Merging Multilingual Information Retrieval Results Based on Prediction of Retrieval Effectiveness

Wen-Cheng Lin and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN
denislin@nlg.csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

Abstract

This paper deals with Chinese, English and Japanese multilingual information retrieval (MLIR). Merging problem in distributed MLIR is studied. The prediction of retrieval effectiveness is used to determine the merging weight of each intermediate run. The translation penalty and collection weight are considered to improve merging performance. Several merging strategies are experimented. Experimental results show that the performance of normalized-by-top-k merging with translation penalty and collection weight is similar to that of raw-score merging and better than that of the other merging strategies.

Keywords: *Merging Strategy, Multilingual Information Retrieval, Query Translation*

1. Introduction

Multilingual Information Retrieval (MLIR) facilitates the uses of queries in one language to access documents in various languages. It attracts the attentions of researchers in recent years due to the fact that large amount of multilingual information is created and disseminated nowadays. How to retrieve multilingual information efficiently becomes indispensable in daily life. Most of the previous approaches [14] focused on how to unify the language usages in queries and documents. The adaptation of traditional information retrieval systems has been considered. Query translation and document translation methods have been introduced. The resources used in the translation have been explored.

There are two possible architectures in MLIR – say, centralized and distributed [10]. In a centralized architecture, document collections in different languages are viewed as a single document collection and are indexed in one huge index file. In contrast, documents in different languages are indexed and retrieved separately in a distributed MLIR. The ranked lists of all monolingual and cross-lingual runs are merged into one multilingual ranked list.

In real world, multilingual document collections are distributed from various resources, and managed by information retrieval system of various architectures. How to integrate the results from heterogeneous resources is one of the major issues in MLIR. Merging result lists of individual languages is a commonly adopted approach. The goal of result lists merging is to include as more relevant documents as possible in the final result list and to make relevant documents have higher ranks. Several attempts were done on this problem [7, 15, 16]. The simplest merging method is *raw-score merging*, which sorts all the documents by their original similarity scores, and then selects the top ranked documents. The second approach, *round-robin merging*, interleaves the results of each run based on the rank of each document. The third approach is *normalized-score merging*. For each topic, the similarity score of each document is divided by the maximum score in each result list. After adjusting scores, all results are put into a pool and sorted by the normalized score.

Moulinier and Molina-salgado [13] proposed collection-weighted normalized score to merge result lists. The normalized collection score is used to adjust the similarity score between a document and a query. Collection score only reflects the similarity of a (translated) query and a document collection. This method could fail if a query is not translated well. Savoy [18] used logistic regression to predict the relevance probability of documents according to the document score and the logarithm of the rank. This method does not consider the quality of query translation either. Furthermore, the relationship between the rank and the relevance of a document is not strong. Braschler, Göhring and Schäuble [2] proposed feedback merging that interleaves the results according to the propositions of the predicted amount of relevant documents in each document collection. The amount of relevant information was

estimated by the portion of overlap between the original query and the ideal query constructed from the top ranked documents. The experimental results showed that feedback merging had little impact.

Lin and Chen [9, 10, 11] proposed several merging strategies to integrate the result lists of document collections in different languages. We assume that the importance of each intermediate run depends on their retrieval performance. Three factors affecting the retrieval effectiveness, i.e., the degree of translation ambiguity, the number of unknown words and the number of relevant documents in a collection for a given query, were introduced. We proposed *normalized-by-top-k merging* to avoid the drawback of normalized-score merging. Translation penalty and collection weight are also considered during merging result lists. Experimental results showed that the performances of the proposed merging strategies were similar to that of raw-score merging and were better than that of normalized-score and round-robin merging in single IR system environment.

In this paper, we extend our past works and modify the methods for computing translation penalty and collection weight to improve merging performance. The rest of this paper is organized as follows. Section 2 describes our merging strategies. Section 3 describes the indexing method and query translation process. Section 4 shows the experiment results. Section 5 concludes the remark.

2. Merging strategies

We aim to include as more relevant documents as possible in the final result list and to make relevant documents have higher ranks during merging. We assume that the importance of each intermediate run depends on their retrieval performance. If a result list contains many relevant documents in the top ranks, i.e., it has good performance, the top ranked documents should be included in the final result list. On the other hand, if a result list has few or even no relevant documents, the final result list should not contain many documents from this list. Thus, the higher performance an individual run has, the more important it is. However, without the priori knowledge of a query in advance, it is challenging to predict the performance of an individual run for each document collection. The similarity score between a document and a query is one of a few clues that are common used. The basic idea of our merging strategies is adjusting the similarity scores of documents in each result list to make them more comparable and to reflect the confidence in retrieval effectiveness.

In our previous works [9, 10, 11], we proposed *normalized-by-top-k* to normalized the similarity scores of documents. The original score of each document is divided by the average score of top *k* documents in the result list. Translation penalty and collection weight are proposed to predict the retrieval effectiveness of intermediate runs. The similarity scores are adjusted by the following formula.

$$\hat{S}_{ij} = S_{ij} \times \frac{1}{\overline{S}_{ik}} \times W_i \quad (1)$$

where S_{ij} is the original similarity score of the document at rank j in the ranked list of query q_i,

$\hat{S}_{ij}$ is the adjusted similarity score of the document at rank j in the ranked list of query q_i,

$\overline{S}_{ik}$ is the average similarity score of top k documents in the ranked list of query q_i, and

W_i is the merging weight of query q_i in an intermediate run.

The merging weight W_i of an intermediate run is determined by translation penalty and collection weight which are discussed in the following subsections.

2.1 Translation penalty

Similarity score reflects the degree of similarity between a document and a query. A document with higher similarity score seems to be more relevant to the desired query. However, if the query is not formulated well, e.g., inappropriate translation of a query, a document with high score may still not meet user's information need. When the result lists are merged, those documents that have high, but incorrect scores should not be included in the final result list. Thus, the effectiveness of each individual run has to be considered in the merging stage.

When query translation method is used to deal with the unification of language usages in queries and documents, queries are translated into target language and then the target language documents are retrieved. We could predict the multilingual retrieval performance based on the translation quality. Intuitively, using Chinese to access Chinese collection is expected to have better performance than using it to access other collections. Similarly, using a bilingual dictionary of more coverage is expected to be better than that of less coverage. Less ambiguous queries have also higher tendency to achieve better translation than more ambiguous queries. Two factors, i.e., the degree of translation ambiguity and the number of unknown words, are used to model the translation performance. For each query, we compute the average number of translation equivalents of query terms and the number of unknown words in each language pair, and use them to compute the translation penalty of each cross-lingual run. The following formula is proposed to determine translation penalty.

$$W_{tpi} = c_1 + \left[c_2 \times \left(\frac{51 - T_i}{50} \right)^2 \right] + \left[c_3 \times \left(1 - \frac{\sum_{t=1}^{n_{ui}} w_{ut}}{n_i} \right) \right] \quad (2)$$

where W_{tpi} is the translation penalty of query q_i in a cross-lingual run,

T_i is the average number of translation equivalents of query terms in query q_i,

w_{ut} is the penalty weight of an unknown word,

n_{ui} is the number of unknown words in query q_i,

n_i is the number of query terms in query q_i, and

c_1, c_2 and c_3 are tunable parameters.

The importance of query terms in different syntactic categories should be different in CLIR. For example, if a proper noun in a query is not translated, the retrieval performance will be poor. On the other hand, if the unknown word is a preposition, it has little impact on retrieval performance. Therefore, we divide query terms into three classes and assign different weight to each class. The first class is named entities including person names, location names and organization names. These named entities are very important and assigned the highest weight. The second class contains nouns and verbs. Nouns and verbs are useful content words, but are not as important as named entities. The words in other syntactic categories belong to the third class which has lowest weight. The penalty weight of an unknown word C_{uj} is determined in the following way:

$$w_{uj} = \begin{cases} 1.5 & \text{if } C_{uj} \text{ is a person name, a location name or an organzation name} \\ 1 & \text{if } C_{uj} \text{ is a noun or a verb} \\ 0.5 & \text{otherwise} \end{cases} \quad (3)$$

The best case of query translation is that each query term has only one translation, that is, the average number of translation equivalents is 1, and the number of unknown words is 0. In such a case, query will be translated correctly, thus the value of translation penalty is 1. As the number of unknown words or average number of translation equivalents increases, the translation quality and retrieval performance are more likely to be worse. Therefore, the value of merging weight is decreased to reduce the importance of this intermediate run.

2.2 Collection weight of individual document collection

How many relevant documents there are in a collection for a given query is also an important factor for measuring retrieval effectiveness. If a document collection contains more relevant documents, it could have more contribution to the final result list. Callan, Lu and Croft [3] proposed CORI net to rank distributed collections of the same language for a query. Moulinier and Molina-salgado [13] used collection score to adjust the similarity score between a document and a query.

In our approach, we try to estimate the proportion of possible relevant documents of a query in a document collection. If the proportion of relevant documents is high, they have more chance be retrieved. Therefore, a document collection with higher portion of relevant documents is assigned higher weight. Given a query q_i, the collection weight of a document collection D_m is estimated by the following formula.

$$CW_{im} = \frac{avg\_DF_{im}}{N_m} \quad (4)$$

where CW_{im} is the collection weight of document collection D_m corresponding to query q_i,

avg_DF_{im} is the average document frequency of query terms in query q_i, and

N_m is the number of documents in collection D_m.

Combining translation penalty and collection weight, the merging weight W_i is measured by the following formula.

$$W_i = W_{tpi} + c_4 \times CW_{im} \quad (5)$$

where c_4 is a tunable parameter.

3. Indexing and query translation

The document set used in NTCIR-4 MLIR task consists of Chinese (C), Japanese (J), Korean (K), and English (E) documents [8]. The participants can use two types of multilingual document collections, i.e., CJKE and CJE collections, as the target language sets. We used CJE collection in the experiments.

Okapi IR system was adopted to index and retrieve documents. The weighting function was BM25 [17]. The <HEADLINE> and <TEXT> sections were used for indexing. For English, the words in these sections were stemmed, and stopwords were removed. Japanese documents were segmented by ChaSen [12]. Chinese documents were segmented by a word recognition system [6]. All words in the above two sections were used as index terms.

In the experiments, the Chinese queries were used as source queries and translated into target languages, i.e., English and Japanese. First, the Chinese queries were segmented by a word recognition system, and tagged by a POS tagger. Name entities were then identified [6]. For each Chinese query term, we found its translations by looking up bilingual dictionaries. The Chinese-English bilingual dictionary is integrated from four resources,

including the LDC Chinese-English dictionary, Denisowski's CEDICT [1], BDC Chinese-English dictionary v2.2 [2] and a dictionary used in query translation in MTIR project [1]. The Chinese-Japanese bilingual dictionary is Dr.eye dictionary[3].

For each query term, we used the first-two-highest-frequency method to select appropriate translations. The first two translations with the highest frequency in the target language document collection are considered as the target language query terms.

4. Experiment results

In NTCIR-4 MLIR task, we submitted four C→CJE multilingual runs. Two runs use <TITLE> field and the other two runs use <DESC> field only. The source Chinese queries and the translated English and Japanese queries were used to retrieve Chinese, English and Japanese documents, respectively. Then, we merged these three result lists. The details of the four runs are described as follows.

1. NTU-C-CJE-T-01
   This run used <TITLE> section only. Formula (1) was used to adjust the similarity score of each document. We used the average similarity score of top 10 documents for normalization. The merging weight W_i was determined by formula (5), i.e., combining translation penalty and collection weight. The values of c_1, c_2, c_3, and c_4 were set to be 0, 0.2, 0.5, and 0.3, respectively. After adjusting similarity score, all results were put in a pool and sorted by the adjusted score. The top 1000 documents were selected as the final results.
2. NTU-C-CJE-T-02
   This run used <TITLE> section only. The merging strategy is similar to run NTU-C-CJE-T-01 except that the merging weight W_i was determined by translation penalty only. The values of c_1, c_2 and c_3 were set to be 0, 0.4 and 0.6, respectively.
3. NTU-C-CJE-D-01
   This run used <DESC> section only. The merging strategy used in this run is the same as run NTU-C-CJE-T-01.
4. NTU-C-CJE-D-02
   This run used <DESC> section only. The merging strategy used in this run is the same as run NTU-C-CJE-T-02.

The results of our official runs are shown in Table 1. Table 2 shows the unofficial evaluation of intermediate monolingual (i.e., Chinese to Chinese) and cross-lingual runs (i.e., Chinese to Japanese and Chinese to English). The rigid relevant set was used to evaluate the performances. The performance of Chinese-Japanese cross-lingual run is poor. One of the reasons is that many query terms have no Japanese translation. In contrast, most query terms have English translations, thus the performance of Chinese-English cross-lingual run is better. The performance of Chinese monolingual run is not good enough. This is probably because that only 14 topics (Topics 001 - 014) are original in Chinese. Figure 1 shows the average precision of each topic in Chinese monolingual runs. The performances of Topics 001 - 014 are better than that of remaining Topics. In Topic 001 - 014, there are 9 and 8 topics which average precision is higher than 0.1 in run ntu-c-c-t and ntu-c-c-d, respectively. In the remaining 46 topics, only 13 and 16 topics have an average precision higher than 0.1 in run ntu-c-c-t and ntu-c-c-d, respectively.

In order to compare the effectiveness of different merging strategies, we also conducted several unofficial runs using the following merging strategies.

1. ntu-c-cje-t-raw-score/ ntu-c-cje-d-raw-score
   We used raw-score merging to merge result lists.
2. ntu-c-cje-t-normalized-score/ ntu-c-cje-d-normalized-score
   The result lists were merged by normalized-score merging strategy. The maximum similarity score was used for normalization.
3. ntu-c-cje-t-round-robin/ ntu-c-cje-d-round-robin
   We used round-robin merging to merge result lists.
4. ntu-c-cje-t-normalized-top10/ ntu-c-cje-d-normalized-top10
   In this run, we used normalized-by-top-*k* merging. The average similarity score of top 10 documents was used for normalization.

The average precision of optimal merging proposed by Chen [4] was regarded as an upper-bound, which was used to measure the performances of our merging strategies. Given the relevance judgments of documents, optimal merging can produce the best merging result under the constraint that the relative ranking of the documents in the individual ranked lists is preserved. The performances of the unofficial runs are shown in Table 3. We used the rigid relevant set to evaluate the unofficial runs. The performance relative to optimal merging is enclosed in parentheses.

[1] The dictionary is available at http://www.mandarintools.com/cedict.html

[2] The BDC dictionary is developed by the Behavior Design Corporation (http://www.bdc.com.tw)

[3] The Dr.eye dictionary is developed by Inventec Corporation (http://www.dreye.com)

Table 1. The results of official runs

Run	Scoring Mode	Average Precision	Recall	F-score
NTU-C-CJE-T-01	Rigid	0.0640	3411 / 14328	0.1009
	Relax	0.0838	5916 / 24688	0.1242
NTU-C-CJE-T-02	Rigid	0.0629	3386 / 14328	0.0994
	Relax	0.0809	5765 / 24688	0.1202
NTU-C-CJE-D-01	Rigid	0.0521	2874 / 14328	0.0827
	Relax	0.0625	4837 / 24688	0.0948
NTU-C-CJE-D-02	Rigid	0.0519	2858 / 14328	0.0824
	Relax	0.0609	4739 / 24688	0.0925

Table 2. The performances of intermediate runs

Run	# Topic	Average Precision	Recall	F-score
ntu-c-c-t	59	0.1158	622 / 1318	0.1860
ntu-c-j-t	55	0.0399	1214 / 7137	0.0646
ntu-c-e-t	58	0.1303	3006 / 5866	0.2078
ntu-c-c-d	59	0.1273	733 / 1318	0.2072
ntu-c-j-d	55	0.0319	1268 / 7137	0.0541
ntu-c-e-d	58	0.1061	2412 / 5866	0.1687

Table 3. The results of unofficial runs

Run	Average Precision	Recall	F-score
ntu-c-cje-t-raw-score	0.0697 (71.71%)	3433 / 14328	0.1080 (75.42%)
ntu-c-cje-t-normalized-score	0.0529 (54.42%)	3112 / 14328	0.0851 (59.43%)
ntu-c-cje-t-round-robin	0.0481 (49.49%)	3083 / 14328	0.0786 (54.89%)
ntu-c-cje-t-normalized-top10	0.0541 (55.66%)	3188 / 14328	0.0870 (60.75%)
c-cje-t-optimum merging	0.0972	3892 / 14328	0.1432
ntu-c-cje-d-raw-score	0.0503 (61.95%)	2796 / 14328	0.0800 (66.28%)
ntu-c-cje-d-normalized-score	0.0473 (58.25%)	2857 / 14328	0.0765 (63.38%)
ntu-c-cje-d-round-robin	0.0434 (53.45%)	2609 / 14328	0.0701 (58.08%)
ntu-c-cje-d-normalized-top10	0.0466 (57.39%)	2839 / 14328	0.0755 (62.55%)
c-cje-d-optimum merging	0.0812	3371 / 14328	0.1207

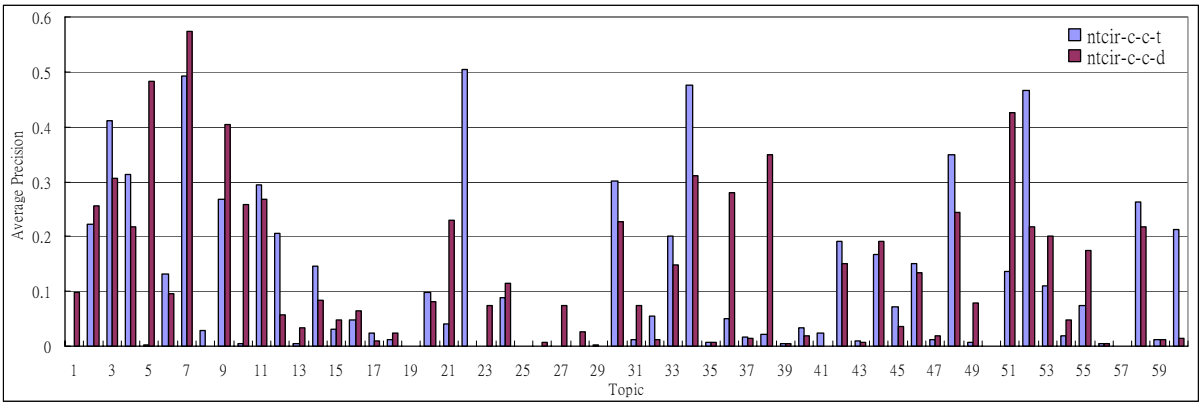

Figure 1. The average precision of each topic in Chinese monolingual runs

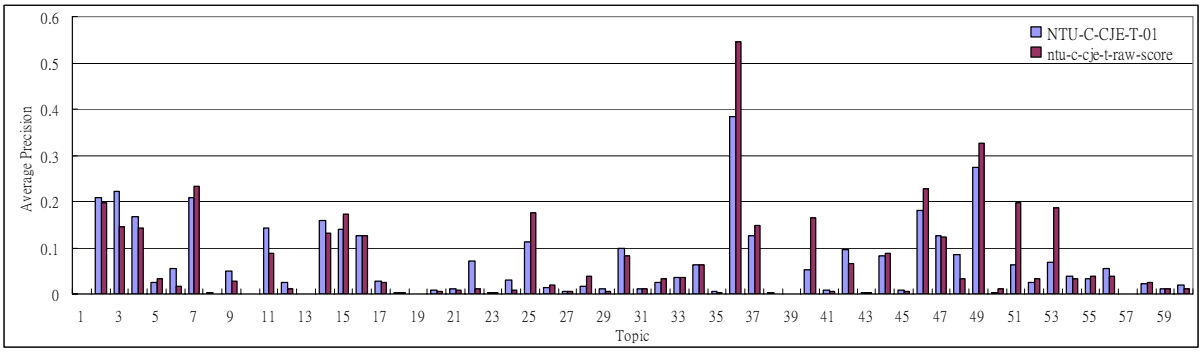

Figure 2. The average precision of each topic in run NTU-C-CJE-T-01 and ntu-c-cje-t-raw-score

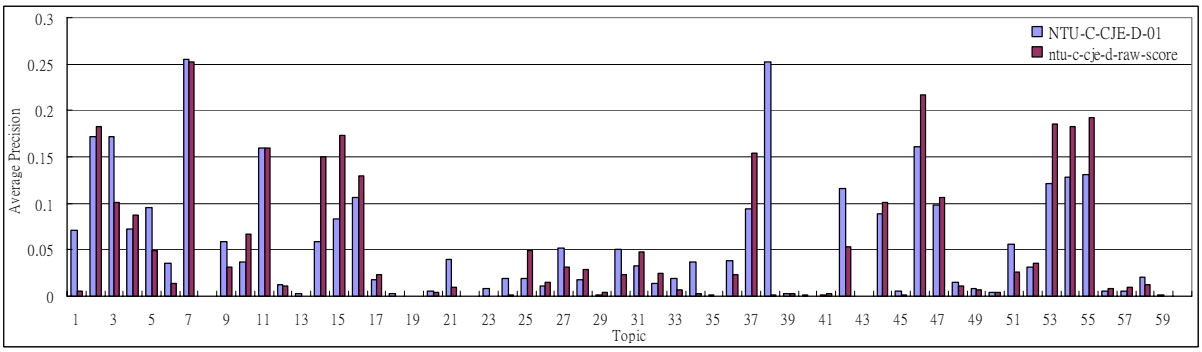

Figure 3. The average precision of each topic in run NTU-C-CJE-D-01 and ntu-c-cje-d-raw-score

From Table 1, the performance of run NTU-C-CJE-T-01 is slightly better than that of run NTU-C-CJE-T-02. Similarly, run NTU-C-CJE-D-01 is slightly better than run NTU-C-CJE-D-02. It shows that the contribution from collection weight is limited. The performances of the runs using title field are better than that of the runs using description field. When using title field, raw score merging performs best. Normalized-by-top-*k* merging with translation penalty and collection weight is slightly worse than raw-score merging. The performances of normalized-score and round-robin merging are worse than raw-score merging and our approaches. In the experiments that using description field, normalized-by-top-*k* merging with translation penalty and collection weight performs better than raw-score merging. However, the difference is not significant.

We further analyze the performances of our approaches query by query. There are 32 and 30 topics that Normalized-by-top-*k* merging with translation penalty and collection weight performs better than raw-score merging when using title and description field, respectively. In contrast, there are 27 and 25 topics that raw-score merging performs better. Figures 2 and 3 show the average precision of each topic in title and description experiments, respectively. For many queries, the difference in average precision between these two approaches is large. We found that if the performance of a topic in Chinese monolingual run is good, normalized-by-top-*k* merging with translation penalty and collection weight is likely better than raw-score merging. Take Topic 005 as an example. When using description field, the average precision of Chinese-Chinese, Chinese-Japanese and Chinese-English runs are 0.4820, 0.0161 and 0.1755, respectively. The performance of normalized-by-top-*k* merging with translation penalty and collection weight (0.0952) is better than that of raw-score merging (0.0489). This is because that the similarity scores of Chinese documents are smaller than that of English and Japanese documents. Raw-score merging will prefer English and Japanese documents, thus irrelevant documents are placed in higher rank. In run NTU-C-CJE-D-01, similarity scores are normalized, and the merging weight of Chinese monolingual run is higher than that of cross-lingual runs. Therefore, Chinese documents are preferred.

On the other hand, if the performance of Chinese monolingual run is poor, and either cross-lingual run performs well, raw-score merging performs better. As described above, the similarity scores of Chinese documents are smaller, thus they are placed in lower ranks when using raw-score merging. Since monolingual run does not have translation penalty, its merging weight is high. Chinese documents result in higher ranks in our approach.

5. Concluding Remarks

Merging problem is critical in distributed multilingual information retrieval. This paper proposed several merging strategies to integrate the result lists of collections in different languages. The prediction of retrieval effectiveness is used to determine the merging weight of each intermediate run. We modify the methods for computing translation penalty and collection weight to improve merging performance.

The experimental results showed that the performance of normalized-by-top-*k* with translation penalty and collection weight was better than that of normalized-score merging and round-robin merging, and is similar to that of raw-score merging. When the performance of monolingual intermediate run is good, normalized-by-top-*k* with translation penalty and collection weight tends to be better than raw-score merging. When the performance of monolingual run is poor, the merging weight is overweighted. How to estimate the merging weight of monolingual run more precisely will be further investigated.

References

[1] Bian, G.W. and Chen, H.H. Cross Language Information Access to Multilingual Collections on the Internet. *Journal of American Society for Information Science*, 51(3):281-296, 2000.

[2] Braschler, M., Göhring, A. and Schäuble, P. Eurospider at CLEF 2002. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.), *Advances in Cross-Language Information Retrieval - Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002*. Lecture Notes in Computer Science, Vol. 2785. Springer, 2003 (pp. 164-174).

[3] Callan, J.P., Lu, Z. and Croft, W.B. Searching Distributed Collections With Inference Networks. In Fox, E.A., Ingwersen, P. and Fidel, R. (Eds.) *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 1995 (pp. 21-28).

[4] Chen, A. Cross-language Retrieval Experiments at CLEF-2002. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.), *Advances in Cross-Language Information Retrieval - Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002*. Lecture Notes in Computer Science, Vol. 2785. Springer, 2003 (pp. 28-48).

[5] Chen, H.H., Bian, G.W. and Lin, W.C. Resolving translation ambiguity and target polysemy in cross-language information retrieval. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1999 (pp. 215-222).

[6] Chen, H.H., Ding, Y.W, Tsai, S.C., and Bian, G.W. Description of the NTU System Used for MET2. In *Proceedings of 7th Message Understanding Conference*. 1998.

[7] Kando, N., Oyama, K. and Ishida, E. (Eds.). *Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering*. Tokyo, Japan: National Institute of Informatics, 2003.

[8] Kishida, K., Chen, K.H., Lee, S., Kuriyama, K., Kando, N., Chen, H.H., Myaeng, S.H., and Eguchi, K. Overview of CLIR Task at the Forth NTCIR Workshop. In this volume.

[9] Lin, W.C. and Chen, H.H. Description of NTU Approach to NTCIR3 Multilingual Information Retrieval. In Kando, N., Oyama, K. and Ishida, E. (Eds.), *Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering*. Tokyo, Japan: National Institute of Informatics, 2003.

[10] Lin, W.C. and Chen, H.H. Merging Mechanisms in Multilingual Information Retrieval. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.), *Advances in Cross-Language Information Retrieval - Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002*. Lecture Notes in Computer Science, Vol. 2785. Springer, 2003 (pp. 175-186).

[11] Lin, W.C. and Chen, H.H. Merging Results by Predicted Retrieval Effectiveness. In Peters, C. (Ed.), *Results of the CLEF 2003 Cross-Language System Evaluation Campaign*. 2003.

[12] Matsumoto, Y., Kitauchi, A., Yamashita, T., and Hirano, Y. Japanese Morphological Analysis System ChaSen version 2.0 Manual. Technical Report NAIST-IS-TR99009, Nara Institute of Science and Technology Technical Report, 1999.

[13] Moulinier, I. and Molina-salgado H. Thomson Legal and Regulatory experiments for CLEF 2002. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.), *Advances in Cross-Language Information Retrieval - Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002*. Lecture Notes in Computer Science, Vol. 2785. Springer, 2003 (pp. 155-163).

[14] Oard, D. and Diekema, A. Cross-Language Information Retrieval. *Annual Review of Information Science and Technology*, 33:223-256, 1998.

[15] Peters, C. (Ed.). *Results of the CLEF 2003 Cross-Language System Evaluation Campaign*. 2003.

[16] Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.). *Advances in Cross-Language Information Retrieval - Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002*. Lecture Notes in Computer Science, Vol. 2785. Springer, 2003.

[17] Robertson, S.E., Walker, S. and Beaulieu, M. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive. In Voorhees, E.M. and Harman, D.K. (Eds.), *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*. National Institute of Standards and Technology, 1998 (pp. 253-264).

[18] Savoy, J. Report on CLEF 2002 Experiments: Combining Multiple Sources of Evidence. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.), *Advances in Cross-Language Information Retrieval - Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002*. Lecture Notes in Computer Science, Vol. 2785. Springer, 2003 (pp. 66-90).