CLARITECH

CLARIT TREC-8

CLIR Experiments

(The Eighth Text REtrieval Conference)

Yan Qu, Hongming Jin,

Alla N. Eilerman, Emilia Stoica, David A. Evans

January 31, 2000

Report No. CC-TR-00-03

CLARITECH Corporation
Technical Report

5301 Fifth Avenue
Pittsburgh, PA 15232-2124
(412) 621-0570

To Be Published As:

[Qu et al. 2000] Qu, Y., H. Jin, A. Eilerman, E. Stoica, and D. A. Evans (2000). CLARIT TREC-8 CLIR Experiments. In E.M. Voorhees and D.K. Harman (Eds.), *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, (NIST Special Publication) Washington, DC: U.S. Government Printing Office.

CLARIT TREC-8 CLIR Experiments

Yan Qu, Hongming Jin, Alla N. Eilerman, Emilia Stoica, David A. Evans

CLARITECH Corporation

Abstract In the TREC-8 cross-language information retrieval (CLIR) track, we adopted the approach of using machine translation to prepare a source-language query for use in a target-language retrieval task. We empirically evaluated (1) the effect of pseudo relevance feedback on retrieval performance with two feedback vector length control methods in CLIR and (2) the effect of multilingual data merging either before or after retrieval. Our experiments show that, in general, pseudo relevance feedback significantly improves cross-language retrieval performance, and that post-retrieval merging of retrieval results can outperform pre-retrieval merging of multilingual data collections.

1 Introduction

TREC-8 marks the first occasion for CLARITECH to participate in the CLIR track. For commercial reasons, we have developed technology for English, Japanese, and Chinese CLIR. With our TREC-8 submission, we are in a position to assess how well our techniques extend to European languages.

Our approach to CLIR takes advantage of machine translation (MT) to prepare a source-language query for use in a target-language retrieval task. We developed a parameterized cross-language retrieval evaluation environment, integrating the functionality of natural language processing, retrieval, (pseudo) relevance feedback, feedback vector length optimization, MT, and data merging. For MT, we use SYSTRAN Enterprise, a commercial client-server based translation product.

Pseudo relevance feedback (PRF) has been shown, in general, to improve retrieval performance in monolingual and in cross-language retrieval using bilingual dictionaries (Ballesteros & Croft 1996). In CLIR, feedback-based query expansion can occur before query translation, after query translation, or at both places. In our pre-TREC-8 experiments, we observed that, in general, pseudo relevance feedback significantly improved retrieval performance for all the selected language pairs (English-French, English-German, and English-Italian). We calibrated our system with TREC-6 and TREC-7 CLIR topics to determine the optimal points for pseudo relevance feedback and the optimal parameter settings for the individual language pairs.

In our TREC-8 submissions, we compared two methods for controlling feedback vector length: one with a uniform number of thesaurus terms for all the topics, and the other with a varying (query-dependent) number determined by vector length optimization.

Multilingual data merging needs to be addressed in this work because the CLIR track requires a single ranked list of retrieved documents from data collections in four languages. We distinguish pre-retrieval and post-retrieval data merging methods. Pre-retrieval data merging refers to the merging of data collections in different languages into a single multilingual data collection, while post-retrieval data merging refers to the merging of retrieval results obtained from separate data collections in different languages. Retrieval from a merged multilingual collection using multilingual topics eliminates the need for merging retrieval results, but the method can degrade the system's capability to process individual languages optimally. The post-retrieval merging method, on the other hand, allows optimization of retrieval performance for each language pair, but it requires merging of retrieval results. Our TREC-8 results show that post-retrieval merging of retrieval results can outperform pre-retrieval merging of multilingual data collections.

In the following sections, we first describe the system and the language resources employed for the TREC-8 CLIR track. Then we describe our experiments with pseudo relevance feedback and experiments in multilingual data merging, and present the evaluation results. Finally, we summarize our work.

2 System Description

We adopted MT-based query translation as our way of bridging the language gap between the source language (SL) and the target language (TL).

We implemented three methods of pseudo relevance feedback (PRF) for bilingual retrieval. The simple MT-based query translation and the PRF methods are illustrated in Figure 1. Figure 1(a) illustrates query translation without expansion. In this configuration, the topics in a source language are translated using the MT engine into texts in the designated target language, which are then used for retrieval from a target language database. Figure 1(b) illustrates query expansion prior to translation. Here each topic in a source language (SL) is first augmented with N thesaurus terms extracted from the top M subdocuments retrieved from a SL database. The top M subdocuments are assumed to be relevant to the query. The resulting topic, which consists of the original query text and the additional thesaurus terms in SL, is then sent to the MT engine. The translation of the source language query text

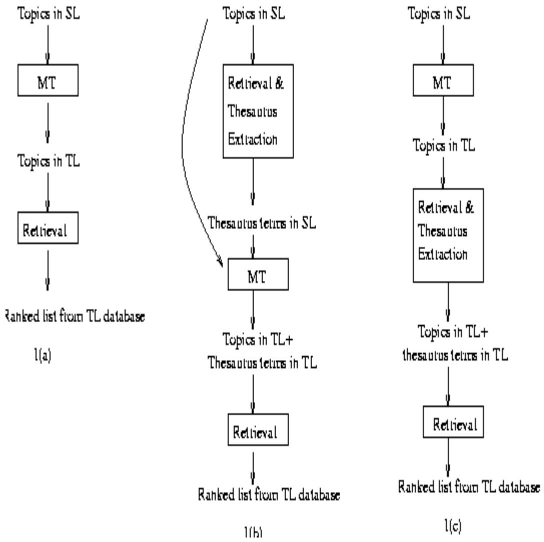

Figure 1. CLIR with MT-based query translation and pseudo relevance feedback

and the thesaurus terms is used for retrieval from a target language database. In post-translation query expansion (Figure 1(c)), the original query text is first translated via the MT engine. Then the translated query text is augmented using the feedback process. The resulting topic, which consists of the translated query topic and the thesaurus terms in the TL, is then used for retrieval from a TL database. The combined feedback method unites the feedback process prior to translation in Figure 1(b) and the feedback process after translation in Figure 1(c). For details on the CLARIT term extraction methods and the retrieval engine, the reader is referred to (Milic-Frayling et al. 1998).

For CLIR involving more than two languages, we decompose the task into bilingual retrieval from the source language to the individual target languages, then merge the retrieval results.

3 Linguistic Resources

For processing the English corpus and queries, we used the CLARIT English NLP module, which consists of a parser and a morphological analyzer that utilize the English lexicon and grammar to identify linguistic structures in texts (Milic-Frayling et al. 1998). The CLARIT NLP module supports discovery of various types of linguistic structures, such as simplex and complex noun phrases (NPs), verbs, and other selected constituents.

The English grammar was adapted for use in German, French, and Italian NLP. Necessary modifications were made to accommodate specific categories of each language.

For German NLP, we automatically extracted a core lexicon from the German lexicon distributed by the Linguistic Data Consortium (LDC). The resulting lexicon, with 318,809 entries, specifies word surface forms, their parts of speech, and normal forms.

For French and Italian NLP, we manually developed lexicons of closed-class categories that are sufficient to achieve mostly correct phrase segmentation. In addition, punctuation marks and special symbols, as found in multilingual texts, were collected and used to supplement the core lexicons. No morphological normalization was done for either language, even though a design for the French normalization had been completed.

For all four languages, we also manually constructed lexicons of stop words, which included extraneous words and their inflected forms (e.g., *document, relevant, report* in English; *document, pertinent, rapport, rapporter* in French; *Dokument, relevant, Bericht* in German; and *documento, rilevante, rapporto* in Italian). The stop words were selected from the TREC-6 and TREC-7 topics.

For all the experiments reported in this paper, we indexed the data collections of individual languages using simplex NPs and all attested sub-terms. The English topics and their French/German/French translations were processed similarly into simplex NPs and decomposed into all attested sub-terms.

We used the SYSTRAN Enterprise software for translating the queries. The client-server configuration of this software allows us to integrate SYSTRAN's translation capability into our evaluation environment by calling the client API. The client API takes as input the source language query (plus feedback terms if feedback is used) stored in a file and the specific language pair for translation, and returns a file with the translation of the source text to the application program. Query translation is a black box process to the application program. The language pairs selected for the TREC experiments included English-French, English-German, and English-Italian.

4 Pre-TREC System Calibration

In preparation for the TREC-8 CLIR track, we performed experiments to calibrate the components of the CLIR evaluation environment. We focused on testing the effectiveness of pseudo relevance feedback on the English monolingual retrieval and the English-to-French/German/Italian bilingual retrieval. Two feedback-vector length control methods were tuned: one with uniform vector length, i.e., the same number of feedback terms for all topics, and the other with varying vector length optimized for individual topics. We used Rocchio as the thesaurus term extraction method, as it was observed to generate the greatest improvement of retrieval performance in our previous monolingual experiments. We conducted experimental runs over TREC-6 and TREC-7 CLIR topics to obtain: (1) the optimal place for pseudo relevance feedback, and (2) the optimal parameter settings for the two feedback vector length control methods.

4.1 Feedback with uniform-length vectors

For PRF using uniform-length vectors, we focused on two parameters: (1) N_d, the number of subdocuments selected for thesaurus extraction, and (2) N_t, the number of terms extracted from the set of subdocuments for augmenting the original query vector. For bilingual retrieval, we also tested the optimal place (pre-translation, post-translation, and combined positions) or pseudo relevance feedback should apply. The experiments were conducted using English as the query language. For the English monolingual retrieval and the English-to-French/German bilingual retrieval, we used TREC-6 English topics, and evaluated the results using relevance judgments for TREC-6 topics. For the English-to-Italian bilingual retrieval, we used TREC-7 English topics and TREC-7 relevance judgments.[1]

In the English-to-French/German/Italian bilingual retrieval, we observed that, for all the language pairs, all three pseudo relevance feedback methods significantly improved Average Precision and Recall, compared to their respective no feedback (NF) baseline runs. In particular, post-translation query expansion yielded the greatest improvement in both average precision and recall for all the language pairs. The optimal settings obtained from the calibration are N_d = 50 subdocuments and N_t = 75 terms for English monolingual, and N_d = 25 subdocuments and N_t = 50 terms for English-to-French/German/Italian, respectively.

4.2 Feedback with optimized-length vectors

The uniform-length vector method adds the same number of terms to each profile. In contrast to this, the optimized-length vector method dynamically computes the number of terms to be added for query expansion, using the curve of terms' weights. The algorithm was developed based on the observation that there seems to be a correlation between the change in slope of the curve of the terms' weights and the average precision of a query.

The algorithm uses the first N weights (arranged from highest to lowest weights), and adds a term for query expansion if its weight satisfies the following condition:

$w(t) >= min + perc * (max-min)$

where *min* is the smallest weight, *max* is the largest weight, and *perc* is a constant. The method aims to provide the maximum benefit from feedback, while reducing the number of terms required for feedback.

For the optimized-length vector method, we tuned two parameters: (1) N_m, the maximum number of terms extracted from a set of subdocuments, for which we experimented with N as 80 and 250, and (2) *perc*, for which we tried the values 0.25, 0.1, 0.05, and 0.01. For bilingual retrieval experiments, we selected the top 25 subdocuments to be used for term extraction and the post-translation feedback method, as they were observed to give the best retrieval performance in general for the uniform-length vector method. The training experiments were conducted using English as the query language. We used the same set of topics and data collections as described in section 4.1. Compared with the experiments with no feedback (NF), PRF using the optimized-length vector method also demonstrated significant improvements in Average Precision and Recall. The optimal settings obtained from the calibration are *perc* = 0.25 and N_m = 250 terms for English monolingual, and *perc* = 0.05 and N_m = 80 terms for English-to-French/German/Italian, respectively.

5 TREC-8 Experiments

All of our CLIR submissions used automatic query processing, with English as the topic (source) language and with the combined fields of title, description, and narrative as the body of the query.

5.1 Experiments using Pseudo Relevance Feedback

To evaluate the effectiveness of the two vector control approaches in CLIR, we conducted a baseline run (*CLARITrmnf*) by first obtaining the French, German, and Italian translations of the source English topics, and then performing mono-lingual retrieval for the four languages from their respective databases without using any feedback mechanism. Then we combined the four retrieved result lists into a combined result list using their raw similarity scores.

We submitted two runs, *CLARITrmwf1* (PRF with uniform-length vectors) and *CLARITrmwf3* (PRF with optimized-length vectors), to compare the effectiveness of vector length optimization. With both runs, we first indexed each data collection individually, and obtained a ranked list from each collection. The result lists were then merged based on raw similarity scores. The two runs were conducted using the PRF settings specified in sections 4.1 and 4.2.

Our experiments in Table 1 demonstrated that, in general, PRF using both vector length control methods improved retrieval performance. In particular, both methods yielded significant improvement in Recall, Average Precision, Exact Precision, and Precision at 100 documents. Only Initial Precision decreased. The optimized-length vector method outperformed the uniform-length vector method in Average Precision, Initial Precision, and Exact Precision, but underperformed the uniform-length method in Recall and Precision at 100 documents. Such results are consistent with our observations with TREC-7 topics. Vector-length

[1] However, since the relevance judgments for TREC-7 topics were made based on the combined result list rather than results for individual languages, we treated the Italian results only as suggestive.

optimization seems to be a promising technique, but requires more research into its effectiveness.

5.2 Experiments on Data Merging

We evaluated three multilingual data merging methods to obtain a single ranked list for the purpose of TREC-8 CLIR track submission.

The first experiment (*CLARITdmwf*) used pre-retrieval data merging, i.e., we merged collections of English, French, German, and Italian documents into a single multilingual data collection, and indexed the multilingual collection. The topics were translated from the source language to the target languages and were merged together to form multilingual topics. Retrieval was done using the multilingual topics to obtain a single result list from the multilingual data collection. Pseudo relevance feedback was conducted for obtaining the optimal retrieval performance. For text processing, we used a combined lexicon consisting of all the lexicons for four languages and an adapted version of the English grammar. This run was designed as a baseline to be compared with two runs using post-retrieval result merging (*CLARITrmwf1* and *CLARITrmwf2).*

In *CLARITrmwf2*, we used normalized similarity scores rather than raw similarity scores as in *CLARITrmwf1*. First, we indexed each collection individually and obtained a ranked list from each collection. Then we reconstructed new databases using the N documents from each ranked list (in TREC, N = 1000) and re-computed the similarity scores for each new database. We then merged ranked lists into a single combined ranked list based on the recomputed similarity scores. Table 2 presents the retrieval performance statistics for the three runs.

The performance statistics demonstrate that post-retrieval merging of retrieval results can outperform pre-retrieval merging of data collections. Specifically, post-retrieval merging significantly improved Recall, Average Precision (except in *CLARITrmwf2*), and Exact Precision. Initial Precision was decreased for both set of topics.

The experimental results for TREC-8 topics are consistent with our observations with TREC-7 topics: merging with score normalization underperformed merging using raw similarity scores. One possible reason is that the new databases for score re-computation are too small (i.e., N = 1000 documents) for the similarity scores to be reliable. Another possible reason is that in the CLARIT system, the *idf* scores are computed using subdocuments. If the document lengths vary greatly across databases, the number of subdocuments used for *idf* computation will vary greatly even when the number of documents selected is uniform across databases. We intend to do further research on this issue in our future work.

6 Summary

Our TREC-8 experiments demonstrated that pseudo relevance feedback can be used to improve retrieval performance significantly in MT-based CLIR. The feedback vector length optimization method yields promising results, but requires more research into its effectiveness.

Post-retrieval result merging allows the optimization of retrieval performance for each language pair and has been demonstrated to outperform the pre-retrieval data merging method. However, effective techniques for score normalization for result merging require further investigation.

References

[Ballesteros & Croft, 1996] Ballesteros, L., and Croft, W.B. 1996. Dictionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems*, 791—801.

[Milic-Frayling et al. 1998] Milic-Frayling, N.; Zhai, C.; Tong, X.; Jansen, P.; and Evans, D. A. 1998.Experiments in query optimization, the CLARIT system TREC-6 report. In *Proceedings of the 6th Text REtrieval Conference (TREC-6)*, 415—454.

Run	Recall	Avg. Precision	Initial Precision	Exact Precision	Prec. 100 docs
1. CLARITrmwf1 (incr./Decr. Over (3))	1807 (13.5%)	0.2297 (25.0%)	0.6198 (-10.9%)	0.2717 (24.0%)	0.2661 (23.5%)
2. CLARITrmwf3 (incr./Decr. Over (3)) (incr./Decr. Over (1))	1789 (12.4%) (-1.0%)	0.2357 (28.3%) (2.6%)	0.6865 (-1.3%) (10.8%)	0.2809 (28.2%) (3.4%)	0.2475 (14.9%) (-7.0%)
3. CLARITrmnf (unofficial, baseline)	1592	0.1837	0.6953	0.2191	0.2154

Table 1: Performance statistics for CLARITrmwf1, CLARITrmwf3, and CLARITrmnf

Run	Recall	Avg. Precision	Initial Precision	Exact Precision	Prec. 100 docs
1. CLARITrmwf1 (incr./Decr. Over (3))	1807 (25.8%)	0.2297 (8.0%)	0.6198 (-7.9%)	0.2717 (9.9%)	0.2661 (17.9%)
2. CLARITrmwf2 (incr./Decr. Over (3)) (incr./Decr. Over (1))	1626 (13.2%) (-10.0%)	0.2036 (-4.3%) (-11.4%)	0.6032 (-10.3%) (-2.7%)	0.2514 (1.7%) (-7.5%)	0.2429 (7.6%) (-8.7%)
3. CLARITdmwf (baseline)	1436	0.2127	0.6726	0.2473	0.2257

Table 2: Performance statistics for CLARITrmwf1, CLARITrmwf2, and CLARITdmwf