Automatically Extracting Synonymous Relationship of Technical Terms using a Graph-based Method

Akiko AIZAWA, Kyo KAGEURA, Noriko KANDO
National Center for Science Information Systems
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
akiko@rd.nacsis.ac.jp,kyo@rd.nacsis.ac.jp,kando@rd.nacsis.ac.jp

Keita Tsuji
Tokyo University
3-7-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
i34188@m-unix.cc.u-tokyo.ac.jp

Abstract

Multilingual keyword clusters with similar meanings are recognized to be crucial linguistic resources in information retrieval across language or cultural boundaries. However, the construction of such resources could easily become a bottleneck in academic research fields since most of the keywords are domain-specific. Based on the background, we report in this paper a method for automatically generating Japanese and English keyword clusters of technical terms, using bilingual keyword corpus automatically extracted from databases of academic papers.

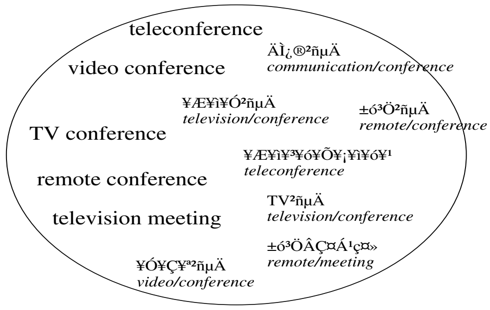

Figure 1: Example of a bilingual keyword cluster.

1 Introduction

The explosive growth of online documents has increased the need for information retrieval (IR) systems that cross language or cultural boundaries. In recent years, the topic has been one of the most actively pursued in IR research field. One especially interesting and important research in relation to this is automatic construction of domain-dependent multilingual thesaurus.

In this paper, we report our attempt to automatically generate *Japanese-English bilingual keyword clusters*, each of which is composed of technical terms with similar meanings from both languages. An example of such keyword cluster is shown in Figure 1 where more than one synonymous words are collected from each language. Our preliminary investigation showed that such keyword clusters not only help searchers of multilingual scientific databases but also be utilized in expanding query terms in monolingual search.

The information source used in our study is one-to-one correspondences of Japanese-English keyword translation pairs automatically extracted from the target databases. We have observed that the extracted translation pairs, also referred to as *bilingual keyword corpus* in this paper, possibly maintain very high degree of precision, which means considerable amount of information useful for IR is embedded in low frequency pairs. Some of the keyword pairs extracted from our keyword corpus is shown in Table 1. Note that although the majority of the pairs are correct, the corpus does contain a few errors such as $<$ information retrieval, (*keyword*) $>$.

The basic idea of our clustering strategy is to adopt a graph-based approach instead of the statistical ones, which seem dominant in the conventional corpus-based studies. The advantage of the graph theoretic view is that it provides us a way to distinguish improper translation pairs from the correct ones utilizing not only the frequency

Table 1: Example of Japanese-English keyword correspondences extracted from the database.

Japanese keywords	English keywords	frequency
	information retrieval	1
	keyword	39
	keyword	1
	information retrieval	1
	information gathering	4
	information retreival	1
	information retrieval	320
	information search	5
	information gathering	6
	information retrieval	1

statistics but also the graph topological information. This enables us to maintain rich expressional variations and minor examples in the generated cluster. In the following, the outline of our clustering method together with some of the recent research topics is briefly introduced.

2 Keyword Cluster Generation Procedure

In the proposed approach, the bilingual keyword pairs are first transformed into a tangled graph of Japanese and English keywords. Next, significant keywords and their translation links are detected using their frequency information. Then, applying the "minimum edge cut detection algorithm" of the traditional graph theory, improper translation links in the graph are identified and eliminated. Consequently, the sub-graphs of the original keyword graph are partitioned into semantically associated clusters of bilingual keywords [1][3][4][8][9]. The outline of the proposed clustering method is summarized in Figure 2.

Applying the method to the keyword corpus with 60,186 different translation pairs, we have obtained 27,918 keyword clusters. The execution time needed for the calculation was about 4 to 5 minutes using Ultra2 workstation, which shows the applicability of the proposed graph-based method to large scale problems. The detected errors were analyzed manually for the evaluation, and the results showed that only 27 out of the examined 951 pairs are eliminated mistakenly. Based on this and other experimental results including the one using a standard IR test collection, we have concluded that the proposed method successfully generates keyword clusters useful for IR applications. Our current system provides about 270,000 clusters generated automatically from keywords of 350,000 academic papers registered to NACSIS Academic Conference Database. The details of the clustering method and other experimental results are described in references [1]-[10] presented by the authors.

3 Related Research Topics

3.1 Automatic Extraction of Translation Pairs

Though the information source used in our present study initially was the Japanese and English keyword lists assigned to academic papers by the authors, the proposed graph-based method is fairly general and applicable to wide range of roughly aligned bilingual term pairs.

The applicability was confirmed by our recent study where bilingual keyword corpora automatically extracted from Japanese and English titles of academic papers are used for cluster generation [10]. In the study, we first apply morphological and lexical processing to titles in each languages and identify domain specific keyword units. Then, bilingual pairs of these units are automatically extracted from the sentence- (title-) level aligned corpus by adopting a simple statistical weighing scheme as well as some heuristics to reduce noises. The resulting pairs are then used to generate an initial keyword graph, based on which bilingual keyword clusters are generated using the graph-based method we have already explained in the previous section.

The experimental results showed that the quality of the generated clusters are fairly good. In addition, it became clear that the conditions used in cluster generation have significant influence on the performance. We are now improving the results by incorporating dynamic parameter adjustment at the cluster generation stage.

3.2 Context Sensitive Keyword Sense Disambiguation

Although the proposed method primarily uses frequency and graph topological information for the detection of improper translation pairs, semantic disambiguation is still needed when (i) selecting translation links for deletion when more than one candidates exist, and also when (ii) determining the cluster for keywords associated with homonymous words. Since the number of candidates is greatly reduced by the graph-based procedure, the disambiguation process can be more time-consuming.

For this purpose, we have examined a method for calculating association scores between keyword pairs utilizing co-occurrence, or the "context", information of the target corpus [6][7]. In the pro-

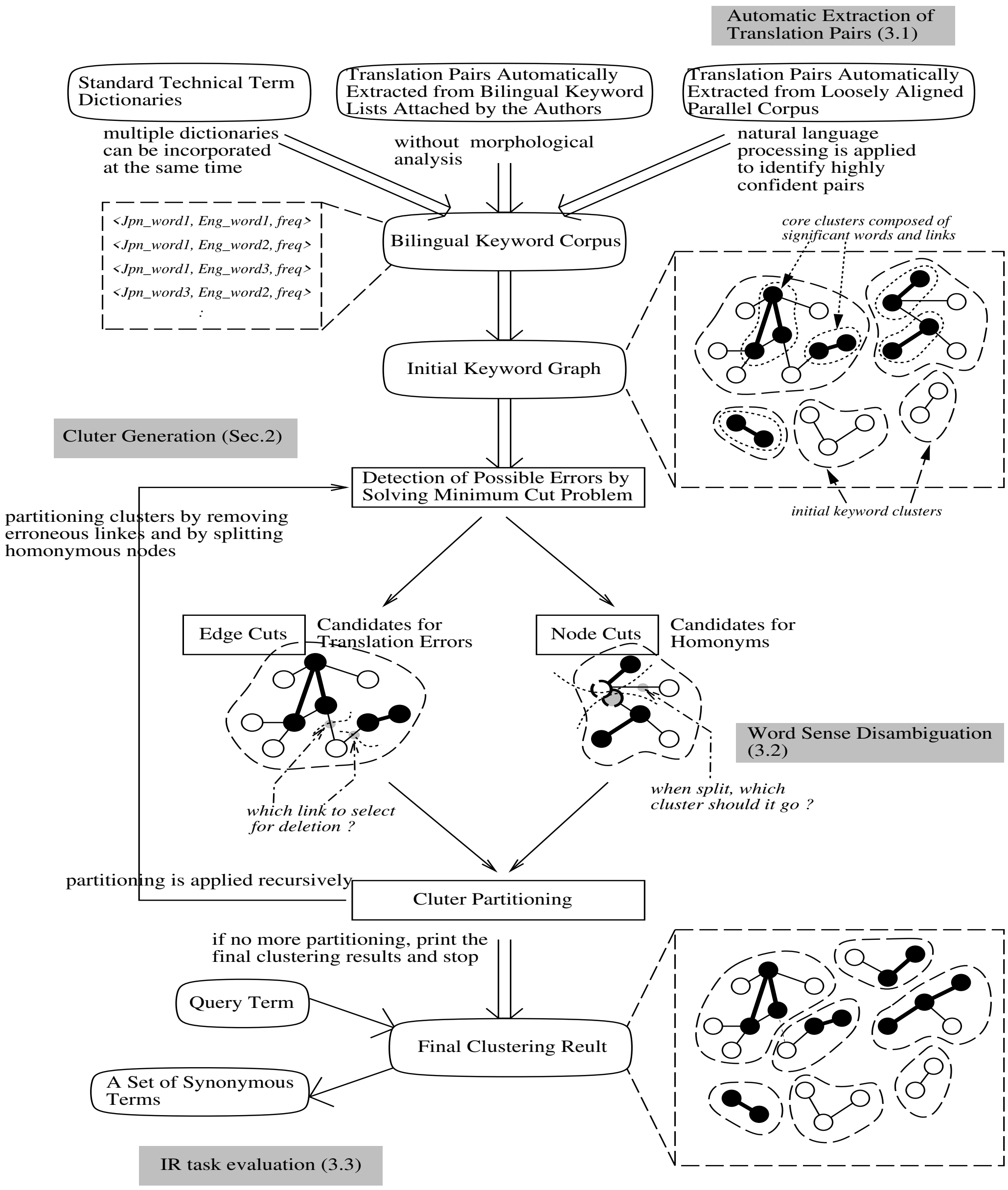

Figure 2: Overview of the proposed keyword clustering procedure.

posed method, a set of keywords that appear in the same text segment is connected by a single association link, which results in the huge and tangled keyword graph similar to the one used in generating bilingual keyword clusters. Since most keywords in the graph are connected to each other in this case, the association score of a given keyword pair can be calculated as the average path length between the selected keywords on the graph.

In the experiments, we have actually generated a large-scale keyword association graph from NACSIS Academic Database using the keyword lists of the papers as text segments. The generated graph was then applied to text classification tasks where the similarity between a document and a category was calculated based on the association score between the keywords on the graph. The result showed that the proposed association measure works effectively compared with other existing measures such as mutual information or context vector, especially when the sparseness problem exists. Though we have not yet integrated the two different types of keyword graphs, the bilingual keyword graph and the keyword association graph, we manually tested a few example cases and confirmed that the proposed association measure is also useful to improve the clustering results.

3.3 Query Expansion in Real Scale Information Retrieval

The generated clusters were also incorporated into the IR system now under development at NACSIS, and the applicability to real-scale IR application was evaluated using (i) the monolingual IR task (J-J task) where Japanese documents are retrieved using Japanese queries, and also (ii) the cross-lingual IR task (J-E task) where English documents are retrieved using Japanese queries. The search performance is tested against the test version of the "NACSIS Test Collection 1" which contains 338,668 Japanese and 186,809 English documents [2][5].

In the evaluation, queries were initially submitted as Japanese natural language sentences. These sentences were first segmented into words using a Japanese morphological analyzer. Then, query terms were automatically identified using several patterns defined over part-of-speech tags. Next, each query term was expanded using the generated bilingual keyword cluster. The retrieval results produced by Opentext 6 showed 11.7% improvement for the J-J task using 11 point average search effectiveness over the *baseline*, which is the performance without query expansion. For the J-E task, the 11 point average performance was 46.6% compared with the case where the original Japanese query terms are applied to the Japanese corresponding of the target English documents. The average term expansion per a query was 10.7 for J-J task and 6.81 for J-E task.

The performance is comparable to the best of the existing cross-lingual IR method without relevance feedback (40 - 60 %). This is considerably good considering that we have not employ any fine-tuning for the target IR task. Based on the results, we conclude that the bilingual clusters generated by our method are quite promising both in cross-lingual IR and monolingual IR. We have also examined the IR performance against different clustering conditions, and it was shown that the graph-topological distance is more effective than the simple frequency information in selecting terms for expansion.

4 Conclusions

In this paper, we have briefly introduced our approach to automatically generate multilingual keyword clusters using bilingual keyword corpus automatically extracted from databases of academic papers. After evaluating the generated clusters by manual analysis and through real-scale IR applications, we recognized that the followings points are important to improve the performance: (i) to adaptively control the clustering conditions depending on the size of the clusters, (ii) to identify the centrality or significance of the keyword within the specified cluster, (iii) to enable incremental update of keyword graph as well as interactive extraction of a cluster which contains a submitted query term. These issues are left for future study.

Acknowledgement

The research reported here is a part of the research project "A Study on Ubiquitous Information System for Utilization of Highly Distributed Information Resources", granted by the Japan Society for the Promotion of Science.

References

[1] Aizawa, A. and Kageura, K. (1998) *An Approach to Automatic Generation of Multilingual Clusters of Technical Terms based on Japanese-English Keywords given by Authors of Academic Papers.* in Proc. of the 4th Annual Meeting of the Association for Natural Language Processing, p.66-69.

[2] Kando, N., Aizawa, A., Tsuji, K., Kageura, K. and Kuriyama, K. (1998) *Cross-Language Information Retrieval and Automatic Construction of Multilingual Lexicons.* 1998 Annual Meeting of the American Society for Information Science.

[3] Aizawa, A., Kando, N., and Kageura, K. (1998) *A Graph-based Method for Automatic Generation of Multilingual Keyword Clusters and Its Applications.* International Joint Workshop on Digital Library (I-JWDL'98).

[4] Aizawa, A. and Kageura, K. (1998) *An Approach to the Automatic Generation of Multilingual Keyword Clusters.* Proceedings of the First Workshop on Computational Terminology (COMPTERM'98), p.8-14.

[5] Kando, N. and Aizawa, A. (1998) *Cross-lingual information retrieval using automatically generated multilingual keyword clusters.* the 3rd International Workshop on Information Retrieval with Asian Languages.

[6] Aizawa, A. and Kageura, K. (1999) *Evaluation of Graph-based Similarity Measure through Automatic Classification of Academic Texts.* Proc. of the 5th Annual Meeting of the Association for Natural Language Processing, p.217-220.

[7] Aizawa, A. and Kageura, K. (1999) *Calculating association between terms utilizing keyword lists of academic papers.* the Special Interest Group Notes of IPSJ, NL-131-8, p.55-62.

[8] Aizawa, A. and Kageura, K. (2000) *Automatic Generation of Clusters of Synonyms Utilizing Japanese-English Keyword Lists of Academic Papers.* Journal of Information Processing Society of Japan (to appear).

[9] Aizawa, A. and Kageura, K. (2000) *An Approach to the Automatic Generation of Multilingual Keyword Clusters.* in Recent Advances in Computational Terminology edited by Bouligault, D., Jacquemin, C. and l'Homme, M-C., John Benjamins (to appear).

[10] Kageura, K., Tsuji, K., and Aizawa, A.(2000) *Automatic Construction of Bilingual Thesauri.* Proc. of the 6th Annual Meeting of the Association for Natural Language Processing (to be presented).