An Approach to the Automatic Generation of Multilingual Keyword Clusters

Akiko N. AIZAWA and Kyo KAGEURA
National Center for Science Information Systems
3-29-1 Otsuka, Bunkyo-ku, Tokyo, 112-8640 Japan
E-Mail: {akiko,kyo}@rd.nacsis.ac.jp

Abstract

In this paper, we report an effective method for generating Japanese and English *bilingual keyword clusters* using the keyword lists assigned to academic papers by the authors where each of the generated clusters contains keywords with similar meanings from both languages. The advantages of the method are that (i) various domain-dependent keyword pairs useful for IR can be automatically extracted and clustered, with (ii) practically reasonable computational cost, and that (iii) low-frequency keywords can be properly treated and maintained for later use in IR applications. We apply the method to a set of Japanese and English keywords extracted from academic conference papers in computer science, and show that our method gives a very promising result.

1 Introduction

Recently, we have witnessed a growing interest in the application of natural language processing methods to extend and enhance the function and performance of IR systems. Since the first workshop on cross-lingual information retrieval in 1996 (Grefenstette, Smeaton & Sheridan 1996), one of the most actively pursued topics in this area has been the automatic identification of bilingual lexical pairs. The topic is of special interest and importance in scientific IR, where multilingual databases are growing rapidly and the construction of domain-dependent lexical linguistic resources has become a bottleneck (e.g. Dagan & Church 1994; Daille, Gaussier, & Langé 1994; Frantzi & Ananiadou 1995; Kageura, Tsuji & Takasu 1996).

In this paper, we propose a graph-based approach to the automatic generation of multilingual keyword clusters, each of which consists of keywords with similar meanings from different languages. The keyword clusters generated by our method can be utilized as domain-specific linguistic resources in IR applications, for example, as a dictionary for query expansion. Here, we have specifically noted that keywords attached by the authors to scientific papers constitute an important and useful information source for the following reasons:

(1) The majority of scientific papers published in Japan have (and *are encouraged to have*, to make the academic results available to foreign countries) both Japanese and English keywords given by the authors. Correspondingly, these bilingual keywords are included in databases and are thus readily available for a great many domains.

(2) The correspondences between the Japanese and English keywords are maintained to such a level that a sufficiently large number of bilingual pairs can be extracted mechanically, though there is some accompanying noise.

(3) Various keyword pairs are found that may not be listed in the standard terminological dictionaries, but are nevertheless meaningful and useful in IR; they may be a valuable complement to the existing terminological resources, if the noise is removed properly.

In the next section, we first summarize the nature of the Japanese and English keywords of the academic papers we use in this study.

2 Basic Nature of the Data

2.1 Extraction of Keyword Pairs

The basic data used in the current study is the Japanese and English keywords assigned by the authors to their papers, extracted from the NACSIS Academic Conference Database (NACSIS, 1997). Of the 28,122 papers we selected

from the computer science field, 23,540 (about 84 %) have both Japanese and English keywords. Since the authors are not explicitly required to provide precise conceptual correspondence between the two languages, we first examine the *rough* correspondence (adequacy of the translation is not strictly considered here) of the lists the possibility of automatically extracting sufficiently reliable keyword pairs.

Of the 23,540 papers considered, 1,339 (about 6%) have different numbers of keywords in the two languages. The mismatch is caused by a variety of factors including missing one or more keywords in either of the languages, or the noise of the database such as a single keyword being separated into two fields. On the other hand, the papers with the same number of Japanese and English keywords (about 94%) are most likely to maintain one-to-one correspondence if we simply couple them according to the listed order. For example, we found only a single ordering error in the total 452 keyword pairs automatically coupled from 100 papers randomly selected.

Based on the above observation, we extracted a total of 112,364 keyword pairs (60,186 different ones) using only the papers with equal numbers of Japanese and English keywords, and used them as a basic bilingual corpus throughout the paper. Table 1 shows some examples of the extracted keyword pairs. For each Japanese keyword, the most general English correspondence is marked with '∗'. (only the word <広域情報検索 (*wide-area information retrieval*)> does not have its exact English counterpart .)

2.2 Comparison with the Standardized Technical Dictionaries

Next, we compared the extracted keyword pairs with 22,690 different term pairs obtained from four dictionaries and handbooks in computer science. Table 2 summarizes the result.

The table shows that the number of common elements between our keyword data and the dictionaries is relatively small, i.e. many of the keyword pairs assigned by the data are not listed in the standard technical dictionaries. It can also be observed that the average number of translations is smaller in the dictionaries than in the keyword data. This may be because the dictionaries are to some extent prescriptive while the keyword pairs reflect particular views and concepts of the authors. At the same time, it is also expected that the keyword data contain some noise due to simple input errors, errors of correspondence, related but improperly linked words, etc. Because of these many-to-many bilingual correspondences, the overall set of keyword pairs constitute a complex network with rich correspondences as well as noise.

From the viewpoint of possible application to cross-lingual information retrieval, the following points should be taken into consideration in the processing of the keyword pairs:

(1) In terms of the precision of the search, it is important to eliminate erroneous or improper translations.

(2) In terms of the recall of the search, it is advantageous to maintain expressional variations and (to some extent) minor errors.

Recognizing the nature of the keyword data stated above, we have decided to adopt a graph-theoretic method, instead of the statistical approaches, which seem dominant in corpus-based studies today.

Table 1: Example of keyword pairs.

Japanese keywords	English keywords	frequency
キーワード	information retrieval	1
キーワード	keyword	39*
テキスト検索	information retrieval	1
テキスト検索	text retrieval	6*
テキスト検索	text search	3
検索指示語	keyword	1
広域情報検索	information retrieval	1
情報検索	information gathering	4
情報検索	information retreival	1
情報検索	information retrieval	320*
情報検索	information search	5
情報収集	information gathering	6*
情報収集	information retrieval	1
文献検索	bibliographic search	1*
文献検索	document retrieval	11
文書検索	document retrieval	19*
文書検索	text retrieval	1

Table 2: Comparison of the technical dictionaries and the bilingual keyword data.

	dictionary terms	corpus pairs	common for both
Japanese words	20,636	37,170	3,966
English words	19,562	49,918	2,814
translation pairs	22,690	60,186	2,066
avr.no.of trans.(Jpn)	1.10	1.62	−−
avr.no.of trans.(Eng)	1.16	1.21	−−
max.no.of trans.(Jpn)	7	86	−−
max.no.of trans.(Eng)	6	29	−−

† notation variation not considered here.

3 Related Work

3.1 Identification of Bilingual Lexical Pairs

There have been many recent studies on the automatic identification of translation pairs in

bilingual textual corpora. Methodologically, the studies can be categorized into the following two types. The first is based on summary/test statistic measures such as the mutual information or ϕ^2 (Gale & Church 1991; van der Eijk 1993; Fung 1995; Smadja et al. 1996), and the other is based on the idea of some kinds of iterative optimizations (Kupiec 1993; Melamed 1996).

All of the above techniques assume a bilingual textual corpus only loosely aligned, (for example, at sentence level) and use statistical measures to calculate the most plausible correspondences. As such, they can extract proper pairs only for lexical items with relatively high frequencies. Pairs with less significant occurrences, though they sometimes include useful information such as rare translation examples or allomorphs, are in general neglected. In addition, the processing of a textual corpus needs much computation, even for such simple and easy-to-calculate summary statistic measures as ϕ^2 or mutual information [1].

In contrast, our method, which is based on the bilingual keyword lists with much less noise in their correspondences, is much easier and much less costly to execute. Though the types of pairs obtained from the textual and the keyword corpora do not exactly match, it can be assumed at least that the keywords given by the authors constitute the most important (sub-)set of the possible keyword pairs extractable from the corresponding textual data.

3.2 Automatic Thesaurus Generation

Another important research area related to our study is the automatic generation of thesaurus. Many studies have been carried out recently on this subject mainly using the corpus-based statistical methods (Finch, 1993; Grefenstette, 1994; Hughes, 1994; Schütze & Pederson 1997; Strzalkowski, 1994). One of the most widely known methods in IR is the Latent Semantic Indexing (Deerwester et. al, 1990) where Singular Value Decomposition is applied to a term by document matrix extracted from a textual corpus in order to calculate the similarity between documents and documents, terms and terms, and also documents and terms.

The same idea can also be applied to the task of extracting term clusters from a bilingual textual corpus (Dumais, Landauer & Littman 1996). However, as in the statistical extraction of bilingual term pairs, the method totally depends on the frequency of co-occurrence of the terms and thus is unsuitable to treat low-frequency terms, which include minor but valuable translation examples as well as errors.

[1] in the case of Japanese, morphological preprocessing is essential because Japanese does not have explicit word boundaries.

4 Generation of Multilingual Keyword Clusters

4.1 Initial Keyword Graph

The initial graph expression of a bilingual keyword corpus is easily derived by representing Japanese and English keywords as nodes and their translation pairs as links. The frequency of appearance of each keyword pair in the corpus is expressed as the capacity of the corresponding link. Figure 1 shows the initial keyword graph generated from Table 1.

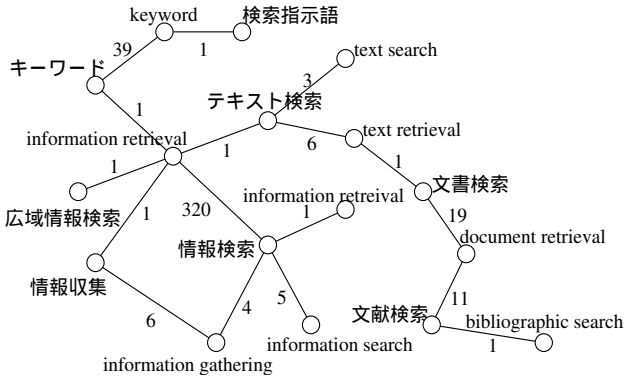

Figure 1: Example of initial keyword graph.

The initial keyword graph is composed of a number of disjoint sub-graphs, which we define as *bilingual keyword clusters*. Thus, by our definition, every pair of nodes connected to each other is regarded as synonymous, regardless of the capacity of the connecting links. For example, in Figure 1, the initial keyword graph contains a single keyword cluster at the initial stage. The detection and deletion procedure described below is applied for each keyword cluster recursively until no more pairs can be removed.

4.2 Detection of Possible Correspondence Errors

The initial keyword graph constructed from all 60,186 different translation pairs contains a huge keyword cluster with as many as 20,659 nodes (about 34%). Since we restrict our subject to a specific academic field *computer science*, the major cause for this is the existence of

improper translation pairs, which connects otherwise disjoint keyword clusters into one [2]. A typical example found in Figure 1 is the link <キーワード (*keyword*) , information retrieval >.

The detection algorithm proposed in this paper is based on a simple principle that *sets of links, which decompose a connected keyword cluster into disjoint sub-clusters when they are removed from the original cluster, are candidates for improper translations.* In conventional graph theory, such a link set is called an *edge cut* and the edge cut with the minimal total capacity among all the edge cuts obtained for a given graph is called a *minimum edge cut.* A minimum edge cut doesn't necessarily imply a single translation error. For example, in the case shown in Figure 2, two links <テキスト検索 (*text retrieval*) , information retrieval> and <情報検索 (*information retrieval*), text retrieval> must be removed simultaneously in order to divide the cluster.

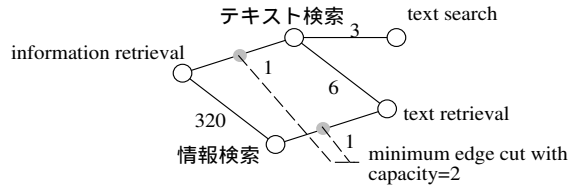

Figure 2: Example of a keyword cluster with two simultaneous translation errors.

The minimum edge cut problem is one of the principal problems in graph theory and has been studied quite extensively since the 1950's (Nagamochi 1993). The complexity of the simple Edmonds & Karp algorithm used in our current implementation is $O(n^2m)$ for the graph with n nodes and m edges, which guarantees sufficient performance for practical use.

4.3 Partitioning Keyword Clusters

The minimum edge cut of a keyword cluster is a candidate for, but does not always represent, imprecise translation. Removing correct pairs inevitably causes oversplitting, i.e. generating more than one cluster with similar meanings.

The distinction between corresponding and associated but non-corresponding keyword pairs depends on the application and is difficult even for human experts. For example, the pair < テキスト検索 (*text retrieval*) ,information retrieval > may be improper in view of strict terminological definition, but not necessarily incorrect for searchers of academic paper databases. Since in our keyword data the frequency of the link directly represents the number of papers with the translation in their keyword fields, it is natural to use that information to avoid oversplitting.

Our procedure first checks links that should *not* be eliminated using the following conditions:

($C1$) the frequency is no less than N_α,
($C2$) the Japanese and English notations are identical,
($C3$) either of the Japanese or English expression has only one corresponding translation (i.e. the link itself).

All the above conditions are easily tested at the pre-processing stage before the minimum-cut detection algorithm is applied. ($C1$) represents the minimum support for a translation pair to be considered correct regardless of the context. In total, 2,723 pairs (about 5%) satisfy this condition for $N_\alpha = 3D5$. ($C2$) refers to proper nouns or abbreviations. A total of 1,336 pairs (about 2%) match this condition. ($C3$) is a condition to guarantee that every cluster contains at least one Japanese and one English keyword. A total of 49,233 pairs (82%) are marked by this condition. Though this percentage seems large, most of them occur with low frequency and do not much affect the performance of IR even if they are actually improper matches.

Next, our procedure partitions the clusters that satisfy both of the following conditions:

($C4$) the total capacity of the minimum cut is less than N_ϵ,
($C5$) for all sub-clusters generated by the minimum cut, there exists at least one translation link whose capacity is no less than N_β

The partitioning is repeated recursively until no cluster exists that satisfies conditions ($C4$) and ($C5$).

To summarize, N_ϵ , N_α , N_β are the three parameters we use in our procedure. Presently, we use the same values for all the clusters.

4.4 Clustering Results

Figure 4 shows the result of the partitioning of the keyword cluster given in Figure 2, where $N_\epsilon = N_\alpha = N_\beta = 5$. Three keyword pairs, <キーワード (*keyword*) , information retrieval>, <テキスト検索 (*text retrieval*) , information retrieval>, <文書検索 (*document retrieval*) , text retrieval>, are removed, and four clusters A, B, C, and D are newly created. The bold lines show that the links are non-removable from ($C3$). It follows that such

[2] This is in contrast to general cases where homonyms are much more common.

pairs as <情報検索 (*information retrieval*) , information retreival> (spelling error), <検索指示語 (*keyword*) , keyword> (rare case), and <広域情報検索 (*wide-area information retrieval*) , information retrieval> (related but not equivalent pair) are retained even after the partitioning.

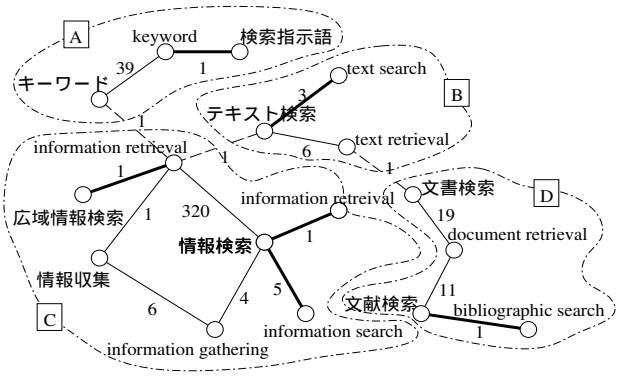

Figure 3: Example of cluster partitioning.

If we let $N_\beta = 10$, then group B cannot be an independent cluster due to $(C5)$, and is merged *either* with group D or C. The ambiguity is caused by the existence of more than one minimum cut in the cluster, and there is no topological cue to decide which link is more plausible for deletion. The group to which B should belong is currently determined either randomly or by similarities calculated from histograms of composing characters, but we are examining the possibility of using information concerning the constituent elements of keywords.

Applying the method to the whole corpus with $N_\epsilon = N_\alpha = N_\beta = 5$, a total of 1,469 of the 60,186 keyword pairs were deleted, generating 27,918 clusters. The run time for this task was about 4 minutes and 30 seconds by Ultra 2 (200MHz), which shows the computational efficiency of our method. The number of keyword pairs included in the biggest cluster was reduced to 159 (43 Japanese and 72 English keywords), from the 20,659 of the initial graph. Example 1 shows the keywords in this biggest cluster with frequencies ≥ 3 (the number in parenthesis indicates the frequency of the keyword).

Example 1: *Keywords in the largest cluster.*
Japanese: 並列処理 *(740)*, 並列 *(62)*, 並列化 *(56)*, 並列計算 *(29)*, 並行処理 *(19)*, 並列性 *(10)*, 並列システム *(6)*, 多重処理 *(6)*, 並行システム *(5)*, マルチプロセス *(5)*, 並列プロセス *(4)*, 並列度 *(3)*, 多重プロセス *(3)*, マルチプロセシング *(3)*, パラレル処理 *(3)*
English: *parallel processing(672), parallel(74), parallelization(44), concurrent processing(20), parallel computing(18), parallel computation(18), parallelism(14), parallel process(8), multi process(8), concurrent system(7), parallel system(6), parallel processings(6), multi processing(6), multiprocessing(5), multiprocess(5), concurrent(5), future(4), parallelize(4), parallel processes(4), parallel operation(4), paralell processing(4)*

As can be seen, the closely related Japanese and English keywords are clustered together while maintaining rich variations inherent in the original academic papers, which is in itself quite promising for IR.

4.5 Remarks

So far, we have primarily focused on the treatment of improper translations in bilingual keyword clusters. Other aspects of the issue are now also remarked upon.

Homonyms

Though homonyms do not seem to occur frequently in a specific scientific domain such as computer science, we have still observed several cases. Possibly homonymous keywords can also be detected utilizing topological features of the keyword cluster. It can be assumed that *homonymous nodes are the ones that decompose the cluster when the node and all the edges starting from the node are removed.* Thus, the problem is transformed again to the well-known node cut problem of graph theory. Once a candidate node for homonym is obtained using existing graph algorithms, partitioning is done by splitting the node into newly generated clusters as shown in Figure 4.

However, homonyms require extra considerations. First, we need deeper lexical knowledge to avoid oversplitting. Since most of the homonyms we have observed are abbreviations, we presently consider only keywords composed of English capital characters and symbols as candidates for node cuts, but this may be insufficient in some cases. Secondly, when a homonymous node is split, every edge starting from the node should be classified into one of the newly generated clusters. However, there exists no topological clue for an isolated edge like <金融自動機 (*automatic teller machine*), ATM> in Figure 4. Thus, the attribution of the isolated pairs must be decided by other means. We presently use notational similarity between keywords, but again, deeper lexical knowledge may be needed for better performance.

Standard Technical Dictionaries

Japanese and English translation pairs extracted from standard technical dictionaries can

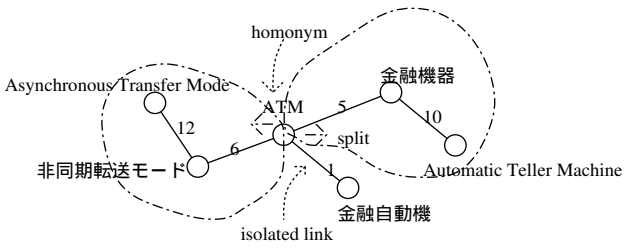

Figure 4: Example of a homonymous keyword.

easily be integrated into the original corpus data (with the frequencies set to infinite) as a complement.

Natural Language Processing

Our present implementation employs only a little language-dependent knowledge, i.e. trivial normalization rules to deal with notational variations, the definition of abbreviations, and the similarity measure based on the histogram of character codes. Though our original intention is the automatic extraction of bilingual keyword clusters for various domains and languages, the standard natural language processing technique can also be incorporated easily.

First, natural language processing can be employed at the screening stage to mark non-removable links. Since this pre-processing stage should not be computation-intensive, simple string matching with existing standard dictionaries, for example, can be a good candidate. Second, natural language processing can be used at the disambiguation stage. Disambiguation includes (i) selecting an edge cut for deletion when more than one candidate exists, and (ii) determining the cluster for the isolated edge of the cluster size and the selected node cut. Since the number of candidates is greatly reduced by graph-theoretic means, this stage can be more time-consuming.

The overall procedure including those discussed here is illustrated in Figure 5.

5 Discussion

Currently, we are refining the basic methods at each stage in Figure 5. For example, we have used only the keyword-level correspondences so far, since Japanese does not have clear word boundaries. Given that many keywords are complex, we can expect more information by utilizing the morpheme- (or word-) level correspondences. Also, we have so far emphasized that an interesting aspect of the clusters is that they include useful but non-standard pairs. In IR, this may well be an advantage, but at the same time, it may be quite useful if we can extract canonical bilingual pairs among the clustered data. The simple co-occurrences already show interesting information, but this will be further refined by considering other factors.

Lastly, we are planning to evaluate the applicability of the generated clusters as a cross-lingual thesaurus for IR systems using the test collection currently under construction (Kageura et al. 1997; Kando et al. 1998). Part of the result will be reported in (Kando and Aizawa 1998).

Acknowledgement

This research is a part of the research project "A Study on Ubiquitous Information Systems for Utilization of Highly Distributed Information Resources", funded by the Japan Society for the Promotion of Science.

References

[1] Dagan, I. and Church, K. (1994) "Termight: Identifying and translating technical terminology." *Proc. of the Fourth ANLP.* p.34–40.

[2] Daille, B., Gaussier, E. and Langé, J. M. (1994) "Towards automatic extraction of monolingual and bilingual terminology." *COLING'94.* p. 515–521.

[3] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990) "Indexing by latent semantic analysis." *JASIS.* 41(6), p. 391–407.

[4] Dumais, S.T., Landauer, T.K., and Littman. M.L. (1996) "Automatic cross-linguistic information retrieval using Latent Semantic Indexing." *SIGIR'96 Workshop on Cross-Linnguistic Information Retrieval.* p. 16–23.

[5] Eijk, van der P. (1993) "Automating the acquisition of bilingual terminology." *Proc. of the 6th EACL.* p. 113–119.

[6] Finch, S. P. (1993) *Finding Structure in Language.* PhD Thesis. Edinbourgh: University of Edinbourgh.

[7] Frantzi, K. T. and Ananiadou, S. (1995) "Statistical measures for terminological extraction." *Proc. of 3rd Int'l Conf. on Statistical Analysis of Textual Data.* p. 297–308.

[8] Fung, P. (1995) "A pattern matching method for finding noun and proper noun translations from noisy parallel corpora." *Proc. of 33rd ACL.* p. 233–236.

[9] Gale, W. A. and Church, K. W. (1991) "Identifying word correspondences in parallel texts." *Proc. of DARPA Speech and Natural Language Workshop.* p. 152–157.

[10] Grefenstette, G. (1994) *Explorations in Automatic Thesaurus Discovery.* Boston: Kluwer Academic.

[11] Grefenstette, G., Smeaton, A. and Sheridan, P. (eds.) (1996) *Workshop on Cross-Linguistic Information Retrieval.* Switzerland, ETH, August 22, 1996.

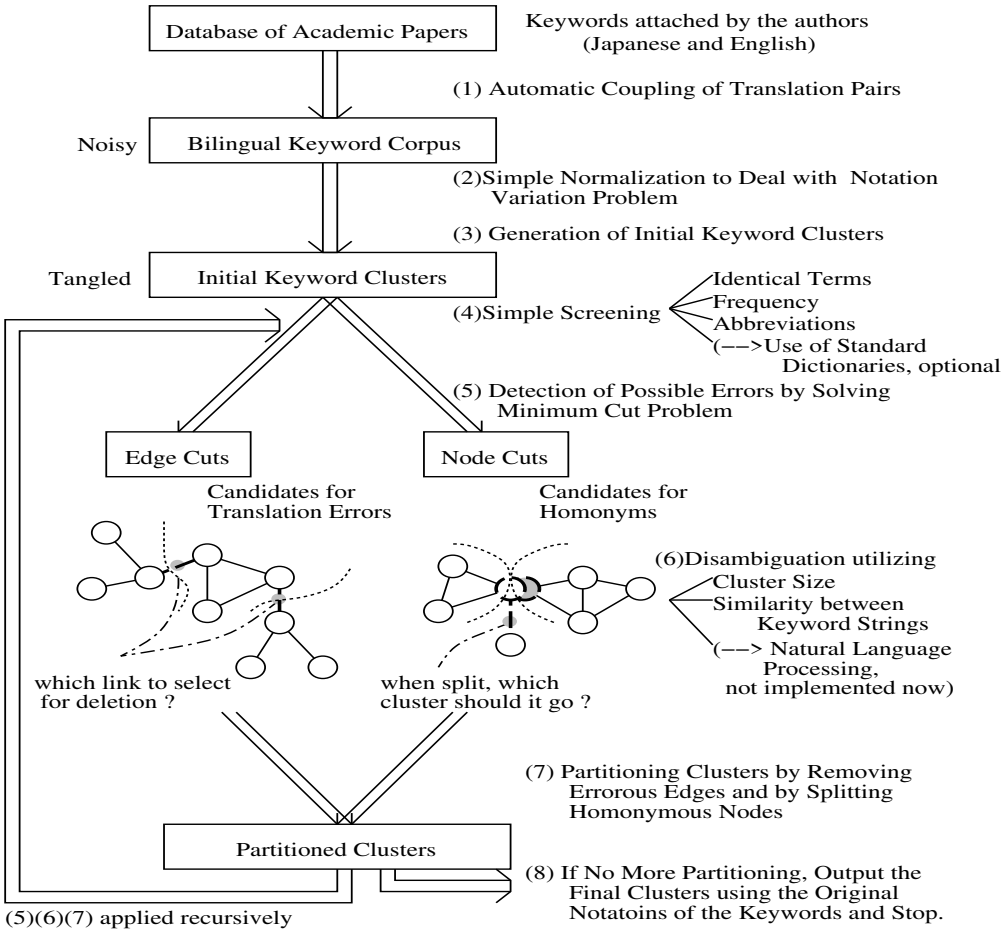

Figure 5: Overview of the proposed keyword clustering procedure.

[12] Hughes, J. (1994) *Automatically Acquiring a Classification of Words.* PhD Thesis. Leeds: University of Leeds.

[13] Kageura, K., Tsuji, K. and Takasu, A. (1996) "Some statistical characterisations of terminological and non-terminological elements." *Proceedings of TKE'96.* p. 131–138.

[14] Kageura, K., Koyama, T., Yoshioka, M., Takasu, A., Nozue, T. and Tsuji, K. (1997) "NACSIS corpus project for IR and terminological research." *NLPRS'97*, p.493-496.

[15] Kando,N., Koyama,T., Oyama,K., Kageura,K., Yoshioka,M., Nozue,T., Matsumura,A. and Kuriyama,K. (1998) "NTCIR : NACSIS Test Collection Project [Poster]." the 20th Annual BCS-IRSG Collquium on Information Retrieval Research.

[16] Kando, N., Aizawa, A., Tsuji, K., Kageura, K., and Kuriyama, K. (1998) "Cross-Language Information Retrieval and Automatic construction of multilingual lexicons." 1998 Anuual Meeting of the American Society for Information Science (accepted).

[17] Kupiec, J. (1993) "An algorithm for finding noun phrase correspondences in bilingual corpora." *Proc. of 31st ACL.* p.17–22.

[18] Melamed, I. D. (1996) "Automatic construction of clean broad-coverage translation lexicons." *2nd Conference of the Association for Machine Translation in the Americas.*

[19] NACSIS (1997) *Introduction to the National Center for Science Information Systems.* Tokyo: NACSIS.

[20] Nagamochi, H (1993) "Minimum cut in a graph." In: Fujisige, S. (ed.) *Discrete Structure and Algorithms II* (Chapter 4). Tokyo: Kindaikagakusha.

[21] Schütze, H. and Pedersen, J.O. (1997) "A cooccurrence-based thesaurus and two applications to information retrieval." *Information Processing and Management.* 33(3), p.307–318.

[22] Smadja, F., McKeown, K. R. and Hatzivassiloglou, V. (1996) "Translating collocations for bilingual lexicons: A statistical approach." *Computational Linguistics.* 22(1), p. 1–38.

[23] Strzalkowski, T. (1994) "Building a lexical domain map from text corpora." *COLING'94*, p.604–610.