ELSEVIER

Information Processing and Management 40 (2004) 145–159

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Implicit ambiguity resolution using incremental clustering in cross-language information retrieval

Kyung-Soon Lee [a,*], Kyo Kageura [a], Key-Sun Choi [b]

[a] *National Institute of Informatics (NII), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan*

[b] *Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, 373-1 Kusung, Yusong, Taejon 305-701, South Korea*

Received 4 April 2002; accepted 3 April 2003

Abstract

This paper presents a method to implicitly resolve ambiguities using dynamic incremental clustering in cross-language information retrieval (CLIR) such as Korean-to-English and Japanese-to-English CLIR. The main objective of this paper shows that document clusters can effectively resolve the ambiguities tremendously increased in translated queries as well as take into account the context of all the terms in a document. In the framework we propose, a query in Korean/Japanese is first translated into English by looking up bilingual dictionaries, then documents are retrieved for the translated query terms based on the vector space retrieval model or the probabilistic retrieval model. For the top-ranked retrieved documents, query-oriented document clusters are incrementally created and the weight of each retrieved document is re-calculated by using the clusters. In the experiment based on TREC CLIR test collection, our method achieved 39.41% and 36.79% improvement for translated queries without ambiguity resolution in Korean-to-English CLIR, and 17.89% and 30.46% improvements in Japanese-to-English CLIR, on the vector space retrieval and on the probabilistic retrieval, respectively. Our method achieved 12.30% improvement for all translation queries, compared with blind feedback for the probabilistic retrieval in Korean-to-English CLIR. These results indicate that cluster analysis help to resolve ambiguity.

Keywords: Implicit ambiguity resolution; Cross-language information retrieval; Incremental clustering; Document context; Document re-rank

[*] Corresponding author.

E-mail addresses: kslee@nii.ac.jp (K.-S. Lee), kyo@nii.ac.jp (K. Kageura), kschoi@cs.kaist.ac.kr (K.-S. Choi).

URLs: http://research.nii.ac.jp/~kslee, http://research.nii.ac.jp/~kyo, http://kschoi.kaist.ac.kr/

0306-4573/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0306-4573(03)00028-1

1. Introduction

This paper describes a method of applying dynamic incremental clustering to resolve query ambiguities implicitly in Korean-to-English and Japanese-to-English cross-language information retrieval. The method uses the clusters of retrieved documents as a context for re-weighting each retrieved document and for re-ranking the retrieved documents.

Cross-language information retrieval (CLIR) enables users to retrieve documents written in a language different from a query language. The methods used in CLIR fall into two categories: statistical IR approaches and translation approaches. Statistical IR methods establish cross-lingual associations without language translation (Dumais, Letsche, Littman, & Landauer, 1997; Rehder, Littman, Dumais, & Landauer, 1997; Yang, Carbonell, Brown, & Frederking, 1998). They require large-scale bilingual corpora. In translation approach, either queries or documents are translated. Though document translation is possible when high quality machine translation systems are available (Kwon, Kang, Lee, & Lee, 1997; Oard & Hackett, 1997), it is not very practical. Query translation methods (Hull & Grefenstette, 1996; Davis, 1996; Gilarranz, Gonzalo, & Verdejo, 1997; Eichmann, Ruiz, & Srinivasan, 1998; Yang et al., 1998; Jang, Myaeng, & Park, 1999; Chun, 2000) based on bilingual dictionaries, multilingual ontology and thesaurus are much more practical. Many researches adopt dictionary-based query translation method because it is simpler and practical, given the wide availability of bilingual or multilingual dictionaries. In order to achieve a high-performance CLIR using dictionary-based query translation, however, it is necessary to solve the problem of increased ambiguities of query terms. To resolve query translation ambiguities from bilingual dictionary, the mutual information method or its variation based on co-occurrence statistics have been suggested (Ballesteros & Croft, 1998; Jang et al., 1999; Gao, Zhou, Nie, He, & Chen, 2002). The translation term pair with the highest value is selected. The co-occurrence information has been used with some success for phrasal translations (Smadja, McKeown, & Hatzivassiloglou, 1996; Kupiec, 1993). Jang et al. (1999) used the mutual information not only to select the best candidate but also to assign weights to query terms. Gao et al. (2002) extended the basic co-occurrence model by adding a decaying factor that decreases the mutual information when the distance between the terms increases.

Automatic query expansion via blind relevance feedback has been known to be effective, especially when an initial query is a short. In a multilingual information retrieval, Ballesteros and Croft (1998) compared the parallel corpora disambiguation method and co-occurrence based disambiguation method. The post-translation expansion via local context analysis (Xu & Croft, 1996) after disambiguation based on co-occurrence showed to be helpful.

Document clusters, widely adopted in various applications such as browsing and viewing of document results (Hearst & Pedersen, 1996) or topic detection (Allan, Carbonell, Doddington, Yamron, & Yang, 1998), also reflect the association of terms and documents. Lee, Park, and Choi (2001) showed that incorporating a document re-ranking method based on document clusters into the vector space retrieval achieved significant improvement in monolingual IR, as it contributed to resolving ambiguities caused by polysemous query terms.

The noise or ambiguities produced by dictionary-based query translation in CLIR is much larger than the polysemous ambiguities in monolingual IR. For example, a Korean term '은행 [eun-haeng]' is a polysemous term with two meanings: 'bank' and 'ginkgo'. The English term 'bank' itself is polysemous, so the translated query ends up having magnified ambiguities. We will show that the method we propose, i.e. implicit ambiguity resolution using incremental clustering, is highly effective in dealing with the increased query ambiguities in CLIR.

The rest of the paper is organized as follows: Section 2 presents the basic system architecture, Section 3 describes the method of implicit ambiguity resolution by incremental clustering. Section 4 shows our experiments using the TREC CLIR test collection and the analyses of the results. We will conclude our paper in Section 5.

2. Basic system architecture

Fig. 1 shows the overall architecture of our system which incorporates the implicit ambiguity resolution method based on query-oriented document clusters. In the system, a query in Korean or Japanese is first translated into English by looking up Korean-to-English or Japanese-to-English dictionaries, and documents are retrieved based on the vector space method or probabilistic retrieval method for the translated English query. For the top-ranked retrieved documents, document clusters are incrementally created and the weight of each retrieved document is re-calculated by using clusters with preference. Below, we will briefly describe each module in the system. As the incremental clustering and implicit ambiguity resolution constitutes the core part of our system, we will explain this part in detail in the next section.

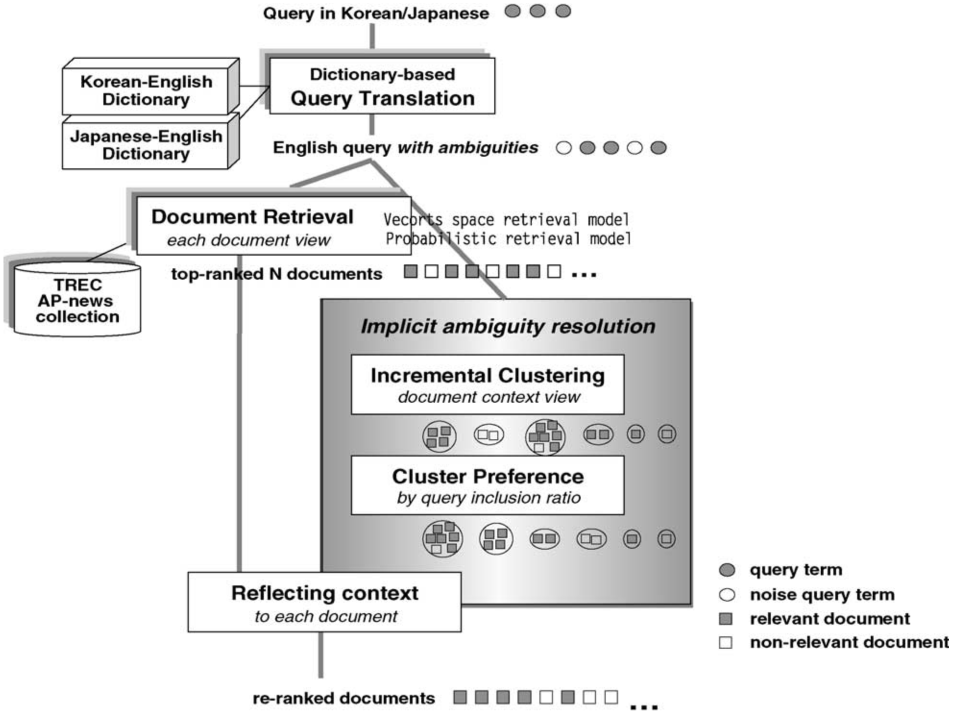
Query in Korean/Japanese ● ● ●
Korean-English Dictionary
Japanese-English Dictionary
Dictionary-based Query Translation
English query with ambiguities ○ ● ● ○ ●
Document Retrieval
each document view
Vecorts space retrieval model
Probablistic retrieval model
TREC AP-news collection
top-ranked N documents ■ □ ■ ■ □ ■ ■ □ ...
Implicit ambiguity resolution
Incremental Clustering
document context view
Cluster Preference
by query inclusion ratio
Reflecting context
to each document
re-ranked documents ■ ■ ■ ■ □ ■ □ □ ...
● query term
○ noise query term
■ relevant document
□ non-relevant document

Fig. 1. System architecture of implicit ambiguity resolution by incremental clustering.

2.1. Dictionary-based query translation and ambiguities

Queries are written in natural language in Korean/Japanese. We first apply morphological analysis and part-of-speech tagging to a query, and select keywords based on the POS information. For each keyword, we look up Korean–English/Japanese–English dictionaries, and all the English translations in the dictionaries are chosen as query terms.

For dictionary-based query translation, we used a general-purpose bilingual dictionary and technical bilingual dictionaries for the Korean–English translation. All in all, they have 282,511 Korean entries and 505,003 English translations. For the Japanese–English translation, we used the EDICT (Breen, 2003) which is a freely available Japanese/English Dictionary in machine-readable form. It includes 161,806 Japanese entries and 283,177 English translations.

Since a term can have multiple translations, the list of translated query terms can contain terms of different meanings as well as synonyms. While synonyms can improve retrieval effectiveness, terms with different meanings produced from the same original term can degrade retrieval performance tremendously.

At this stage, we can apply statistical ambiguity resolution method based on mutual information (Church & Hanks, 1990). In the experiment below, we will examine two cases, i.e. with and without ambiguity resolution at this stage.

2.2. Document retrieval

For the translated query, documents are retrieved based on the vector space retrieval model or the probabilistic retrieval model.

2.2.1. Based on vector space retrieval model

The vector space retrieval model simply checks the existence of query terms, and calculates the similarity between the query and documents.

A document and a query vector are respectively represented as follows: $D = \langle w_{d1}, w_{d2}, \ldots, w_{dn} \rangle$, $Q = \langle w_{q1}, w_{q2}, \ldots, w_{qt} \rangle$. The query-document similarity of each document is calculated by vector inner product of the query and document vector:

$$\mathrm{sim}(q,d) = \sum_{i=1}^{t} w_{qi} \cdot w_{di} \tag{1}$$

Terms in queries and documents are weighted by following weighting scheme which yields the best retrieval result in Lee et al. (2001) among several weighting schemes in SMART system. The weight of a term w_{qi} and w_{di} in a query and a document vector are calculated as follows:

$$w_{qi} = tf \cdot \log(N/df) \cdot 1 \Big/ \sqrt{\sum_{\text{vector}} w_j^2} \tag{2}$$

$$w_{di} = (\ln(tf) + 1) \cdot \log \frac{N}{df} \tag{3}$$

where N is the number of documents in the collection, df is the number of documents containing the term, tf is the frequency of occurrence of the term within a specific document.

2.2.2. Based on probabilistic retrieval model

In the probabilistic retrieval model, the probability that a specific document will be judged relevant to a specific query, is based on the assumption that the terms are distributed differently in relevant and non-relevant documents.

We used the Okapi BM25 formula which incorporates the Robertson–Sparck Jones weights (Robertson & Walker, 1999; Robertson & Spark Jones, 1976).

$$\mathrm{sim}(q,d) = \sum_{i \in Q} w^{(1)} \frac{(k_1 + 1)tf_{di}}{k_1((1-b) + b \cdot dl/avdl)tf_{di}} \frac{(k_3 + 1)tf_{qi}}{k_3 + tf_{qi}} \tag{4}$$

where Q is a query, containing terms $i, w^{(1)}$ is the Robertson/Sparck Jones weight of i in Q, which reduces to an inverse collection frequency weight without relevance information ($R = r = 0$).

$$w^{(1)} = \log \frac{(r + 0.5)/(R - r + 0.5)}{(n - r + 0.5)/(N - n - R + r + 0.5)} \tag{5}$$

N is the number of documents in the collection, n is the number of documents containing the term, R is the number of documents known to be relevant to a specific topic, r is the number of documents containing the term. k_1, b and k_3 are parameters which depend on the on the nature of the queries and possibly on the collection; k_1, b and k_3 default to 1.2, 0.75 and 7 respectively. tf_{di} is the frequency of occurrence of the term i within a specific document d. tf_{qi} is the frequency of the term i within the topic from which Q was derived. dl and $avdl$ are the document length and average document length measured in some suitable unit, respectively.

As the translated query can contain noises, non-relevant documents may be retrieved with high ranks; they may have higher ranks than relevant documents.

2.3. Dynamic incremental clustering to resolve ambiguities and to reflect contexts

In order to exclude the non-relevant documents from higher ranks, we take top N documents to create clusters incrementally and dynamically, and use the similarity between the clusters and the query to re-rank the documents. Basic idea is: each cluster created by clustering of retrieved documents can be seen as giving a context of the documents belonging to the cluster; by calculating the similarity between each cluster and the query, therefore, we can spot the relevant context of the query; documents that belong to more relevant context or cluster are likely to be relevant to the query.

It should be noted here that the static global clustering is not practical in the current setup, because it takes much computational time and the document space is too sparse (see Anick and Vaithyanathan (1997) for the comparison of static and dynamic clustering).

The method of incremental centroid clustering and of reflecting contexts to each document will be explained in Section 3.

3. The method of implicit ambiguity resolution by incremental clustering

3.1. Dynamic incremental centroid clustering

We make clusters based on incremental centroid method. There are a few variations in the agglomerative clustering method. The agglomerative centroid method joins the pair of clusters with the most similar centroid at each stage (Frakes & Baeza-Yates, 1992). Incremental centroid clustering method is straightforward as shown in Table 1.

As the degree of matching of evidences in documents is higher, the two documents are more similar. In document clustering, similar documents are classified as one cluster. Therefore, relevant documents are in the same cluster according to the *cluster hypothesis* (van Rijsbergen, 1979) which states that relevant documents tend to be more similar to each other than to non-relevant documents are.

3.2. Cluster preference

Similarities between the clusters and the query, or query-cluster similarities, are calculated by the combination of the query inclusion ratio and vector inner product between the query vector and the centroid vectors of the clusters.

$$\mathrm{sim}(q,c) = \frac{|c_q|}{|q|} \cdot \sum_{i=1}^{t} w_{qi} \cdot w_{ci} \tag{6}$$

where $|q|$ is the number of terms in the query, $|c_q|$ is the number of query terms included in a cluster centroid, $|c_q|/|q|$ is the query inclusion ratio for the cluster. w_{qi} is the weight of term i in the query vector, w_{ci} is the weight of term i in the cluster centroid vector. The inner product between the query and the centroid vector are calculated. The documents included in the same cluster have the same query-cluster similarity.

Cluster preferences are influenced by the query inclusion ratio, which prefers the cluster whose centroid includes more various query terms. Thus incorporating this information into the

Table 1
Incremental centroid clustering according to the ranks of top N documents

Input	The input document of incremental clustering proceeds according to the rank of the top N documents obtained from the vector space retrieval for a query: Documents and cluster centroids are represented in vectors
Step 1	For the first input document (rank 1), crate one cluster whose member is itself
Step 2	For each consecutive document (ranks $2,\ldots,N$), compute the cosine similarity between the document and each cluster centroid in the already created clusters If the similarity between the document and a cluster is above the threshold, then add the document to the cluster as a member and update the cluster centroid. Otherwise, create a new cluster with the document

Note that one document can be a member of several clusters as shown in Fig. 2 (solid lines show that the document belongs to the cluster).

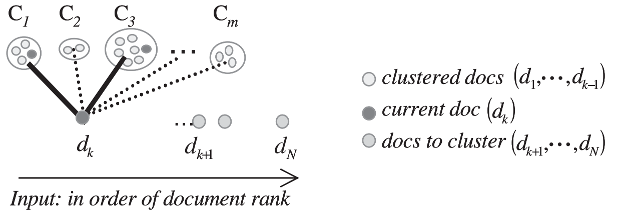
C1
C2
C3
Cm
...
dk
... dk+1
dN
○ clustered docs (d1,···,dk-1)
● current doc (dk)
○ docs to cluster (dk+1,···,dN)
Input: in order of document rank

Fig. 2. Incremental centroid clustering for the top-ranked N documents.

weighting of each document means adding information which is related to the behavior of terms in documents as well as the association of terms and documents into the evaluation of the relevance of each document; it therefore has the effect of ambiguity resolution.

3.3. Reflecting the cluster information to the documents

Using the query-cluster similarity, we re-calculate the relevance of each document according to the following equation:

$$\mathrm{sim}(q,d)' = \mathrm{sim}(q,d)\cdot \mathrm{Max}_{d\in c}\mathrm{sim}(q,c) \tag{7}$$

where $\mathrm{sim}(q,d)$ is a query-document similarity by the vector space retrieval as defined in Eq. (1) or the probabilistic retrieval as defined in Eq. (4). $\mathrm{sim}(q,c)$ is a query-cluster similarity of a document d defined in Eq. (6). Since each document can be a member of several clusters, we assign the highest query-cluster similarity value to the document. The new document similarity, $\mathrm{sim}(q,d)'$, is calculated by the multiplication of the cluster similarity and the document similarity. Based on this new weight $\mathrm{sim}(q,d)'$, we re-rank the retrieved documents.

Through this procedure, we can effectively take into account the contexts of all the terms in a document as well as of the query terms. Thus, even if a document which has a low query-document similarity can have a high query-cluster similarity thanks to the effect of neighboring documents in the same cluster. The reverse can be true as well.

4. Experiments and evaluations

4.1. Experimental setup

We evaluated the effectiveness of proposed method on TREC-6 and TREC-8 CLIR test collection which contains 242,918 English documents (AP news from 1988 to 1990) and 52 English queries. English queries are translated to Korean/Japanese queries manually. We use the title field of a query which consists of three fields: <title>, <description> and <narrative>. Table 2 gives the statistics of the test collection. In dictionary-based query translation, one query term have multiple translations. Table 3 shows the degree of ambiguities. Table 4 gives the examples of translated Korean and Japanese queries by human for the original English query.

In our experiment, we only use 42 queries which consist of more than one term to observe the real effects of proposed method. This is because, if a query consists of more than one term, human

Table 2
The statistics of TREC-6 and TREC-8 CLIR test collection

The number of documents	242,918
The number of queries	52
The number of average relevance document	45.12
The average length of document	479.08
The average length of query (title)	2.92

Table 3
The degree of ambiguities for 52 queries

	Korean query	Japanese query
The number of query terms	124	124
The number of translated English terms	585	268
The average number of translations	4.72	2.16

Table 4
The examples of translated Korean and Japanese query by human for English query

English	Korean	Japanese
Swiss speed limits	스위스 속도 제한	スイス 制限 速度
Effects of logging	벌채 효과	伐採 影響
Solar powered cars	태양열 자동차	ソーラー 発電 車
Organic cotton	유기 면화	有機 綿
Middle-east peace process	중동 평화 절차	中東 和平 プロセス
International terrorism	국제 테러리즘	国際 テロリズム

can select the correct meaning of the term by their neighbors. But if a query consists of one term such as ‘bank’ and the term is polysemous, no one can resolve ambiguities without considering additional external information. The rest 10 queries which consist of one term are used to decide the threshold in incremental clustering.

For document retrieval, we use SMART system (Salton, 1989) developed at Cornell as a vector space retrieval model. As the probabilistic retrieval model, we use Lemur’s Okapi (Paul & Callan, 2001).

For comparisons with resolving ambiguity in CLIR, we evaluate disambiguation method based on co-occurrence, and post-translation blind relevance feedback method in case with/without ambiguity resolution.

In disambiguation method, the best translation is selected among all translation terms based on co-occurrence information (Chun, 2000) on the hypothesis that the correct translations of query terms will co-occur. Co-occurrence $\mathrm{cooc}(x,y)$ is defined as following:

$$\mathrm{cooc}(x,y) = \sqrt{\frac{N \cdot f(x,y)}{f(x) + f(y)}} \tag{8}$$

where $f(x)$ and $f(y)$ are frequency of term x and term y, respectively. Co-occurrence frequency of term x and term y, $f(x,y)$, is taken in window size 6 for AP 1988 news documents. The value of N is 10,000,000.

To select the correct translation terms, calculate cooc(x, y) value for the all possible set $\{x, y\}$ such that x is the translation of a source term a and y is the translation of a source term b. Each set is ranked by cooc score and the highest ranking set is taken as the appropriate translation.

For the blind feedback on the vector space retrieval, the top k terms with the highest weights are added to the original query by summing the weights of the term in the top r documents which are assumed to be relevant. For the blind feedback for the probabilistic retrieval, we used the Okapi's feedback formula since it gave better performance than that of blind feedback by sum.

We compared proposed method with the monolingual retrieval, with translated queries without disambiguation, and with the translated queries after disambiguation, with the translated queries expanded by blind relevance feedback. We experimented on Korean-to-English CLIR and Japanese-to-English CLIR. The followings are the brief descriptions for comparison methods:

(1) *Monolingual*: the performance for original English queries as the monolingual baseline.
(2) *t_all_base*: the performance for translated English queries which have all possible translations in bilingual dictionaries without ambiguity resolution.
(3) *t_all_blindf*: the performance of blind feedback for the retrieved documents of t_all_base.
(4) *t_all_rerank*: the performance of proposed method using dynamic incremental clusters for the retrieved documents of t_all_base.
(5) *t_one_base*: the performance for translated queries with the best translations after ambiguity resolution based on co-occurrence information.
(6) *t_one_blindf*: the performance of blind feedback for the retrieved documents of t_one_base.
(7) *t_one_rerank*: the performance of proposed method using dynamic incremental clusters for the retrieved documents of t_one_base.

't_all_rerank' and 't_one_rerank' use our implicit disambiguation method. The number of top N documents used in dynamic incremental clustering is 300 and the threshold for incremental centroid clustering is set as 0.34 in both t_all_rerank and t_one_rerank. The parameters for blind relevance feedback of the results by Okapi are determined by the best performances of blind feedback for original English queries. Ten terms in ten documents are selected. For performance of blind feedback on vector space retrieval, the results are the best performance for several parameters since the performance was very sensitive to the parameters.

4.2. Results

The main objective of this paper is to observe the performance change by incremental clusters for translated queries with ambiguities (t_all_base and t_all_rerank). The results are summarized in Tables 5 and 6 for Korean-to-English CLIR and Japanese-to-English CLIR.

In Korean-to-English CLIR, the proposed method (t_one_rerank) achieved 39.41% and 36.79% improvements for all translation queries (t_all_base) compared with the vector space retrieval and the probabilistic retrieval, respectively. The proposed method (t_all_rerank) achieved 6.81% and 12.30% performance improvements for all translation queries, compared with blind feedback (t_all_blindf) in Korean-to-English CLIR, on the vector space retrieval and on the probabilistic retrieval, respectively.

Table 5
The performance comparisons in Korean-to-English CLIR

	Vector space retrieval (SMART)		Probabilistic retrieval (OKAPI BM25)	
	11-pt avg precision	chg%	Set avg precision	chg%
(1) Monolingual	0.267	–	0.274	–
(2) t_all_base	0.170	–	0.158	–
(3) t_all_blindf	0.191	12.35%	0.179	13.29%
(4) t_all_rerank	0.204	20.00%	0.201	27.22%
(5) t_one_base	0.210		0.207	
(6) t_one_blindf	0.217	3.33%	0.216	4.35%
(7) t_one_rerank	0.237	12.86%	0.216	4.35%

Table 6
The performance comparisons in Japanese-to-English CLIR

	Vector space retrieval (SMART)		Probabilistic retrieval (OKAPI BM25)	
	11-pt avg precision	chg%	Set avg precision	chg%
(1) Monolingual	0.267	–	0.274	–
(2) t_all_base	0.190	–	0.151	–
(3) t_all_blindf	0.206	9.47%	0.173	14.57%
(4) t_all_rerank	0.216	13.68%	0.187	23.84%
(5) t_one_base	0.202	–	0.175	–
(6) t_one_blindf	0.211	4.46%	0.203	16.00%
(7) t_one_rerank	0.224	10.89%	0.197	12.57%

In Japanese-to-English CLIR, the proposed method achieved 17.89% and 30.46% improvements for all translation queries compared with the vector space retrieval and the probabilistic retrieval, respectively. For the probabilistic retrieval, the proposed method showed better performances for all translation queries, but similar performances for disambiguated queries.

Fig. 3 shows the performance changes depending on the parameters in blind feedback and the thresholds of incremental clustering in proposed method, which are for the results of the probabilistic retrieval on Korean-to-English CLIR. It shows that the performances of blind feedback are very sensitive to the parameters. The performance improvement by proposed method is more stable than that by blind feedback.

The cluster-based implicit disambiguation method, therefore, is more effective for performance improvement than the query disambiguation method based on co-occurrence information; if used together, it shows yet further improvement.

4.3. The analysis of the results

We examined the effects of our method in case of an ambiguous query and an unambiguous query after bilingual dictionary-based term translation.

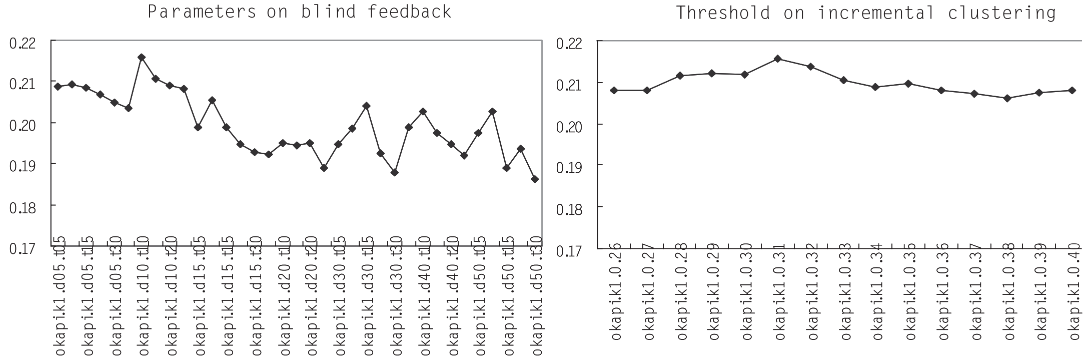

Fig. 3. The performance changes depending on the parameters on blind feedback and the thresholds on clustering for the probabilistic retrieval in Korean-to-English CLIR.

4.3.1. *In case of an ambiguous query*

The Korean query is '자동차 [ja-dong-cha] 공기 [gong-gi] 오염 [o-yeom]' whose original English query is 'automobile air pollution'. The translated query with all the possible translations in Korean-English dictionaries for this query is as follows:

자동차 [ja-dong-cha]	Car, automobile, autocar, motorcar
공기 [gong-gi]	Air, *atmosphere, empty vessel, bowl, jackstone, pebble, marbles*
오염 [o-yeom]	Contamination, pollution

In this query, the term '공기' is polysemous which has several meanings such as <air>, <atmosphere>, <jackstone>, <co-occurrence>, and <bowl>. This is the cause of degrading system performance.

146 clusters were created for the retrieved 300 documents of this query. The token number of documents in he clusters was 435. The distribution of cluster members is shown in Fig. 4. Most non-relevant documents had a tendency to make singleton cluster, and most relevant documents made large group clusters.

We examined inside the clusters how to see cluster give effects to resolve ambiguity and reflect context. Cluster C4 in Fig. 4 has 60 members, which contains 56 relevant documents and 4 non-relevant documents, among 209 relevant documents for this query. This cluster centroid includes following terms related to the query:

Car	0.069
Automobile	0.127
Air	0.082
Atmosphere	0.018
Pollution	0.196
Contamination	0.064

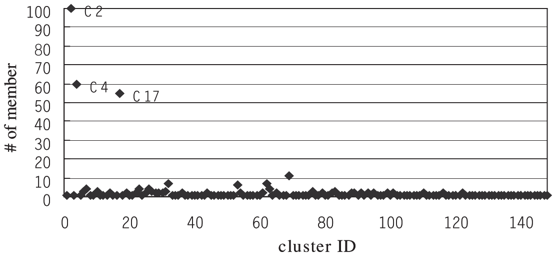

Fig. 4. The distribution of cluster members for a query with translation ambiguities.

Although this centroid includes a noise term 'atmosphere', its weight is low. The other terms are appropriate to the query; they are synonyms. Since all of the query terms are included in the centroid, query inclusion ratio is 1 and all synonyms affect positively to the vector inner product value. Therefore, since this cluster preference is high, the ranks of all documents in this cluster changed higher. The cluster performed as a context of the documents relevant to the query. Cluster C85 is a singleton whose centroid includes one of three query terms:

Bowl	0.101
Marble	0.191

Since query inclusion ratio is low, the cluster preference is low. Therefore this cluster's effect is weak to the document.

Fig. 5 presents the rank changes, calculated by subtracting ranks by our method (t_all_rerank) from those by vector space retrieval (t_all_base) for each relevant document of the query. The ranks of most documents are changed higher through cluster analysis, although the ranks of some documents are changed lower. Fig. 6 shows the recall/precision curves for the performances of original English query (monolingual; 0.6783 in 11-pt. avg. precision), translated query without

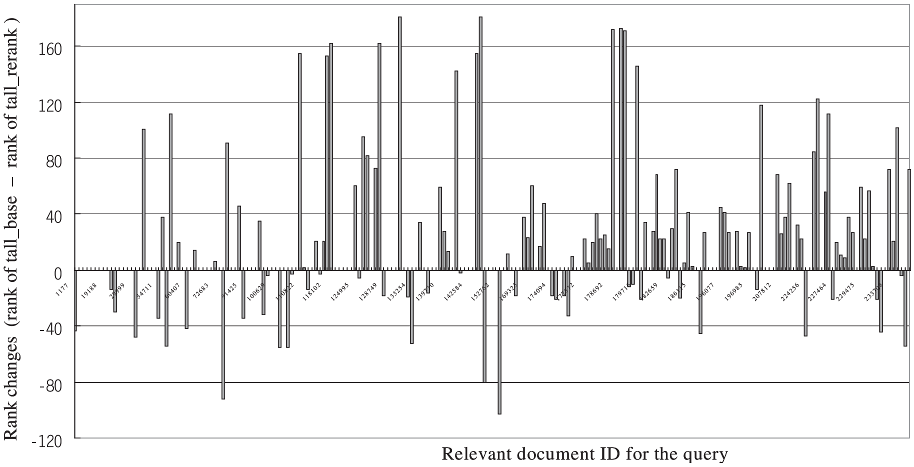

Fig. 5. The rank changes of tall_rerank from rank of tall_base for each relevant document of the query.

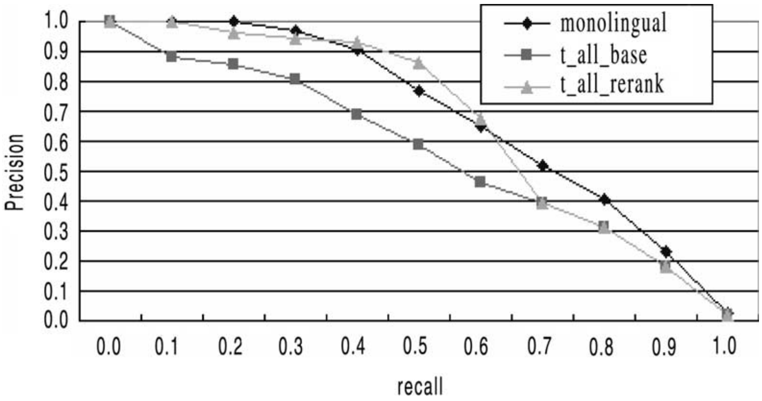

Fig. 6. The performance comparisons for the ambiguous query.

disambiguation (t_all_base; 0.5635), and our method (t_all_rerank; 0.6622). For increased query ambiguity, we could achieve 97.62% performance compared to the monolingual retrieval.

4.3.2. *In case of an unambiguous query*

If a query is not ambiguous, the translated query can include synonyms. In this case, the effect of our method shows the query expansion and reflecting context.

The Korean query is '쓰레기 [sseu-le-gi] 재활용 [jae-hwal-yong]' whose original English query is 'reusage of garbage'. The translated query with all the possible translations in Korean–English dictionaries for this query is as follows:

쓰레기 [sseu-le-gi]	Waste, garbage, refuse, rubbish, trash
재활용 [jae-hwal-yong]	Recycle, reusage

In this query, the two terms have synonyms, which show query expansions.

As shown Fig. 7, the performance of t_all_base and t_all_rerank is higher than that of original English query. The recall/precision curves for the performances of original English query

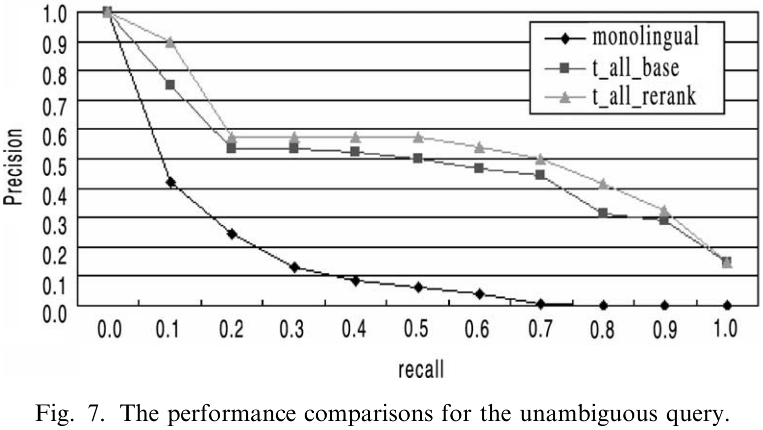

Fig. 7. The performance comparisons for the unambiguous query.

(monolingual; 0.1815 in 11-pt. avg. precision), the translated query (t_all_base; 0.4993), and our method (t_all_rerank; 0.5564). We could achieve 206.56% performance improvement compared to the monolingual retrieval.

These results indicate that cluster analysis help to resolve ambiguity in case of ambiguous query and to reflect context to documents in case of unambiguous query. Thus, we could effectively take into account the context of all the terms in a document as well as the query terms.

5. Conclusion

In this paper, we have proposed the method of applying dynamic incremental clustering to the implicit resolution of query ambiguities in Korean-to-English and Japanese-to-English cross-language information retrieval. The method used the clusters of retrieved documents as a context for re-weighting each retrieved document and for re-ranking the retrieved documents.

Our method was evaluated on TREC-6 and TREC-8 CLIR test collection. This method achieved 39.41% and 36.79% performance improvements for translated queries without ambiguity resolution in Korean-to-English CLIR, and 17.89% and 30.46% improvements in Japanese-to-English CLIR, on the vector space retrieval and on the probabilistic retrieval, respectively. The proposed method achieved 6.81% and 12.30% performance improvements for all translation queries, compared with blind feedback in Korean-to-English CLIR, on the vector space retrieval and on the probabilistic retrieval, respectively. The performance changes by the thresholds on clustering in proposed method are more stable than those by the parameters in blind relevance feedback. These results indicate that cluster analysis help to resolve ambiguity greatly, and each cluster itself provide a context for a query. We have shown the effectiveness of our method on Korean-to-English and Japanese-to-English CLIR. The method is a language independent model which can be applied to any language retrieval.

We expect that our method will further improve the results, although further research is needed on combining a method to improve recall such as query expansion and relevance feedback.

Acknowledgement

This research was supported by the Ministry of Science and Technology in Korea.

References

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study: final report. In *Proceedings of the DARPA broadcast news transcription and understanding workshop* (pp. 194–218).

Anick, P. G., & Vaithyanathan, S. (1997). Exploiting clustering and phrases for context-based information retrieval. In *Proceedings of the 20th ACM SIGIR conference (SIGIR'97)*.

Ballesteros, L., & Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st ACM SIGIR conference (SIGIR'98)*.

Breen, J. (2003). EDICT Japanese/English dictionary file. The Electronic Dictionary Research and Development Group, Monash University. Available: http://www.csse.monash.edu.au/~jwb/edict_doc.html.

Chun, J. H. (2000). Resolving ambiguity and English query supplement using parallel corpora on Korean–English CLIR system. MS thesis, Department of Computer Science, KAIST (in Korean).

Church, K. W., & Hanks, P. (1990). Word association norms mutual information and lexicography. *Computational Linguistics, 16*(1), 23–29.

Davis, M. (1996). New experiments in cross-language text retrieval at NMSU's computing research lab. In *Proceedings of the fifth text retrieval conference (TREC-5)*.

Dumais, S. T., Letsche, T. A., Littman, M. L., & Landauer, T. K. (1997). Automatic cross-language retrieval using latent semantic indexing. In *Proceedings of the AAAI symposium on cross-language text and speech retrieval*.

Eichmann, D., Ruiz, M. E., & Srinivasan, P. (1998). Cross-language information retrieval with the UMLS metathesaurus. In *Proceedings of the 21st ACM SIGIR conference (SIGIR'98)*.

Frakes, W. B., & Baeza-Yates, R. (1992). *Information retrieval: data structures and algorithms*. New Jersey: Prentice Hall (pp. 435–436).

Gao, J., Zhou, M., Nie, J.-Y., He, H., & Chen, W. (2002). Cross-language information retrieval: resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relation. In *Proceedings of the 25th ACM SIGIR conference (SIGIR'01)*.

Gilarranz, J., Gonzalo, J., & Verdejo, F. (1997). An approach to conceptual text retrieval using the EuroWordNet multilingual semantic database. In *Proceedings of the AAAI spring symposium on cross-language text and speech retrieval*.

Hearst, M. A., & Pedersen, J. O. (1996). Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *Proceedings of the 19th ACM SIGIR conference (SIGIR'96)*.

Hull, D. A., & Grefenstette, G. (1996). Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th ACM SIGIR conference (SIGIR'96)*.

Jang, M. G., Myaeng, S. H., & Park, S. H. (1999). Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of the 37th annual meeting of the association for computational linguistics*.

Kupiec, J. M. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting of the association for computational linguistics*.

Kwon, O.-W., Kang, I. S., Lee, J.-H., & Lee, G. B. (1997). Cross-language text retrieval based on document translation using Japanese-to-Korean MT system. In *Proceedings of the NLPRS'97* (pp. 101–106).

Lee, K. S., Park, Y. C., & Choi, K. S. (2001). Re-ranking model based on document clusters. *Information Processing and Management, 37*(1), 1–14.

Oard, D. W., & Hackett, P. (1997). Document translation for the cross-language text retrieval at the University of Maryland. In *Proceedings of the sixth text retrieval conference (TREC-6)*.

Paul, O., & Callan, J. (2001). Experiments using the Lemur toolkit. In *Proceedings of the tenth text retrieval conference (TREC-10)*.

Rehder, B., Littman, M. L., Dumais, S., & Landauer, T. K. (1997). Automatic 3-language cross-language information retrieval with latent semantic indexing. In *Proceedings of the sixth text retrieval conference (TREC-6)*.

Robertson, S. E., & Spark Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*, 129–146.

Robertson, S. E., & Walker, S. (1999). Okapi/Keenbow at TREC-8. In *Proceedings of the eighth text retrieval conference (TREC-8)*.

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Reading, Pennsylvania: Addison-Wesley.

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics, 22*(1), 1–38.

van Rijsbergen, C. J. (1979). *Information retrieval* (second edition). London: Butterworths.

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th ACM SIGIR conference (SIGIR'96)*.

Yang, Y., Carbonell, J. G., Brown, R. D., & Frederking, R. E. (1998). Translingual information retrieval: learning from bilingual corpora. *AI Journal Special Issue*, 323–345.