Characteristics of the Korean Test Collection for CLIR in NTCIR-3

Sukhoon Lee*, Sung Hyon Myaeng**, Hyeon Kim***, Jerry H.Seo***, Buil Lee*, Sukhyun Cho**
*Department of Statistics, Chungnam National University
**Department of Computer Science, Chungnam National University
***Dept. of Information Technology, Korea Institute of Science and Technology Information
shelee@stat.cnu.ac.kr, shmyaeng@cs.cnu.ac.kr, hyeon@kisti.re.kr, jerry@kisti.re.kr, buillee@stat.cnu.ac.kr, shcho@enya.cnu.ac.kr

Abstract

This paper describes characteristics of the Korean Test Collection constructed for Cross-Language Information Retrieval (CLIR) task at the third NTCIR workshop. We first give summary statistics and then provide further analyses of the collection using the search results and relevance judgments. More specifically, we report on our preliminary analyses on exhaustivity of relevant documents included in the collection, topic characteristics in terms of difficulty, and various factors that may influence the system rankings.

Keywords: *Test Collection, Relevance Judgement, Topic, System Rankings*

1. Introduction

The Korean Test Collection constructed for the Cross Language Information Retrieval (CLIR) task consists of 66,146 Korean documents, the news articles in the Korean Economic Newspaper published in 1994, 30 topics, and relevance assessments for each topic. The topics are written in four different languages: Chinese, English, Japanese and Korean. Considering the cross-language aspects that should have some bearing on cultural differences, the topics were collected from three countries, Taiwan, Japan, and Korea and from the topic set by CLEF.

The relevance judgments were done in four grades, according to the agreement made by the CLIR executive committee. The topic set was divided into five subsets, each of which was assessed by a judge according to his background. Relevance judgment results contain not only grades but also justifications for the grades assigned to the documents.

The documents used for relevance judgments were collected from the runs submitted by the participants. Eight groups have conducted monolingual Korean retrieval tasks, submitting 17 runs, and two groups participated in cross language retrieval tasks, submitting 6 runs. The total of 46,731 judgments were made by the judges. Although multiple judgments would have been more desirable for consistency and a higher quality of the results, our limited resource didn't make it possible.

After the collection was completed, we have attempted to evaluate the quality of the test collection and describe some characteristics so that more in-depth analysis of the system being evaluated can be done, other than basic statistics such as recall and precision and comparative figures based on the measures. Evaluation of the test collection has been carried out for the following aspects [1]: exhaustivity of the relevance judgments, categorization of topics, and invariability of the collection for comparative system evaluations.

2. Description of the Korean Test Collection

This Section describes some details of the collection in terms of:

(1) the documents
(2) the topics
(3) relevance judgments of documents for the topics

2.1. Documents

The document set in the Korean collection consists of the articles randomly selected from the entire collection of the Korea Economic Newspaper articles published in 1994. The document collection contains 66,146 documents that are written in Korean.

Each document consists of several fields: document ID, header, title, text, and a keyword list.
The list below shows the basic statistics of the document lengths.

- The number of documents: 66,146
- The total length of the collection: 92,976,140 bytes
- The minimum length among the documents: 15 bytes
- The maximum length among the documents: 23,257 bytes
- The mean length of the documents: 1405.62 bytes
- The median length of the documents: 993 bytes
- The standard deviation of lengths: 1,211.8 bytes
- The skewness: 3.7
- The kurtosis: 22.9

2.2. Topics

Topics in a test collection generally represent users' information needs with descriptions in natural language. We categorize users into four classes, Chinese, Japanese, Korean and European, in order to make the topics more appropriate for the evaluation of cross-language information retrieval systems. Each topic generated from the source nation was translated into the other languages to form four sets of topics corresponding the four languages.
More specifically, some IR researchers from each country created some topics based on the events or issues both in Korea and in the world in 1994 as a stratified topic-sample of the whole world-wide users. In order to remove the topics with too few or too many relevant documents, several dry runs have been conducted. The topics expected to have basically more than ten but no more than 50 relevant documents among the top 100 topics have been selected.
The collection contains 30 topics, each of which is written in Chinese, Japanese, English, or Korean. A topic consists of a title, description, narrative, and concept field, all of which are normally defined in the IR field.
Details about topic creation and selection process can be found in an overview paper in the proceedings.
2.3. Relevance Assessments
For each topic a pool was made from the union of the top 1,000 documents from each submitted run. Most of the pools were made with the top 200 documents (i.e. depth 200) but a few with the depth less than 200. The relevance judgments were made by one judge assigned to each topic, who classified each document in the pool into one of the four categories: "highly relevant (S)", "relevant (A)", "partially relevant (B)", and "irrelevant (C)". The basic philosophy of relevance judgment is found in the "Relevance Judgment Manual for CLIR Tasks of NTCIR Workshop 3".

3. Exhaustivity of the Relevance Judgements

As one of primary features of the test collection, we examined the degree to which relevance documents have been identified for each topic by evaluating how many documents were identified at different pool depth by the pooling method. This analysis is in line with previous research [4].
Let $n(P)$ be the number of relevant documents found in a new pool of depth P, which were not included in the pool of depth $P-1$. We computed $n(P)$, P = 1, 2, …, 200 and plotted n against P to figure out how many new relevant documents were found as pool depth was increased and also how fast the rate of new document appearance diminishes.
From the plot for all the relevant documents totalled over the 30 topics we found that the appearance rate of new relevant documents diminished fast up to the depth of from 50 to 60 but it becomes much slower after the depth 100.
Topics were classified into two groups based on their patterns of the curve showing the number of new documents. Topics belonging to one group (type I) have many new relevant documents at the very beginning (up to the depth 50 for most of them) but very few new ones after the depth 100. Those topics belonging to the other group (type II), on the other hand, show that the new relevant documents continue to appear with the same rate from the beginning to the point where we stoped. Unlike the type I group, it wasn't possible to fit the pattern with an exponential-like curve. We characterize the type II topics as the ones with more than 10% of relevant documents appearing after the depth 150.
In order to estimate the total number of relevant documents for the topics of type II, which possibly have a certain amount of relevant documents but unjudged and so regarded as irrelevant, we tried to fit the exponential curves to the data from all the 30 topics first and the data from the topics of type I. Since the plots for the entire topic set and those for the type I topic set strongly showed an exponential trend, we regarded the corresponding difference as the total number of relevant documents for the topics of type II.
In the following, we give more detailed results for two different cases: the rigid case with the scores above 3 being relevant and the relaxed case with the scores above 2 being relevant.

Table 1. Curves fitted for rigid and relaxed cases

1) Rigid Case

The curve fitted: $\ln(n+1) = a + b\ln(P)$

	a	b	R^2	Depth 210	Depth 250	Depth 300
For all	4.378 (0.109)	-0.502 (0.025)	0.68	40	200	400
For type I topics	4.153 (0.159)	-0.727 (0.036)	0.68	0	0	0
For type II topics				40	200	400

2) Relaxed Case

The curve fitted: $\ln(n+1) = a + b\ln(P)$

	a	b	R^2	Depth 210	Depth 250	Depth 300
For all	4.993 (0.079)	-0.497 (0.018)	0.80	90	440	840
For type I topics	4.994 (0.148)	-0.828 (0.033)	0.76	10	50	50
For type II topics				80	390	790

The tables show that the number of new relevant documents of type II topics can be estimated as 400 and 790, respectively, depending on the cut-off criteria, when we make a pool of depth 300. Due to the limitation of extrapolation we did not estimate for pools of depth more than 300.

4. Analysis of Topics

4.1. Domain-based Topic Categorization

In order to see the properties of the topics in the collection from the user's point of view, we categorized them on the basis of the following characteristics: creator's country, major domains (politics, economics, social issues, culture, science, etc.), usages of pronouns for a person name, country name, or organization name, and the relationship between the countries and the issues.

Out of the 30 topics, 11 are about political issues, 12 about economical issues, 5, 3, 2, 1 about social, cultural scientific, environmental issues, respectively. To further investigate the characteristics of the topics based on their difficulties, we conducted the t-test and found out that there was no significant difference in difficulties with respect the issues listed above.

However, there was some difference in terms of the nationality contained in the topics. 10 topics were about Korean issues only among the 24 topics that mention some Korean issues to a varying extent. A t-test showed that the mean of average precision for the 10 topics was significantly lower than that of the 24 topics. On the other hand, the topics containing issues related to the other countries have no significant differences in the mean average precision.

The pronoun effect was investigated to see if the topics with pronouns representing person name, country name, and organization name show lower average precision values than those without pronouns. There was no significant difference between two groups of topics in terms of average precision.

We also investigated on possible differences among the topics created by people in different countries, i.e., Japan, Taiwan, Korea, and Europe. We also found that there was no significant difference in average precision.

4.2. Run-Based Topic Categorization

Based on the submitted runs, we obtained three clusters with the following statistical features. Here we used only the 17 mono-lingual retrieval runs. The result might be a clue for a further study on the analysis of topic difficulties [3]. Details are in Tables 2 and 3.

Table 2. Mean Average Precision of Clusters - Rigid case

Cluster 1 (1, 3, 15)	Cluster 2 (5, 7, 9, 11, 13, 14, 16, 17, 18, 21, 25, 28, 30)	Cluster 3 (2, 4, 6, 8, 10, 12, 19, 20, 22, 23, 24, 26, 27, 29)
0.5684	0.2294	0.0737
0.7487	0.2193	0.0404
0.7433	0.4240	0.1140
0.7525	0.5165	0.2064
0.7609	0.5584	0.1967
0.7189	0.2250	0.0817
0.4042	0.1557	0.0634
0.6463	0.2556	0.0655
0.6992	0.4302	0.1190
0.7172	0.4187	0.0956
0.6100	0.2984	0.0833
0.4251	0.2943	0.0979
0.3272	0.2498	0.0717
0.4156	0.2903	0.0608
0.3181	0.0839	0.0546
0.6559	0.2512	0.1124
0.5389	0.3605	0.1264

Table 3. Mean Average Precision of Clusters – Relaxed case

Cluster 1 (1, 3, 7, 9, 13, 14, 15, 18, 25, 28)	Cluster 2 (2, 4, 5, 8, 11, 16, 17, 24, 30)	Cluster 3 (6, 10, 12, 19, 20, 22, 23, 26, 27, 29)
0.5075	0.1993	0.0845
0.5989	0.2710	0.0985
0.5989	0.4939	0.0963
0.6884	0.5715	0.2280
0.6897	0.5757	0.2170
0.4214	0.2980	0.0676
0.3037	0.1637	0.0684
0.4599	0.1688	0.0796
0.5579	0.3843	0.1084
0.5474	0.3388	0.1030
0.4522	0.2237	0.1137
0.3825	0.4192	0.0708
0.3470	0.2339	0.0613
0.4486	0.2639	0.0687
0.1849	0.1679	0.0181
0.5216	0.2891	0.1079
0.5971	0.2812	0.1786

We investigated the effects of each cluster on the system rankings as in the next section.

5. Effects on System Rankings

5.1. Effect of pooling depths

As discussed in section 3, some relevant documents for several topics were not found in the pool. Thus we needed to investigate the influence of pool depths on the system evaluation. One of the ways to see this phenomenon is to go over the system rankings based on the different pooling depths [2].
The following table shows that the Kendall's tau coefficient between the rankings based on the pools used in this collection and the rankings based on the pools of depths 30, 40, 50, 80, 100.

Table 4. Kendall's tau coefficient

Depth	Rigid Kendall's tau	Relax Kendall's tau
30	0.913	0.953
40	0.929	0.968
50	0.913	0.960
80	0.921	0.976
100	0.921	0.976

From the Kendall's coefficients computed, we can conclude that using this collection for evaluating the systems would cause no problem even if it is very likely that some relevant documents are buried in the unjudged document set.

5.2. Effect of the Cut-off Criterion for Relevance

The Kendall's coefficient between the system rankings based on the "Relax" and those on the "Rigid" was 0.834. which is lower than those in Table 4. Out of 23 runs, 4 runs have two different rankings where the differences are larger than 2. One extreme case shows that a system ranked at 14 in the rigid judgment was ranked at 7 in the relaxed judgment. This means that it is more likely that system rankings can be altered depending on whether 'relaxed' or 'rigid' criterion is used for relevance judgments. We need further investigations on this issue.

5.3. Effect of the topic types

We checked if system ranks were influenced by the types of the topics, i.e. type I and II. In order to do this, we compared the system rankings based on the topics of type I and those based on the topics of type II. The results show that the ranks of the 8 runs were altered by more than 4 depending on which topic types were used.

5.4. Effect of topic difficulties

Based on the run-based topic categorization result, we investigated the effect of the topic difficulties on the run rankings by comparing the rankings based an all the 30 topics, topics in the cluster 1 (the easy ones) and topics in the cluster 3 (the hard ones). The table shows that some runs with a high ranking from the easy topics had a very low ranking from the hard topics. It would be desirable to do further studies on the algorithms of the runs that show such large differences.

Table 5. Effects of topic difficulties on rankings

Rigid Case

Full	13	14	4	2	1	10	16	11	3	5	7	9	15	12	17	8	6
Cluster1	11	2	4	2	1	5	15	9	7	6	10	13	16	14	17	8	12
Cluster3	11	17	5	1	2	10	14	13	6	8	9	7	12	15	16	6	3

Relaxed Case

Full	10	7	3	1	2	12	16	14	5	6	11	9	15	13	17	8	4
Cluster1	9	3	4	2	1	13	16	10	6	7	11	14	15	12	17	8	5
Cluster3	10	8	9	1	2	15	14	11	5	7	4	12	16	13	17	6	3

6. Conclusions

We examined some properties of the Korean Test Collection of CLIR at the NTCIR Workshop 3 from the point of views of the document set, exhaustivity of the relevance judgement, analysis of topics, and factors influencing the run rankings.

We introduced the basic statistics of the document lengths and studied exhustivity of the relevance judgements by estimating the number of relevant documents not identified in the pool set for judgements. We found two groups of topics: the topics of one group (type Ⅰ topics) have most of their relevant documents in the beginning part of the pools and very few in the bottom of the pools, but the topics of the other (type Ⅱ topics) have the relevant ones appearing uniformly in the pool. A preliminary analysis showed that systems using type Ⅱ topics have significantly lower average precision than those using type Ⅰ topics.

Regarding the analysis of topics we tried two kinds of categorization: domain-based, and run-based. Domain-based categorization adopted the features such as fields of topic domains, the, creator's country, countries in which the issues are raised, usage of pronouns. Our investigation to see the influence of these features on the average precision indicated that all features except Korean related issues did not make any significant influence. Run-based categorization classified the topics into 3 clusters: easy. normal, and hard. This categorization might be a clue to analysis of topic difficulties. As factors that possibly influence the run rankings we considered pooling depths, relevant criterion (rigid, relaxed), the topic types defined in sections 3, and the topic difficulties. The pooling depths, and relevant criterion influenced little on the rankings, but some characteristics of the topics did on the rankings of some runs.

Based on these preliminary findings, we feel that a further study would be necessary for the analysis of topic difficulties, analysis of algorithms used in runs.

References

[1] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, Soichiro Hidaka. Overview of IR tasks at the first NTCIR workshop. In *Proc. Of 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 11-44, 1999

[2] David Banks, Paul Over, and N. Zhang. Blind Men and Elephants: Six Approaches to TREC Data, *Information Retrieval* 1, pp7-34, 1999

[3] Koji Eguchi, Kazuko Kuriyama, Noriko Kando. Analysis of the topic difficulty for NTCIR-1 : NACSIS test collection form information retrieval system-1 (in English of Japanese). In IPSJ SIG *Notes*, number 2000-F1-59 (2000-DD-24), pp25-32, 2000

[4] Justin Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments? *Proceedings of the 21st Annual International ACM SIGIR Conference*, pp307-314, 1998.