Toward a Unified Retrieval Outcome Analysis Framework for Cross-Language Information Retrieval

Jiangping Chen

School of Library and Information Sciences, University of North Texas, P. O. Box 311068, Denton, TX 76203, jpchen@unt.edu

This paper proposes a Retrieval Outcome Analysis Framework, or ROA Framework, to systematically evaluate retrieval performance of Cross-Language Information Retrieval systems. The ROA framework goes beyond TREC-type retrieval evaluation methodology by including procedures focusing on individual queries, especially difficult queries. The framework is comprised of four interrelated components: (1) Overall System Performance Evaluation, (2) Query Categorization, (3) Translation Analysis, and (4) Individual Query Analysis. An example of applying the framework is discussed in detail. The author believes the proposed framework would be especially useful for the development of real-world Cross-Language Information Retrieval systems because the evaluation guided by the framework has the potential to discover causes behind poor retrieval performance.

Introduction

Cross-Language Information Retrieval (CLIR) is a special case of Information Retrieval (IR). It explores solutions to finding relevant documents in a collection of documents written in a different language or languages from users' queries. A CLIR system often behaves quite differently in response to different queries: The system retrieves relevant documents or web pages as top-ranked ones for some queries, but it fails to find any relevant documents, or ranks them very low, for some other queries. In the latter case, the users either cannot obtain the needed information, or they have to study the long list of returned documents to locate what they

want.

CLIR evaluation is an essential part of CLIR system design and development. A well-designed evaluation guided by sound methodology should be able to identify the strengths and the weaknesses of the system, especially the causes of unsatisfactory retrieval performance in response to certain queries, and to provide evidence for system improvement. However, current CLIR evaluation focuses more on the average performance over multiple topics than individual topic, just like monolingual IR system evaluation, as Hu, Bandhakavi, and Zhai have pointed out (2003). Few systems or researchers have performed systematic, in-depth analysis on individual queries or topics. In particular, researchers have paid little attention to those difficult queries or topics for which relevant documents or answers are not found or are ranked very low by IR systems or CLIR systems. Consequently, little is known about why some queries are more difficult then others. Current IR evaluation as conducted by TREC (http://trec.nist.gov/) may help the system to improve overall performance, but produces a limited effect on certain difficult queries because current TREC evaluations lack methods for performing in-depth retrieval analysis.

The researcher believes that it is necessary to explore methodological issues of conducting analysis at individual query level in order to understand the causes behind IR system performance. The investigation would benefit IR systems, especially real-world information access and retrieval systems, by allowing system designers to adjust their retrieval and user interaction strategies to provide better service for their users. In this paper, the author introduces a concept called *Retrieval Outcome Analysis* (ROA). ROA refers to a series of analytical procedures which systematically evaluate information retrieval on individual queries. In contrast to the traditional, TREC-like IR system evaluation paradigm, ROA focuses on exploring the causes behind retrieval performance on individual queries. A well designed ROA should provide more evidence to explain why a system performs well on certain topics and why it does poorly on some others, not just precision and recall scores.

In order to demonstrate the usefulness of the ROA and the procedures involved in it, the author proposes an ROA framework as a methodology for CLIR system evaluation. The ROA framework that is built upon the ROA concept will be presented and illustrated in the remaining part of this paper: The next section, "Related Research," reviews current IR system evaluation strategies and studies that have contributed to IR or CLIR performance analysis methodologies. The following section presents the ROA framework for CLIR. The fourth section provides an example to demonstrate the application of the ROA framework to the performance evaluation of an English-Chinese CLIR system. The paper concludes with a summary and future research direction.

Related Research

Modern IR system evaluation has been well established (Saracevic, 1995) and is mainly conducted in the context of Text Retrieval Conference (TREC). TREC provides test collections for comparing and evaluating IR and CLIR systems. A test collection typically consists of a large collection of documents (more than 1 million documents), a set of test topics (or queries), and their relevance judgments. The relevance judgments, which list all relevant documents for each test topic in the test collection, are TREC's major contribution to large-scale IR evaluation because IR experiments using the test collection can be evaluated automatically. The precision and recall scores and other statistics for the corresponding IR systems can be calculated automatically through a comparison of the retrieval results with the relevance judgments. One can immediately know the overall IR performance of the system from the scores.

Researchers have realized that overall performance measurements such as mean average precision do not help much on difficult queries. A special track--Robust Track--has been sponsored at TREC since 2003 to explore "methods for improving the consistency of retrieval technology by focusing on poorly performing topics" (Voorhees, 2003). Each year, the Robust Track has selected 50 difficult topics from previous TREC topics in addition to a number of new topics to be tested by participating systems. Participants have been required to explore solutions for difficult topics that obtained low median average precision scores but had one or more higher outliers from previous TREC ad hoc tracks (Voorhees, 2003). Also, two new measures have been introduced to indicate the performance of the systems on the difficult topics. One was the "percentage of topics that retrieved no relevant documents in the top ten retrieved"; and the other is a measure for the system's worst X topics, called "area underneath the MAP(X) vs. X curve" where MAP stands for the mean of the average precision score. Starting from the year 2004, systems are required to predict the difficulty level of the topics, which is quite challenging. The TREC Robust Track demonstrates that the IR community has noticed the importance of improving IR system performance on individual queries.

TREC test collection and system performance measures provide excellent resources for IR system evaluation. However, they are not sufficient to discover the causes behind system failure for certain difficult queries. The methodological issue of applying those instruments needs more investigation. The ROA framework described in the next section attempts to provide a strategy for more

effectively employing the TREC test collection and performance measures.

Hu, Bandhakavi, and Zhai (2003) investigated some difficult TREC topics. The questions they tried to explore included whether certain topics were hard for all the systems, and whether the document sets were a factor affecting the IR performance of those topics. Their analysis focused on the document sets. They proposed several measures to assess the influence of sub-TREC collections. They found that some topics received better IR performance from one document subset than others. However, it was still unclear what exactly made a topic difficult.

Diekema (2003) conducted her dissertation study to investigate translation events that affect cross-language information performance. Through content analysis of queries and their translations, she developed a taxonomy of translation events that might affect CLIR performance. Then she conducted and analyzed CLIR experiments to understand the role of the identified translation events in CLIR. In particular, she performed statistical analysis and qualitative query analysis. In the query analysis, the test queries were classified into 12 classes according to the difference in average precision between monolingual run and cross-lingual run, and each class was further examined to understand the causes of the differences in retrieval performance. The study found that translation events were not the only factors affecting CLIR performance. Some queries obtained better IR results even if the translation was poorer. The study concluded that further investigation would be needed to understand the nature of queries.

The above two studies applied both statistical and qualitative analysis techniques to analyze the information retrieval results, which provided some insight into the impact of individual queries on retrieval performance. However, none of them has focused on understanding why the system failed on certain queries. There have been no systematic strategies for analyzing retrieval results and discovering why some queries are more difficult than others for a system. IR experiments demonstrate that some techniques are helpful to some of the queries, but still, they cannot bring significant improvement to some consistently difficult queries. The ROA framework described below attempts to help the systems not only to improve performance on the difficult topics, but also to discover the causes accounting for the failure of the retrieval.

A Framework for CLIR Retrieval Outcome Analysis

The ROA framework for CLIR is designed to achieve the following objectives:

1. **to provide a mechanism to assess the overall performance of the system or IR strategy under investigation as well as to identify the behavior of the system for different types of queries;**
2. **to estimate the effectiveness of translation resources used by the system;**
3. **to identify possible causes behind varying system performance on individual queries; and**
4. **to recommend possible solutions to improve system performance.**

The proposed framework is comprised of four interrelated analytical components: Overall System Performance Evaluation, Query Categorization, Translation Analysis, and Individual Query Analysis. Table 1 summarizes the four components in terms of their purposes, the applicable analytical methods, and the conditions or prerequisites to perform the analysis. Following the table is a further explanation of each of the components.

Table 1. CLIR Retrieval Outcome Analysis Framework

Component	Purposes	Methods	Prerequisites
Overall System Performance Evaluation	Assess overall system performance and provide evidence for query categorization	IR measures such as average precision and significance testing	Test collection
Query Categorization	Facilitate individual query analysis	Classification based on average precision or other criteria	Overall system performance measure
Translation Analysis	Evaluate translation resources and effectiveness involved in CLIR	Manual or automatic judgment on translation correctness	Bilingual evaluator(s) or parallel queries or topics
Individual Query Analysis	Identify and understand causes for system performance on different queries	Miscellaneous	Relevance judgment

Overall System Performance Evaluation

Many systems have conducted Overall System Performance Evaluation using TREC test collection: applying certain appropriate measures to test queries and obtain a score for system performance. This component serves two purposes: One is to test system correctness. Some obvious mistakes which lead to abnormal system performance can be easily identified before spending time on further analysis. The other is to provide criteria for the Query Categorization component. The individual performance measures from which the overall performance is summarized can be used as criteria or as part of the criteria for classifying queries in the next component.

Precision and Recall are widely used measures for IR systems. In addition to the mean average precision score over all topics,

experiments with TREC test collection can be evaluated automatically and 11 measures can be obtained for each topic. Some aspects of the TREC evaluation paradigm have been criticized, such as unrealistic assumptions and the relevance pooling approach (Saracevic, 1995; Blair, 2002). However, the advantages of using TREC test collection are obvious: the evaluation is fast and well accepted by the IR community. This paper will use TREC test collections as instruments to illustrate the application of the framework in next section. User-centered performance evaluation (Dalrymple & Roderer, 1994) should also be able to serve as evidence for system correctness and query categorization, but needs further exploration.

If TREC test collection is employed, methods to achieve an overall system performance assessment include two steps:

1. Perform the TREC standard system evaluation based on average precision. Using the TREC-provided evaluation program, an average precision score can be obtained for each topic, and a Mean Average Precision (MAP) score can be calculated for the system or for a chosen IR strategy in testing.
2. Conduct statistical significance testing. Normally, IR experiments are conducted to compare different systems or different IR strategies. In the case of CLIR, monolingual experiments are conducted as baselines for comparison with cross-lingual ones, and cross-lingual experiments using different translation resources and/or strategies are compared. Statistical significance testing can help to find out whether the systems or approaches under investigation are significantly different. Hull (1993) suggested several statistical tests to be used in IR experiment evaluation under different conditions, such as the paired t-test, the paired Wilcoxon signed-rank test, and the sign test for comparing two groups, and certain analysis of variance (ANOVA) techniques for multiple groups. The significance testing needs to be selected with caution.

Figure 1 presents the formula and major assumptions for applying the paired Wilcoxon test from Conover (1999). It's a non-parametric alternative to the t-test for correlated samples and can be applied to compare two IR results based on similar test topics. To calculate the test statistic, the Wilcoxon signed ranks test first ranks each value of the difference between two group experiments, then the signs (plus or minus) of the difference are assigned to the ranks. The test statistic is obtained by dividing the sum of the ranks by its expected value assuming the two groups are equal.

The Wilcoxon signed ranks test

$$T = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}}$$

Where: R_i = the rank assigned to (X_i, Y_i) if $D_i = X_i - Y_i > 0$

R_i = the negative of the rank assigned to (X_i, Y_i) if $D_i = X_i - Y_i < 0$

Assumption: the distribution of each D_i is symmetric

Figure 1. Wilcoxon signed ranks test Summarized from (Conover 1999, p.353)

Query Categorization

Query Categorization classifies test topics into different categories based on their performance scores obtained in the previous component. Here, a simple faceted classification scheme is proposed. Two facets are considered useful: difficulty and stability. Difficulty reflects how hard it is for the query to return relevant documents. It includes three values: easy, moderate, and hard. Stability measures how stably different systems perform on a particular query. It has two values: stable and unstable. Altogether, a test query or topic can be classified into one of the following six categories: easy & stable, easy & unstable, moderate & stable, moderate & unstable, hard & stable, hard & unstable. In actual operation of the classification scheme, certain thresholds need to be determined for each category. Table 2 presents the classification scheme and the categorization approaches for each category under each facet.

Table 2. Query Categorization Scheme and Possible Strategies

Facet	Category	Categorization Approach
Difficulty	Easy	Queries whose average precision is above certain threshold P_{top}
	Moderate	Queries whose average precision is between thresholds P_{top} and P_{bottom}
	Hard	Queries whose average precision is below certain threshold P_{bottom}
Stability	Stable	Queries of which the differences in average precision scores between experimental groups are within a predefined range.
	Unstable	Otherwise

Translation Analysis

The third component of the ROA framework is Translation Analysis. This component is not necessary for monolingual ROA, but necessary and important for CLIR. A CLIR system typically involves either query translation, which translates users' queries into the language of the documents, or document translation, which translates document collection into the language of the queries. Translation has been considered the major source of the performance difference between a CLIR system and its monolingual counterpart. Translation Analysis aims at assessing the translation resource used by the system and providing evidence for the impact of translation to CLIR. In the case of query translation, the translation analysis can find out how correctly or accurately the translation is performed on the queries.

Translation Analysis typically includes selecting appropriate performance measures, conducting translation evaluation, and performing statistical analysis. The translation evaluation can be conducted manually, automatically, or in a hybrid mode depending on the resources available. If the documents or queries are parallel texts, a computer program can be developed to perform automatic evaluation, or one can apply the popular machine translation (MT) evaluation approach (Papineni, Roukos, Ward & Zhu, 2002). But we need to realize that most CLIR systems apply word-level or phrase-level translation, which is different from real MT systems which typically allow a set of terms with nearly the same meaning for each word or phrase. Finally, the same kinds of

statistical analysis used to evaluate overall system performance can be applied to assure the significance of translation differences among translation resources.

Individual Query Analysis

The above three components of ROA will provide a "big picture" of system retrieval performance and translation effectiveness. However, it is still unclear why a system would do well on certain queries, but poorly on some others. The Individual Query Analysis attempts to provide insights on this issue.

The methods for conducting individual query analysis can be quite diverse, depending on the time and effort allowed. Analysis may focus on the most interesting categories resulting from the Query Categorization component, such as queries which are consistently difficult for different systems or IR strategies. The characteristics of these queries, such as query length, questioner's background, translation accuracy, and document relevance can be analyzed statistically and/or interpretively. Usually, it is necessary and possible to select and control one or two important features to be explored according to the specific context and situation of the CLIR system. Statistical analysis, such as correlation analysis or regression analysis may help to identify variables that account for the differences among various experimental groups.

Applying the ROA Framework: An Example

In order to illustrate the application of the proposed ROA framework, the author applied the framework to analyze a set of English-Chinese Cross Language Information Retrieval (EC-CLIR) experiments. The specific goal of the retrieval outcome analysis is to systematically evaluate a translation resource for EC-CLIR. Additionally, it is expected that the analysis can provide insights which could be used to modify the system for performance improvement. The following will first briefly describe the experimental settings, then present the analysis following the four components of the proposed ROA framework.

General Information about the EC-CLIR System and Retrieval Experiments

An EC-CLIR system accepts English queries and searches for relevant documents from a Chinese collection. The EC-CLIR system, which is built upon a Vector-space IR model (Singhal, Buckley, and Mitra, 1996) and applied query translation strategy, employed TREC 5&6 Chinese test collection for all its experiments. The test collection contains 54 topics, a Chinese document collection, and relevance judgments. Each topic is bilingual with three parts: title, description, and narrative. The description provides descriptors and key terms for the topic, and the narrative contains the criteria for relevant documents. Figure 2 presents one of the 54 topics.

The original experiments were conducted to examine the effectiveness of a Lexical Knowledge Base (LKB) for EC-CLIR. The LKB is constructed by customizing available lexical resources based on the EC-CLIR system's document collection. (Chen, 2003). Four runs (a run is an execution of the retrieval program to return relevant documents for the 54 topics) were conducted and analyzed: a monolingual run and three cross-lingual runs. Each run used all three portions (title, description, and narrative) of a topic and were assigned a run tag for the convenience of analysis. The three cross-lingual runs used different translation resources, namely the LKB; a bilingual dictionary developed by LDC (http://www.ldc.upenn.edu) and used for constructing the LKB; and a machine translation system, Huajian (Kwok, 1997). The run tags for the four runs are: *mono_tdn* for monolingual run, *ldc_tdn* for cross-lingual run using the LDC dictionary, *lkb_tdn* for cross-lingual run using the LKB, and *mt_tdn* for cross lingual run using Huajian system.

<num> Number: CH29
<C-title> 信息高速公路的建设
<E-title> Building the Information Super Highway

<C-desc> Description:
信息高速公路，建设
<E-desc> Description:
Information Super Highway, building

<C-narr> Narrative:
相关文件应提到信息高速公路的建设，包括任何技术上的，或与信息基础设施有关的问题，以及有关发达国家或发展中国家对国际网络的应用计划.

<E-narr> Narrative:
A relevant document should discuss building the Information Super Highway, including any technical problems, problems with the information infrastructure, or plans for use of the Internet by developed or developing countries.

Figure 2. A Sample TREC Test Topic

Overall System Performance Evaluation

The TREC evaluation program was used to assess the overall system performance of the four runs. In addition, the Wilcoxon signed ranks test was applied to compare lkb_tdn with ldc_tdn, and lkb_tdn with mt_tdn. The evaluation results are summarized in Table 3.

Table 3. Overall System Performance

	mono_tdn	*mt_tdn*	*lkb_tdn*	*ldc_tdn*
Number of retrieved relevant documents	4427	3592	3702	3443
Mean Average Precision	0.4174	0.3062 (73.4%)	0.2825 (67.7%)	0.2466 (59.1%)
Median Average Precision	0.4057	0.2549	0.2765	0.1940
Standard deviation	0.218	0.2191	0.2049	0.1967
Range	0.7777	0.7575	0.8142	0.7921
Minimum	0.0414	0.003	0.008	0.001
Maximum	0.8191	0.7605	0.8227	0.7933
p-value		0.425[1]		<0.001[2]

[1] Result of statistical testing using the Wilcoxon signed ranks test to compare run *mt_tdn* and *lkb_tdn*.
[2] Result of statistical testing using the Wilcoxon signed ranks test to compare run *ldc_tdn* and *lkb_tdn*.

The evaluation results show that the monolingual run achieved much better performance than all three cross-lingual runs. Among the cross-lingual runs, lkb_tdn was significantly better than ldc_tdn: the p-value is less than 0.001 in terms of Wilcoxon signed ranks test. However, there is no significant difference on performance between lkb_tdn and mt_tdn.

Even though lkb-tdn is statistically different from ldc_tdn, their Mean Average Precision values are both lower than 30%, which means there is considerable room for improvement. Since LKB is the resource that we would like to evaluate, it would be important to explore what kinds of topics obtained improvement after using the LKB and why the LKB couldn't bring a larger effect to the system's performance. The following components will try to answer those questions.

Query Categorization

The results of the overall performance evaluation show that the LKB did improve performance as compared with the LDC dictionary. But more analysis needs to be performed before the accuracy of the LKB translation and the impact of the translation resources can be understood. The Query Categorization classified the 54 topics based on run lkb_tdn and its comparison with the other three runs. Table 4 is the results of categorization. The threshold values are 0.17 and 0.5 to cut the queries into easy, moderate and hard categories.

Table 4. Query Categorization Based on lkb_tdn and the Comparing Runs

Category	Classification Criteria	Topics in the Category	Total Topics
Hard & Stable	AV (average precision) score was below 0.17 for all four runs	1, 5, 6, 13, 14, 18, 34, 46	8
Hard & Unstable	AV (average precision) score was below 0.17 for *lkb_tdn*, but one or more other runs got higher scores	7, 9, 17, 25, 26, 28, 30, 32, 33, 39, 41, 42, 45, 48,	14
Moderate & Stable	AV (average precision) score was between 0.17 – 0. 5 for all four runs	2, 3, 8, 10, 12, 16, 27, 29, 35, 36, 37, 43, 44, 50, 51, 54	16
Moderate & Unstable	AV (average precision) score was between 0.17 – 0. 5 for *lkb_tdn*, but one or more other runs got higher or lower scores	4, 11, 15, 19, 24, 31, 52	7
Easy & Stable	AV (average precision) score was above 0.5 (include 0.5) for all four runs	20, 22, 23, 38, 40	5
Easy & Unstable	AV (average precision) score is above 0.5 (include 0.5) *lkb_tdn*, but one or more runs got lower scores	21, 47, 49, 53	4

Translation Analysis

Query translations using the LKB and the LDC dictionary were manually evaluated. A human evaluator was given the translation results and instructed to classify the translation into three categories: correct translation, incorrect translation, and missing translation. The evaluation results are summarized in Table 5. Applying the Wilcoxon signed ranks test demonstrated that the LKB achieved significantly more correct translations and fewer missing translations than the LDC dictionary. The difference between the two resources on incorrect translation is not significant.

Table 5. Table 5. Summary of Query Translation using the LKB and the LDC Dictionary

	lkb_tdn	*ldc_tdn*
Total Terms Evaluated	1538	1610
Number of Correct Translations	1204 (78.3%)	1185 (73.6%)
Number of Incorrect Translations	260 (16.9%)	282 (17.5%)
Number of Missing Translations	74 (4.8%)	143 (8.9%)

Individual Query Analysis

In this component, the researcher was interested in what contributed to the good performance of LKB on certain queries, and what caused the failure of LKB on some other queries. The first factor being considered was the translation effectiveness. The analysis above revealed that EC-CLIR using the LKB achieved better retrieval performance than that using the LDC dictionary, and the translation using the LKB was better as well. Superficially, a correlation between the difference in EC-CLIR performance and the percentage difference in correct, incorrect and missing translations can be expected. However, this was not true for the queries tested in this study. A correlation analysis using Spearman's rho found that the difference in average precision between lkb_tdn and ldc_tdn had no correlation with the difference in the percentage of correct, incorrect and missing translations [1].

The researcher then decided to examine two types of topics: Hard & Stable, and Hard & Unstable topics, to explore the reasons behind the above results and other major factors affecting system performance. The analysis of Hard & Stable topics may discover causes of generally hard topics, and the analysis of Hard & Unstable ones may help find possible ways to improve the performance of the CLIR system using the LKB.

Eight topics belong to Hard & Stable category, which had an average precision score lower than 0.17 from all the four runs. They were topics 1, 5, 6, 13, 14, 18, 34, and 46. These queries were resistant to translation errors--the query translation results had little effect on their IR performance. Among them, Topics 1, 5, 14, 18 proved difficult for TREC-5 monolingual participating systems, with median average precision lower than 0.15. In an attempt to find out the reasons, the top 10 retrieved documents returned by the

most precise of the four runs were examined. Table 6 presents some characteristics of those topics, including the run which returned the highest average precision (AP), the magnitude of the AP, query length, number of relevant documents, and the number of relevant documents returned in the top 10 by the top runs. It appears that there were very few relevant documents returned in the top 10 for most of the Hard & Stable topics.

Table. 6. Hard & Stable Topics

Topic ID	The run returning the highest AP score	The highest AP score	Original English query length (in words)	# of Relevant Documents	# of relevant documents in top 10
1	*mono_tdn*	0.1502	56	13	1
5	*ldc_tdn*	0.1069	70	28	3
6	*mono_tdn*	0.1325	37	77	4
13	*ldc_tdn*	0.0787	36	110	0
14	*mono_tdn*	0.0558	45	57	2
18	*lkb_tdn*	0.1214	93	102	1
34	*ldc_tdn*	0.1632	65	95	5
46	*mono_tdn*	0.1443	68	166	6

A manual inspection of the top 10 retrieved documents (including relevant and irrelevant documents) for each topic has been conducted. Table 7 summarizes our observations after comparing the relevant and irrelevant documents in the top 10 sets. It seems to us that most of the 8 topics need a different retrieval strategy from the traditional tf-idf IR model applied by the system. For example, the query from topic 6, “International support of China’s membership in the WTO,” asks for specific nations which support China’s membership in the WTO. Documents about general issues concerning the WTO and China were judged not relevant.

Fourteen topics belong to the Difficult & Unstable Category, which received lower average precision scores from lkb_tdn but higher score(s) from one or more other runs, as specified in Table 4. For most of the 14 queries, the poor performance was mainly caused by translation errors, such as incorrect translation or lack of translations of the important terms. For examples, lkb_tdn failed to correctly translate "fire" for topic 26, which is among the most important words for this topic. Topics 7, 9, 17, 26, 28, 32, 39, 41, and 48 received the highest average precision scores from mt_tdn among the three cross-lingual runs. The Huajian MT system generated accurate translations for the important terms in these queries, such as "forest fires", "Cellular phones", "terrorism", and "Resettlement." These terms were either missing a translation or were incorrectly translated in lkb_tdn.

Table 7. Relevant and Non-relevant Document Analysis for Hard & Stable Topics

Topic ID	Topic Title	Results
1	U.S. to separate the most-favored-nation status from human rights issue in China.	Relevant documents should discuss the reasons. Most retrieved docs don't satisfy that.
5	Regulations and Enforcement of Intellectual Property Rights in China	Relevant documents should mention specific laws established for Intellectual Property Rights protection. Many retrieved documents don't discuss any specific laws.
6	International Support of China's Membership in the WTO	Relevant documents should include names of specific nation(s) that support China's membership. Non-relevant documents don't contain specific nations.
13	China Bids for 2000 Olympic Games	Relevant documents should describe how China bid for the 2000 Olympic Games. Term "bid" needs to be expanded to include activities or events that were held in China for the Games. Most retrieved documents describe not events related to the 2000 Olympic Games, but Chinese athletes' performance at other Olympic Games. No relevant documents found in top 10.
14	Cases of AIDS in China	Relevant documents should specify areas in China that have the highest AIDS cases, and the approaches to prevent AIDS transmission. Most retrieved documents don't satisfy the above.
18	The Mid-East Peace Talks	Most retrieved documents don't center around the peace talks, but provide background information or descriptions of conflicts in the area.
34	The Impact of Droughts in China	Needs numbers to specify the impact of droughts. Half of the retrieved documents are relevant.
46	New advances in the Relationship between	A topic with no specific requirements. The Top 10 was OK since 6 out of 10 retrieved docs are relevant.

		talks, but provide background information or descriptions of conflicts in the area.
34	The Impact of Droughts in China	Needs numbers to specify the impact of droughts. Half of the retrieved documents are relevant.
46	New advances in the Relationship between China and Vietnam	A topic with no specific requirements. The Top 10 was OK since 6 out of 10 retrieved docs are relevant.

What can be concluded from the above individual query analysis? Firstly, the role of translation to IR performance depends largely on the nature of the queries. For most queries, correct translation of important conceptual terms is essential to system performance for a CLIR system. But some queries are resistant to query translation errors. Occasionally, missing translations or incorrect translations even brought benefits to IR performance. Secondly, queries that are generally difficult normally require alternative solutions for retrieval. This is evidenced by the results of the TREC Robust Track in which some participating systems applied certain natural language processing techniques to difficult topics and achieved better performance on those queries than the top systems (Liu, Sun, & Yu, 2004). Thirdly, current IR systems or search engines that apply a unified strategy to handle different queries are not sufficient. A "bag-of-words" approach may work very well for some queries, but is not effective for some others which require identifying named entities or numbers. But applying sophisticated IR approach to all queries may be computationally expensive and not necessary for simple queries. Many question answering systems applied different strategies for answering different types of questions (Moldovan, et al., 2004; Chen, *et al*., 2004) – some questions can be better answered using knowledge base and linguistic patterns and some others can be better answered using the Web or statistical methods. High performance information retrieval may also need to go beyond a single IR strategy per system to handle different types of user queries.

As to the targeted translation resource LKB, the analysis discovered that the LKB did a better job on translating certain named entities and had the capability to reduce missing translations. However, it still generated a lot of false translations which affected the CLIR performance. Solutions need to be explored to validate translation knowledge in the LKB and to perform better translation disambiguation at the actual query translation stage.

Summary and Future Research

This paper introduced a concept about CLIR retrieval evaluation called Retrieval Outcome Analysis (ROA) and proposed a framework for conducting the analysis. The ROA framework consists of four interrelated steps: Overall System Performance Evaluation, Query Categorization, Translation Analysis, and Individual Query Analysis. The application of the framework discussed in the paper demonstrated that the framework could discover more characteristics of the target CLIR system and could provide more evidence about the performance of the system on a particular set of test topics than TREC-type evaluation. The author believes that the ROA framework could be a useful tool to guide real-world CLIR system design and development. A real-world CLIR system designed for real customers needs to understand the types of queries its customers will have and be able to find relevant information for those queries. Evaluation guided by the ROA framework will be able to inform the designer what types of queries can be effectively handled by the system and what types of questions may be difficult for the system. It can also discover the strengths and weaknesses of the translation resources used by the system. Ultimately, the information resulting from the analysis may help the CLIR system designer to improve the system by focusing on exploring solutions to the weak points in the system.

The ROA framework discussed in this paper is just the beginning step toward a systematic retrieval outcome analysis which has the potential to provide more insights into CLIR system performance. More research is needed to further develop and evaluate the framework. Specifically, directions for further research include:

1. investigating appropriate measures as outcomes of the whole analysis and the integrating of the four components to obtain an overall evaluation of a CLIR system. The framework would be more useful if certain measures of retrieval performance could be calculated or produced out of the analysis.
2. exploring automated or semi-automatic techniques for performing the analysis to facilitate rapid system development.
3. refining the Individual Query Analysis component so that a list of analytic techniques can be recommended for use to assess individual queries.
4. evaluating the framework itself through applying it to CLIR systems in different contexts, such as interactive CLIR systems and multi-lingual IR systems.

Currently the researcher is working on exploring appropriate measures as the outcomes of the analysis and automating the process of Translation Analysis and testing statistical approaches to Individual Query Analysis. The ROA framework is expected to be further developed and evaluated to make it a useful tool for CLIR system evaluation.

Acknowledgements

The author would like to thank the anonymous reviewers for their comments and suggestions.

Notes

[1]The values of Spearman's rho between the difference in average precision between lkb_tdn and ldc_tdn and the difference in the percentage of correct, incorrect, and missing translations are 0.144, -0.082, and -0.018 respectively. None is significant.Back

References

Blair, D. C. (2002). Some thoughts on the reported results of TREC. *Information Processing and management, 38*(4), 445-451.

Chen, J. (2003). *The construction, use, and evaluation of a lexical knowledge base for English-Chinese cross language information retrieval*. PhD dissertation, Syracuse University.

Chen, J., Ge, H., Wu, Y., and Jiang, S. (2004). UNT at TREC 2004: question answering combining multiple evidences. *TREC 2004 Conference Note Book*, p. 695-702.

Conover, W. J. (1999). *Practical Nonparametric Statistics*. John Wiley and Sons, 3rd edition.

Dalrymple, P.W & Roderer, N. K. (1994), Database access systems. In Williams, M. E. (Ed.) *Annual Review of Information Science and Technology, vol. 29*, (pp. 137-178).

Diekema, A. (2003). *Translation events in cross-language information retrieval: lexical ambiguity, lexical holes, vocabulary mismatch, and correct translations*. PhD dissertation, Syracuse University.

Hu, X., Bandhakavi, S., and Zhai, C. (2003). Error analysis of difficult TREC topics. *Proceedings of ACM SIGIR 2003* (poster).

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. *Proceedings of the 16th ACM SIGIR*, p. 329-338.

Liu, S., Sun, C., & Yu, C. (2004). UIC at TREC-2004: Robust Track. *TREC 2004 Conference Note Book*, p. 625 – 634.

Moldovan, D., Harabagiu, S., Clark, C., Bowden, M., Lehmann, J. & Williams, J. (2004). Experiments and Analysis of LCC's two QA Systems Over TREC 2004. *TREC 2004 Conference Note Book*, p. 21- 30.

Papineni, K., Roukos, S., Ward, T., & Zhu (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of 40th ACL Annual Conference*, p.311-318. Available at: http://www.ldc.upenn.edu/acl/P/P02/P02-1040.pdf

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. *Proceedings of the 18th ACM SIGIR*. p. 138-146.

Voorhees, E. M. (2003). Overview of the TREC 2003 Robust Retrieval Track. *Proceedings : The Twelfth Text REtrieval Conference*, P. 69 – 77. Available at: http://trec.nist.gov/pubs/trec12/papers/ROBUST.OVERVIEW.pdf

Voorhees, E. M. (2004). Overview of the TREC 2004 Robust Retrieval Track. *TREC 2004 Conference Note Book*, p. 183-190.