

Generating Cross-lingual Concept Space from Parallel Corpora on the Web

Christopher C. C. Yang and Kar Wing Li
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China
{yang, kwli}@se.cuhk.edu.hk

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 and generate a cross-lingual concept space automatically for cross-lingual information retrieval. The alignment method is developed based on dynamic programming to identify the one-to-one Chinese and English title pairs for building parallel corpus. The Hopfield network is then employed to generate the cross-lingual concept space based on the statistical correlation analysis of the semantics (knowledge) embedded in the bilingual press release corpus. The research output consisted of a thesaurus-like, semantic network knowledge base, which can aid in semantics-based cross-lingual information management and retrieval.

1. INTRODUCTION

As the Web-based information systems emerge, searching information on the World Wide Web is on high demand, especially the demand of searching across language boundaries. This highlights the importance to develop a tool to refine a query in cross-lingual information retrieval. The major difficulties to retrieve relevant information are the lack of explicit semantic clustering of relevant information and the limits of conventional keyword-driven search techniques [1]. The traditional approaches normally require a document to share some keywords with the query. In reality, it is known that the users may use some keywords that are different from what used in the documents. There are then two different terms spaces, one for the users, and another for the documents. How to create relationships for the related terms between the two spaces is an important issue. The problem can be viewed as the creation of a concept space to cluster terms of similar concepts. The creation of such relationships would allow the system to match queries with relevant documents, even though they contain different terms.

In this paper, we present the construction of the Chinese-English cross-lingual concept space by using Hopfield network based on parallel corpora. Such concept space is important for solving vocabulary difference problem in cross-lingual information retrieval. Since our approach is developed based on the dynamic parallel corpora extracted from the World Wide Web and the information on the Web is frequently updated, the concept space generated can identify unknown terms that do not appear in dictionaries.

2. CONSTRUCTING PARALLEL CORPORA

Parallel corpora can be generated using *overt translation* or *covert translation*. The overt translation [3] 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) [9]. The covert translation [3] is non-directional. Multilingual documents expressing the same content in different languages are generated by the same source [2], e.g. press release from the government, commentaries on a sports event broadcast live in several languages by a broadcasting organization.

There are three major structures of parallel documents on the World Wide Web, *parent page structure*, *sibling page structure*, and *monolingual sub-tree structure*. The monolingual sub-tree structure 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 [4]. Such structure is usually adopted by parallel corpora generated by covert translation. Press releases from the governments and organizations are generated in different languages for the same content independently using covert translation. As a result, the monolingual sub-tree structure is often used.

Alignment methods are required to map the parallel documents organized in monolingual sub-tree structure since links from the documents no longer provide any information of their counterparts. Length-based approach is typical for aligning bilingual documents. However, it is not practical for English/Chinese parallel documents since these languages are significantly different in grammar and structure. We have developed a text-based approach using the longest common subsequence (LCS) to optimize the alignment of English and Chinese titles [8]. Experiment results show that precision of 0.995 and recall of 0.8096 are achieved.

3. GENERATING CONCEPT SPACE

The automatic Chinese-English concept space generation system consists of three components: i) English phrase extraction, ii) Chinese phrase extraction, and iii) Hopfield network. The Chinese and English phrase extraction identifies important conceptual phrases in the corpora. The Hopfield network generates the cross-lingual concept space with the Chinese and English important conceptual phrases as input.

3.1 English and Chinese Phrase Extraction

The English term segmentation is developed based on Salton [6] approach using stop-word, stemming and term-phrase formation. A stop-word list is used to remove non-semantic bearing words such as the, a, on, in, etc. The Chinese term segmentation is developed based on our previous developed technique, boundary detection [7], since there are not any natural delimiters in Chinese sentences to mark work boundaries.

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

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

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j} \times w_j\right) \quad (1)$$

where df_j is the number of documents containing term j . w_j is the length of term j . For an English term, the length of it is the number of words in it. For a Chinese term, the length of it is the number of characters in it.

Asymmetric co-occurrence function [1] is then used to evaluate the relevance weights among concepts. The co-occurrence weight is computed as follows:

$$d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_j\right) \quad (2)$$

The co-occurrence weight, d_{ijk} , is the weight between term j and term k that are both exist in document i . tf_{ijk} is the minimum between occurrence frequency of term j and that of term k in document i . The weight will be zero if neither term j or term k exists in the document.

The relevance weight is a measure of the association between two terms in a collection.

$$Weight(T_j, T_k) = \frac{\sum_{i=1}^n d_{ijk}}{\sum_{i=1}^n d_{ij}} \times WeightingFactor(T_k) \quad (3)$$

$$Weight(T_k, T_j) = \frac{\sum_{i=1}^n d_{ikj}}{\sum_{i=1}^n d_{ik}} \times WeightingFactor(T_j) \quad (5)$$

$$WeightingFactor(T_j) = \frac{\log \frac{N}{df_j}}{\log N} \quad (4)$$

$$WeightingFactor(T_k) = \frac{\log \frac{N}{df_k}}{\log N} \quad (6)$$

3.2 The Hopfield Network Algorithm

To generate the cross-lingual thesaurus, the Hopfield network is modeled as an associate network and transforms a noisy pattern into a stable state representation. The synaptic weights in the storage phase are generated by the co-occurrence analysis. In the canonical Hopfield Networks, if two nodes behave similarly in a sample pattern, the weight between these nodes is usually adjusted with a higher value. Similarly, the relevance weights that computed by Equation (3) and (5) are assigned as the synaptic weights since the relevance weights correspond to how these nodes are strongly associated. The higher the relevance weights between two terms, the stronger the corresponding nodes are associated. In the retrieval phase, a searcher starts with an English term. The Hopfield network spreading activation process will identify other relevant English term and gradually converge toward heavily linked Chinese term through association (or vice versa).

4. EXPERIMENTS

In our experiment, 4907 parallel documents were aligned from the press releases of the Hong Kong SAR government Web site. 10906 concepts were extracted from the parallel corpus. A user evaluation with 10 subjects was conducted. 50 test descriptors (25 English descriptors and 25 Chinese descriptors) were randomly selected from 10906 extracted concepts and presented to the subjects. In the recall phrase of the experiment, the subjects were asked to generate as many relevant terms as possible. In the recognition phrase, the test descriptors and the associated concepts generated by the Hopfield network were presented to the subjects. Noise terms are added in order to reduce the bias generated by the subjects on the concept space. The subjects were asked to determine if the associated concepts were relevant or irrelevant to the test descriptor. Measurement of concept precision and concept recall are utilized to assess the performance of the generated concept space. The precision is the number of retrieved relevant concepts judged as relevant by the subjects over the number of total number of retrieved concepts. The recall is the number of retrieved concepts judged as relevant by the subjects over the number of relevant concepts judged and suggested by the subjects. The overall concept precision and concept recall are 0.88 and 0.85 respectively. The concept precision and concept recall of the English concepts are 0.90 and 0.86, respectively. The concept precision and concept recall of the Chinese concepts are 0.89 and 0.87, respectively.

5. CONCLUSION

Cross-lingual information retrieval is important for Web searching as the Web pages in languages other English are growing significantly. In this work, we have developed an automatic generated concepts space to support cross-lingual information retrieval. Parallel corpora are automatically constructed from the World Wide Web. The associations between the extracted English and Chinese terms are determined statistically. The cross-lingual concept space is generated by the Hopfield network. The experiments show that high precision and recall is achieved.

6. REFERENCES

1. Chen, H., Lynch, K. J., "Automatic construction of networks of concepts characterizing document database" *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 5, pp. 885-902, Sept-Oct, 1992.
2. Ebeling, J., Contrastive Linguistics, Translation, and Parallel Corpora. *Meta*, Vol 43, Issue 4, 1998, pp.602-615.
3. Leonardi, V., "Equivalence in Translation: Between Myth and Reality," *Translation Journal*, Vol. 4, No.4, 2000.
4. Resnik, P. "Mining the Web for Bilingual Text," 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), College Park, Maryland, June, 1999.
5. Rose, M. G., "Translation Types and Conventions," In *Translation Spectrum: Essays in Theory and Practice*, Marilyn Gaddis Rose, Ed., State University of New York Press, 1981, pp.31-33.
6. Salton, G., *Automatic Text Processing*. Addison-Wesley Publishing Company, Inc., Reading, MA, 1989.
7. Yang, C. C., Luk, J., Yung, S., Yen, J., "Combination and Boundary Detection Approach for Chinese Indexing," In *Journal of the American Society for Information Science, Special Topic Issue on Digital Libraries*, vol.51, no.4, March, 2000, pp.340-351.
8. Yang, C. C., and Li, K. W., "Mining English/Chinese Parallel Documents from the World Wide Web," *Proceedings of the International World Wide Web Conference*, Honolulu, Hawaii, May 7-11, 2002.
9. 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.