Implicit Ambiguity Resolution Using Incremental Clustering in Korean-to-English Cross-Language Information Retrieval

Kyung-Soon Lee[1], Kyo Kageura[1], Key-Sun Choi[2]

[1] NII (National Institute of Informatics)
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo, 101-8430, Japan
{kslee, kyo}@nii.ac.jp

[2] Division of Computer Science, KAIST
373-1 Kusung Yusong
Daejeon, 305-701, Korea
kschoi@cs.kaist.ac.kr

Abstract

This paper presents a method to implicitly resolve ambiguities using dynamic incremental clustering in Korean-to-English cross-language information retrieval. In the framework we propose, a query in Korean is first translated into English by looking up Korean-English dictionary, then documents are retrieved based on the vector space retrieval for the translated query terms. 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 clusters. In experiment on TREC-6 CLIR test collection, our method achieved 28.29% performance improvement for translated queries without ambiguity resolution for queries. This corresponds to 97.27% of the monolingual performance for original queries. When we combine our method with query ambiguity resolution, our method even outperforms the monolingual retrieval.

1 Introduction

This paper describes a method of applying dynamic incremental clustering to the implicit resolution of query ambiguities in Korean-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 approaches and translation approaches. Statistical methods establish cross-lingual associations without language translation (Dumais et al, 1997; Rehder et al, 1997; Yang et al, 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 et al, 1997; Oard and Hackett, 1997), it is not very practical. Query translation methods (Hull and Grefenstette, 1996; Davis, 1996; Eichmann et al, 1998; Yang et al, 1998; Jang et al, 1999; Chun, 2000) based on bilingual dictionaries, multilingual ontology or thesaurus are much more practical. Many researches adopt dictionary-based query translation 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. One way of resolving query ambiguities is to use the statistics, such as mutual information (Church and Hanks, 1990), to measure associations of query terms, on the basis of existing corpora (Jang et al, 1999).

Document clusters, widely adopted in various applications such as browsing and viewing of document results (Hearst and Pedersen, 1996) or topic detection (Allan et al, 1998), also reflect the association of terms and documents. Lee et al (2001) showed that incorporating a document re-ranking method based on document clusters into the vector space retrieval achieved the significant improvement in monolingual IR, as it

contributed to resolving ambiguities caused by polysemous query terms.

The noise or ambiguity 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.

2 Implicit ambiguity resolution using incremental clustering

Figure 1 shows the overall architecture of our system which incorporates implicit ambiguity resolution method based on query-oriented document clusters. In the system, a query in Korean is first translated into English by looking up dictionaries, and documents are retrieved based on the vector space retrieval for the translated 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. This phase is the core of our implicit ambiguity resolution method. Below, we will describe each module in the system.

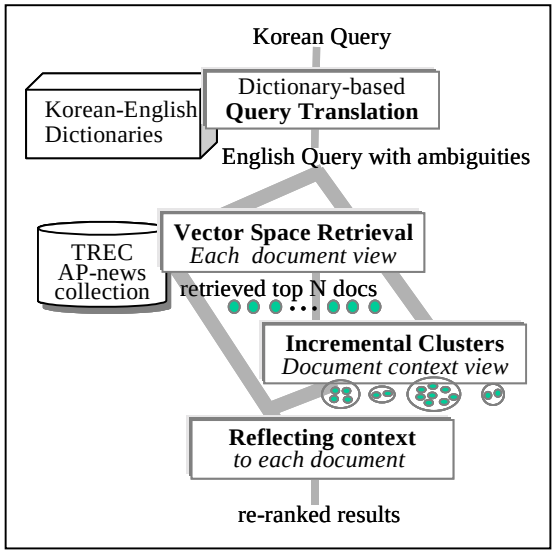

Figure 1. System architecture of implicit ambiguity resolution by incremental clustering.

2.1 Dictionary-based query translation and ambiguities

Queries are written in natural language in Korean. We first apply morphological analysis and part-of-speech (POS) tagging to a query, and select keywords based on the POS information. For each keyword, we look up Korean-English dictionaries, and all the English translations in the dictionaries are chosen as query terms. We used a general-purpose bilingual dictionary and technical bilingual dictionaries (Chun, 2000). All in all, they have 282,511 Korean entries and 505,003 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. In the experiment below, we will examine two cases, i.e. with and without ambiguity resolution at this stage.

2.2 Document retrieval based on vector space retrieval model

For the query, documents are retrieved based on the vector space retrieval method. This method simply checks the existence of query terms, and calculates similarities between the query and documents. The query-document similarity of each document is calculated by vector inner product of the query and document vectors:

$$simD(q,d) = \sum_{i=1}^{t} w_{qi} \cdot w_{di} \qquad \textbf{(1)}$$

where query and document weight, w_{qi} and w_{di}, are calculated by *ntc-ltn* weighting scheme which yields the best retrieval result in Lee et al (2001) among several weighting schemes used in SMART system (Salton, 1989).

As the translated query can contain noises, non-relevant documents may have higher ranks than relevant documents.

2.3 Query-oriented incremental clustering for implicit ambiguity resolution

In order to exclude non-relevant documents from higher ranks, we take top N documents to create clusters incrementally and dynamically, and use similarities 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).

2.3.1 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 and Baeza-Yates, 1992).

Incremental centroid clustering method is straightforward. The input document of incremental clustering proceeds according to the ranks of the top-ranked N documents resulted from vector space retrieval for a query. Document and cluster centroid are represented in vectors. For the first input document (rank 1), create one cluster whose member is itself. For each consecutive document (rank 2, ..., N), compute 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 a threshold, then add the document to the cluster as a member and update cluster centroid. Otherwise, create a new cluster with this document. Note that one document can be a member of several clusters as shown in Figure 2 (sold lines show that the document belongs to the cluster).

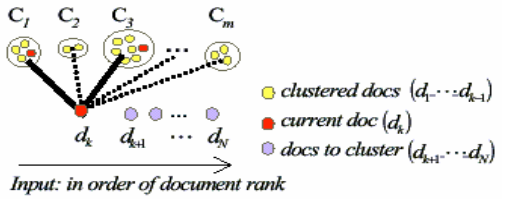

Figure 2. Incremental centroid clustering in order of the top-ranked N documents

2.3.2 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.

$$simC(q,c) = \frac{|c_q|}{|q|} \cdot \sum_{i=1}^{t} w_{qi} \cdot w_{ci} \qquad \textbf{(2)}$$

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. 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 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.

2.4 Reflecting cluster information to the documents

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

$$sim(q,d) = simD(q,d) \cdot MAX_{d \in c} simC(q,c) \quad \textbf{(3)}$$

where *simD(q,d)* is a query-document similarity by vector space retrieval as defined in equation (1) and *simC(q,c)* is a query-cluster similarity of a document *d* defined in equation (2). 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, *sim(q,d)*, is calculated by multiplication of a query-cluster similarity and a query-document similarity. Based on this new

similarity *sim(q,d)*, we re-rank the retrieved documents. In the equation, we tried to use weighted sum of a query-document similarity and a query-cluster similarity. The combination by multiplication showed better performances than that of weighted sum.

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.

3 Experiments

3.1 Experimental environment

We evaluated our method on TREC-6 CLIR test collection which contains 242,918 English documents (AP news from 1988 to 1990) and 24 English queries. English queries are translated to Korean queries manually. We use title field of queries which consist of three fields such as title, description and narrative.

In dictionary-based query translation, one query term has multiple translations. Table 3 shows the degree of ambiguities.

The number of Korean query terms	47
The number of translated terms	149
The average number of translations	3.2

Table 1. The degree of ambiguities for 24 queries.

In our experiment, we only use 14 queries which consist of more than one term to observe real effects of our method. This is because, if a query consists of more than one term, human can select the correct meaning of the term by its neighbours. But if a query consists of one term such as 'bank' and it 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 a threshold in incremental clustering.

We use SMART system (Salton, 1989) developed at Cornell as a vector space retrieval.

3.2 Results

The retrieval effectiveness was evaluated using the 11-point average precision metric.

We compared our method with original English queries, with translated queries with ambiguities, and with translated queries with the best translation after disambiguation. The followings are the brief descriptions for comparison methods:

1) ***monolingual***: the performance of vector space retrieval system for original English queries as the monolingual baseline.
2) ***tall_base***: the performance of vector space retrieval system for translated English queries which have all possible translations in bilingual dictionaries without ambiguity resolution.
3) ***tall_rerank***: the performance of proposed method using dynamic incremental clusters for the retrieved documents of tall_base.
4) ***tone_base***: the performance of vector space retrieval system for translated queries with the best translations for each query term after ambiguity resolution based on mutual information.
5) ***tone_rerank***: the performance of proposed method using dynamic incremental clusters for the retrieved documents of tone_base.

'tall_rerank' and 'tone_rerank' use our implicit disambiguation method. The number of top N documents used in dynamic incremental clustering is 300 and thresholds for incremental centroid clustering are set as 0.41 which are learned from training 10 queries with one term in both tall_rerank and tone_rerank.

The main objective of this paper is to observe the performance change by incremental clusters for translated queries with ambiguities (tall_base and tall_rerank).

Comparison	11-pt avg. precision	C/M (%)	Change (%)
1) monolingual	0.2858	100	-
2) tall_base	0.2167	75.82	-
3) tall_rerank	0.2780	97.27	+28.29
4) tone_base	0.2559	89.54	-
5) tone_rerank	0.3026	105.87	+18.25

Table 2. The retrieval effectiveness for comparison methods.

To observe the effect of clusters, we compared the results after disambiguation based on mutual information (tone_base and tone_rerank). We selected the best translation based on mutual information among all translation terms. Mutual information *MI(x,y)* is defined as following (Church and Hanks, 1990):

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} = \log_2 \frac{N \cdot f(x,y)}{f(x)f(y)} \quad \textbf{(4)}$$

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 11-point average precision value, corresponding result to monolingual (C/M), and performance change are summarized in Table 2. The retrieval effectiveness of tall_rerank is 0.2780, corresponding to 97.27% of monolingual performance. The performance of tone_rerank yields 0.3026 (105.87%). This is even better than the monolingual performance. The performance of our implicit ambiguity resolution method for all translations (tall_rerank) shows 8.63% improvement compared with that of ambiguity resolution based on mutual information (tone_base). The proposed method achieved 28% improvement for all translation queries and 18% for best translation queries compared with the vector space retrieval. Our method after disambiguation (tone_rerank) using mutual information improved about 39.6% over vector space retrieval for all translations queries (tall_base).

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

3.3 Result analysis

We examined the effects of our method for a query with ambiguities increased after bilingual dictionary-based term translation.

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.

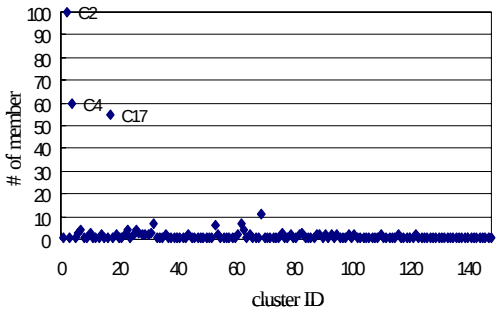

Figure 3. The distribution of cluster members for the query with translation ambiguities.

146 clusters were created for the retrieved 300 documents of this query. The token number of documents in the clusters was 435. The distribution of cluster members is shown in Figure 3. 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 Figure 3 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

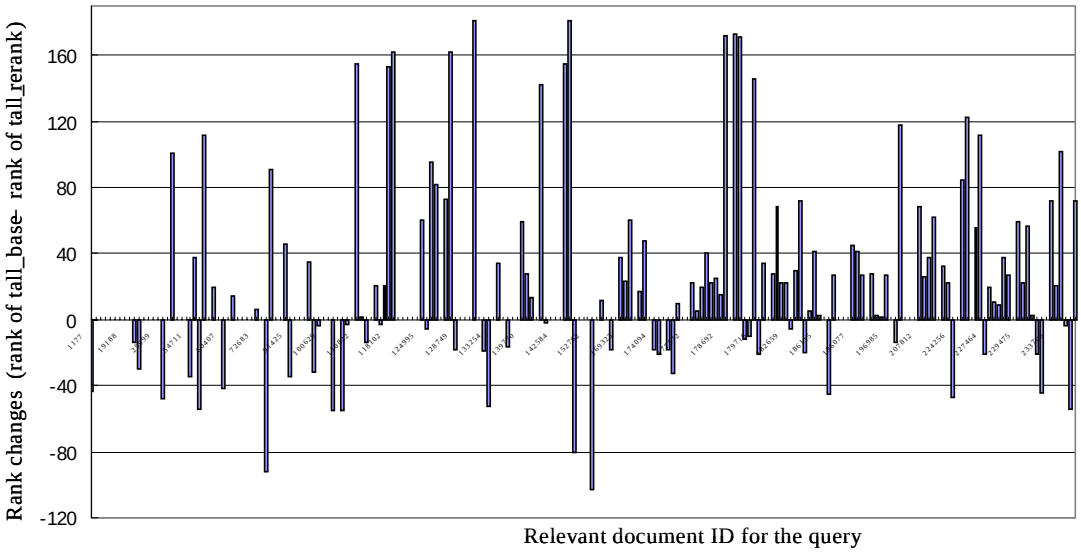

Figure 4. The rank changes of tall_rerank from rank of tall_base for each relevant document of the query.

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.

Figure 4 presents the rank changes, calculated by subtracting ranks by our method (tall_rerank) from those by vector space retrieval (tall_base) for each relevant document of the ambiguous query. The ranks of most documents are changed higher through cluster analysis, although the ranks of some documents are changed lower. Figure 5 shows recall/precision curves for the performances of original English query (monolingual; 0.6783 in 11-pt avg. precision), translated query without disambiguation (tall_base; 0.5635), and our

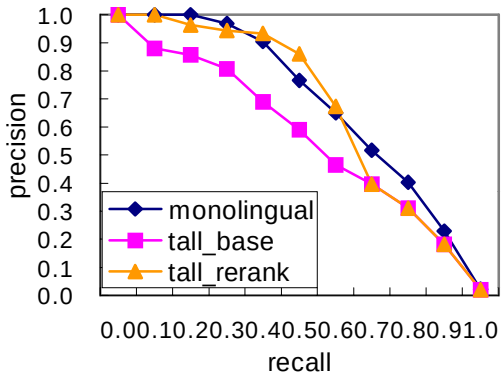

Figure 5. The performance comparison for the ambiguous query.

method (tall_rerank; 0.6622). For increased query ambiguity, we could achieve 97.62% performance compared to the monolingual retrieval.

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

4 Conclusion

We have proposed the method of applying dynamic incremental clustering to the implicit resolution of query ambiguities in Korean-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 CLIR test collection. This method achieved 28.29% performance improvement for translated queries without ambiguity resolution. This corresponds to 97.27% of the monolingual performance. When our method was used with the query ambiguity resolution method based on mutual information, it showed 105.87% performance improvement of the monolingual retrieval. These results indicate that cluster analysis help to resolve ambiguity greatly, and each cluster itself provide a context for a query.

Our 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.

References

Allan, J. Carbonell, J., Doddington, G. Yamron. J. and Yang, Y. (1998) Topic Detection and Tracking Pilot Study: Final Report. In Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, pp.194-218.

Anick, P.G. and Vaithyanathan, S. (1997) Exploiting Clustering and Phrases for Context-Based Information Retrieval. In Proc. of 20th ACM SIGIR Conference (SIGIR'97).

Chun, J.H. (2000) Resolving Ambiguity and English Query Supplement using Parallel Corpora on Korean-English CLIR system. MS thesis, Dept. of Computer Science, KAIST (in Korean).

Church, K.W. and Hanks P. (1990) Word Association Norms Mutual Information and Lexicography. Computational Linguistics, 16(1), pp.23-29.`

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

Dumais, S.T., Letsche, T.A., Littman, M.L. and Landauer, T.K. (1997) Automatic cross-language retrieval using latent semantic indexing. In Proc. of AAAI Symposium on Cross-Language Text and Speech Retrieval.

Eichmann, D., Ruiz, M.E. and Srinivasan, P. (1998) Cross-Language Information Retrieval with the UMLS Metathesaurus. . In Proc. of the 21th ACM SIGIR Conference (SIGIR'98).

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

Gilarranz, J., Gonzalo, J. and Verdejo, F. (1997) An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database. In Proc. of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval.

Hearst, M.A. and Pedersen, J.O. (1996) Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In Proc. of 19th ACM SIGIR Conference (SIGIR'96).

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

Jang, M.G., Myaeng, S.H. and Park, S.H. (1999) Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting. In Proc. of the 37th Annual Meeting of the Association for Computational Linguistics.

Kwon, O-W., Kang, I.S., Lee, J-H and Lee, G.B. (1997) Cross-Language Text Retrieval Based on Document Translation Using Japanese-to-Korean MT system. In Proc. of 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), pp. 1-14.

Oard, D.,W. and Hackett, P. (1997) Document Translation for the Cross-Language Text Retrieval at the University of Maryland. In Proc. of the Sixth Text REtrieval Conference (TREC-6).

Rehder, B., Littman, M.L., Dumais, S. and Landauer, T.K. (1997) Automatic 3-language cross-language information retrieval with latent semantic indexing. In Proc. of the Sixth Text REtrieval Conference (TREC-6).

Salton, G. (1989) Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, Pennsylvania.

Voorhees, E.M. (1986) Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. Information Processing & Management, 22(6), pp. 465-476.

Yang, Y., Carbonell, J.G., Brown, R.D. and Frederking, R.E. (1998) Translingual Information Retrieval: Learning from Bilingual Corpora. AI Journal special issue, pp. 323-345.