature when measuring similarity. For example, two documents are thus neither more nor less similar because neither is about 'politics' or 'entertainment'. The emphasis instead is placed on those features that occur in at least one of the objects being compared. Jaccard coefficient measures the proportion of features occurring in either object that occurs in both.

For a pair of document to be considered parallel, it is expected that these two documents are containing the two corresponding sets of translated terms (i.e. either document will not have any of its index term being unable to map to its foreign counterpart in the other document), and each corresponding term is carrying an identical contextual significance (i.e. occurring with (almost) identical frequency) in each of the document respectively. For this reason, the Jaccard coefficient is more appropriate for the calculation of the similarity score. Ideally, a pair of parallel documents indexed by two set of direct translation equivalents should have their Jaccard coefficient equals to 1.

Given a set T of N pairs of translated terms t and a collection D of M bilingual documents in two languages E and C. Two documents, $e_i \in M$, which is in language E, and $c_j \in M$, which is in language C, are represented by two different column vectors in an $M \times N$ term-document matrix, respectively, as follows:

$$e_i = (w_{1i}, w_{2i}, \ldots, w_{Mi}) \quad (6)$$

where w_{MN} represents the frequency of term t_M being occurred in document e_N.

$$c_j = (w_{1j}, w_{2j}, \ldots, w_{Mj}) \quad (7)$$

where w_{MN} represents the frequency of term t_M being occurred in document c_N.

To compute the content similarity score of these two documents, a similarity scoring function that measures the semantic similarity between two document vectors is defined as follows:

$$sim(e_i, c_j) = \frac{\sum_{k=1}^{M} w_{ki} \cdot w_{kj}}{\sum_{k=1}^{M} w_{ki}^2 + \sum_{k=1}^{M} w_{kj}^2 - \sum_{k=1}^{M} w_{ki} \cdot w_{kj}} \quad (8)$$

These similarity scores will lie between 0 and 1 and depend on the proportion and the occurrence of matching translations in the document vectors. Hence, a pair of parallel documents indexed by two sets of direct translation equivalents respectively should have their similarity score being 1.

5 Experiment

Our purpose is to automatically detect and align pairs of English-Chinese Web documents which are the direct translation of each other. Using a Web spider, called WebZip, 427 Web documents in both English and Chinese are fetched from the Hong Kong Government Web site at http://www.info.gov.hk. These documents are sent to the PTI system to undergo the document alignment. The alignment result will be evaluated against these documents' actual parallelism that has already been indicated by the Webmaster of the Hong Kong government Web site.

Most of the parallel documents we have downloaded from the Hong Kong government Web site are news articles reporting identical news event in both English and Chinese. Each news article is pointed to by an anchor text representing the news headline. For a pair of parallel news, anchor texts pointing to the corresponding language versions are in fact the translation of each other. Therefore, news articles that are parallel to each other can easily be identified by examining their anchor texts. Besides parallel news articles, our experimental document collection also includes parallel documents where each language version is associated with a language-specific out-link, such as 'English' or 'Chinese, pointing to its cross-lingual counterpart. These parallel documents can easily be traced by following their corresponding hyperlinks. As such, all parallel documents pairs existing in our experimental document collection are manually aligned to provide a basis for our system evaluation. Among the 427 documents, 187 pairs of them are found to be truly parallel.

To conduct the experiment, all 427 documents are first sent to the filename comparison module. By this module, 79 pairs of parallel documents are identified and aligned. Those 299 documents left unaligned after the filename comparison process are then sent to the content analysis module.

Theoretically speaking, every pair of parallel documents should exhibit a similarity score equaling to one. However, this is only true if the parallel documents are translated on a word-to-word basis. Daily documents existing in several language versions usually are not

translated in such a strict fashion. The variations of language usage may result in a similarity score of less than one. To take such language usage variations into account while still maintaining reasonable alignment accuracy, we impose an arbitrary threshold of 0.9 on the similarity score. As such, documents with a high content similarity but considerable language usage variations will be picked up. Hence, the parallel corpus archived in this way will be composed of documents that are highly parallel in content while linguistically diverse enough to capture the most prominent variations in real language usage. This linguistic diversity is essential to make a parallel corpus a rich linguistic resource to facilitate an in-depth multilingual lexical analysis that goes beyond the rigid cross-lingual word-to-word mapping. After passing through the content analysis module, 114 pairs of bilingual documents are aligned.

To evaluate the effectiveness of the PTI system, two evaluation measures are defined as follows:

$$\text{Recall} = \frac{\text{No. of pairs being aligned that are parallel}}{\text{Total no. of parallel pairs in the collection}} \quad (9)$$

$$\text{Precision} = \frac{\text{No. of pairs being aligned that are parallel}}{\text{Total no. of pairs being aligned}} \quad (10)$$

The evaluation result shows that the PTI system achieves a recall rate of 0.96 and a precision rate of 0.93. 180 out of 193 pairs of documents aligned by PTI are truly parallel. In other words, the PTI system has failed to detect 7 pairs of parallel documents while it has also erroneously aligned 13 pairs of non-parallel documents. After a close examination of these documents, it is found that those 7 pairs of parallel documents being missed out by PTI are loosely translated. Their extent of parallelism is quite low even though they are semantically similar. On the other hand, 10 document pairs, which are non-parallel, are being considered parallel by the PTI system are found to be the documents coming from a highly similar context. Such document pairs mostly contain highly parallel translated index terms but are in fact describing different events. Moreover, 3 pairs of bilingual documents have also been wrongly aligned based solely on the filename comparison. After all this analysis, the performance achieved by the PTI systems can still be considered promising. This is partially due to the limited domain we are dealing with. To further study the effectiveness of this method on more diversified document collection, a larger scale experiment on documents gathered from various heterogeneous multilingual Web sites will be conducted in future research.

6 Conclusion

In this paper, we have developed an automatic parallel text identification system for aligning parallel Web documents, which are direct translations of each other. Such parallel text identification is significant due to the increasing demand of parallel corpus to be used as a linguistic resource for exploiting in-depth multilingual lexical knowledge for various fast emerging multilingual text management applications. The filename comparison and content similarity scoring are implemented to detect parallel documents based on filename resemblance and semantic closeness. Experimental results obtained from collection of multilingual Web documents show both high precision and recall. This demonstrates the effectiveness of our PTI system as an automatic means for constructing a parallel Web corpus.

7 References

[1] Littman, M. L., Dumais, S., and Landauer, T. K. (1998) Automatic cross language information retrieval using latent semantic indexing. In Grefenstette, G. (ed.) *Cross-Language Information Retrieval*. Chapter 5. Kluwer Academic Publishers, Boston.

[2] Carbonell, J. G., Yang, Y., Frederking, R. E., Brown, R. D., Geng, Y. and Lee, D (1997) Translingual information retrieval: a comparative evaluation. In Pollack, M. E. (ed.) IJCAI-97 *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pp.708-714.

[3] Chau, R. and Yeh, C-H. (2001) Construction of a fuzzy multilingual thesaurus and its application to cross-lingual text retrieval. In N. Zhong, Y. Yao, Liu, S. Ohsuga (Eds.): *Web Intelligence: Research and Development*, First Asia-Pacific Conference, WI 2001, Maebashi City, Japan, October 23-26, 2001, Proceedings. *Lecture Notes in Artificial Intelligence*. Springer-Verlag. Germany. pp. 340-345.

[4] Resnik, P., Smith, N. A. (2002) The Web as a Parallel Corpus. *Technical Report UMIAC-TR-2002-61*, MD: University of Maryland.

[5] Stepforth. Glossary. [Online] Available URL: http://www.stepforth.com/faq/glossary.htm

[6] Chen, J., Nie, J.-Y. (2000) Parallel Web Text Mining for Cross-Language IR. In *Proceedings of RIAO-2000: "Content-Based Multimedia Information Access"*, Paris.

[7] Nie, J. and Ren, F. (1999) Chinese information retrieval: characters or words? *Information Processing and Management*, 35, 443-462.

[8] Salton, G. (1989) Automatic Text Processing: The Transformation, analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading. MA.

[9] Willert, P. (1988) Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, 24(5), pp.577-597.