Overview of CLIR Task at the Third NTCIR Workshop

Kuang-hua Chen[*], Hsin-Hsi Chen[+], Noriko Kando[‡], Kazuko Kuriyama[**]
Sukhoon Lee[++], Sung Hyon Myaeng[‡‡], Kazuaki Kishida[+++], Koji Eguchi[‡], and Hyeon Kim[***]

[*]Department of Library and Information Science, National Taiwan University
Taipei 10617, Taiwan
khchen@ccms.ntu.edu.tw

[+]Department of Computer Science and Information Engineering, National Taiwan University
Taipei 10617, Taiwan
hh_chen@csie.ntu.edu.tw

[‡]National Institute of Informatics (NII)
Tokyo 101-8430, Japan
{kando, eguchi}@nii.ac.jp

[**]School of Literature, Shirayuri College
1-25 Midorigaoka, Chofu, Tokyo 182-8525, Japan
kuriyama@shirayuri.ac.jp

[++]Department of Statistics, Chungnam National University
Taejon 305-764, Korean
shlee@stat.cnu.ac.kr

[‡‡]Department of Computer Science, Chungnam National University
Taejon 305-764, Korean
shmyaeng@cs.cnu.ac.kr

[+++]Faculty of Cultural Information Resources, Surugadai University
Hanno-shi 357, Japan
kishida@surugadai.ac.jp

[***]Department of Information Technology, Korea Institute of Science and Technology Information
P.O.BOX 122, Yusong, Taejon, Korea
hyeon@kisti.re.kr

Abstract

This report is an overview of Cross-Language Information Retrieval Task (CLIR) at the third NTCIR Workshop. There are 3 tracks in CLIR: Single Language IR (SLIR), Bilingual CLIR (BLIR), and Multilingual CLIR (MLIR). The scope, schedule, test collections, search results, relevance judgment, scoring results, and the preliminary analyses are described in the report.

1. Introduction

After the second NTCIR workshop [1, 2], researchers from Japan, Korea, and Taiwan have discussed a much more complicated cross-language information retrieval (CLIR) evaluation task, which is closer to the realistic application of IR environment and is a real challenge to IR researchers. We propose a CLIR task in the third NTCIR workshop and organize an executive committee to fulfill this task. The CLIR Task Executive Committee consists of 9 researchers from Japan, Korea, and Taiwan. These members meet 3 times in Japan to discuss the details of CLIR Task, to make the schedule, and to arrange the agenda.

The goals of CLIR Task at NTCIR are shown as follows.

- Promote CLIR researches
- Initiate complicated and real CLIR task
- Investigate effective techniques for CLIR
- Provide a forum to present research results and share research ideas

There are 3 tracks in CLIR task: 1) Multilingual CLIR (MLIR), 2) Bilingual CLIR (BLIR), and 3) Single Language IR (SLIR). Basically, 4 languages will involve in the CLIR task: Chinese, English, Japanese, and Korean. The test collection used in CLIR Task is composed of news articles in the aforementioned 4 languages. Chinese, English, and Japanese documents are news articles published in 1998 and 1999. In contrast, Korean documents are news articles published in 1994. Therefore, Chinese, English, and Japanese documents could be regarded as one set; Korean documents will be regarded as another set. That is to say, we will have retrieval task of Chinese queries against Chinese and Japanese documents, but we could not provide retrieval task of Chinese queries against Chinese and Korean documents. The possible retrieval tasks will be described in Section 2.

Researchers from all over the world are welcome to participate CLIR Task. The participants could use automatic systems or interactive systems are

welcome. The participants could submit any retrieval task listed in Section 2. However, we have some restrictions on the number and type of submitted runs. Please refer to Section 2 for detailed information.

26 groups from 8 countries or areas, 19 groups from 6 countries or areas, and 15 groups from 6 countries or areas enrolled SLIR track, BLIR track, and MLIR track, respectively. However, not all enrolled groups submit search results. Table 1 shows the distribution of groups enrolling each track and groups submitting search results at final. In total, 189 runs are submitted from 23 groups. Among them, 110 runs are SLIR submitted from 22 groups, 50 runs are BLIR submitted from 14 groups, and 29 runs are MLIR submitted from 7 groups. Table 2 shows the detailed statistics.

Table 1. Distribution of Participants

	Enrolled			Submitted		
	SLIR	BLIR	MLIR	SLIR	BLIR	MLIR
Canada	3	1	3	2	0	0
China	3	4	1	2	3	1
Hong Kong	2	1	0	1	0	0
Japan	7	6	5	6	4	2
Korea	2	1	1	3	2	0
Singapore	1	0	0	1	0	0
Taiwan	1	0	1	2	0	1
USA	7	6	4	5	5	3
Total	26	19	15	22	14	7

Table 2. Participants for CLIR Task

Tracks	Run Types	#Runs Submitted	#Grps Submitted	Total. #Runs Submitted	Total #Grps Submitted
SLIR	C-C	34	15	110	22
	E-E	29	15		
	J-J	30	13		
	K-K	17	8		
BLIR	C-E	3	1	50	14
	C-J	4	2		
	E-C	16	6		
	E-J	11	5		
	E-K	6	2		
	J-C	5	3		
	J-E	1	1		
	K-C	2	1		
	K-E	2	1		
MLIR	C-CE	3	1	29	7
	C-CJ	3	1		
	C-JE	3	1		
	C-CJE	4	2		
	E-CE	6	2		
	E-CJE	4	2		
	E-JE	1	1		
	J-JE	2	2		
	J-CJE	3	1		
Total		189	23	189	23

The rest of this report will focus on the test collections, relevance judgment, and search results. Section 2 will introduce the CLIR task at the third NTCIR workshop. Section 3 will describe the test collections used in CLIR task. Section 4 will give a picture of the evaluation mechanism. Section 5 will analyze the search results in a broad view. Section 6 will give a conclusion.

2. CLIR Task

3 tracks are provided in the CLIR task. All tracks are ad-hoc, that is to say, the document set is fixed against the different topics.

2.1 Schedule

The CFP is distributed on Aug. 2001. The following shows the schedule for CLIR task at the third NTCIR workshop.

2001-09-30	Application Due
2001-10-30	Deliver Dry Run data
2001-11-15	Submit search result of Dry Run
2001-11-30	Deliver evaluation result of Dry Run
2001-12-22	Deliver Formal Run data
2002-01-25	Submit search result of Formal Run
2002-07-01	Deliver the evaluation results
2002-08-20	Paper Due
2002-10-08/ 2002-10-10	NTCIR Workshop 3

2.2 Tracks

Participants could choose to take part in any one, any two, or all of three tracks.

- Multilingual CLIR (MLIR)

The document set of MLIR Track consist of more than two languages. The challenge is that participants have to resolve the complicated multi-language issues. Since the publishing dates of Korean documents are not parallel to those of Chinese, English, and Japanese documents (see Section Test Collection), document set used in MLIR consists of Chinese, English, and Japanese. However, the topic could be Chinese, English, Japanese, and Korean. We will prepare topics in 4 languages, so participants could carry out K→CEJ track. The following shows the possible retrieval tasks in MLIR track.

Table 3. MLIR Track

Topic Set		Doc. Set	Topic Set		Doc. Set
C	→	C, J, E	C	→	J, E
E	→	C, J, E	E	→	J, E
J	→	C, J, E	J	→	J, E
K	→	C, J, E	K	→	J, E
C	→	C, J	C	→	C, E
E	→	C, J	E	→	C, E
J	→	C, J	J	→	C, E
K	→	C, J	K	→	C, E

- Bilingual CLIR (BLIR)

The topic set and document set in BLIR Track are in different languages. The complexity of this track is

less than that of MLIR. However, participants also have to resolve cross-language issues. The following shows the possible retrieval tasks in the BLIR track.

Table 4. BLIR Track

Topic Set		Document Set
C	→	J
E	→	J
K	→	J
C	→	K
E	→	K
J	→	K
J	→	C
K	→	C
E	→	C

- Single Language IR (SLIR)

The topic set and document set of SLIR Track are in the same language. Table 5 shows the SLIR track.

Table 5. SLIR Track

Topic Set		Document Set
C	→	C
J	→	J
K	→	K
E	→	E

2.3 Topic fields and Run Types

Basically, we allow all types of runs using any combination of fields in topic. (Please refer to the section of Test Collection) The participants have to use the 'T' (TITLE), 'D' (DESC), 'N' (NARR), 'C' (CONC) and any combination of these symbols to identify the run types. That is to say, participants can submit T run, D run, N run, C run, TD run, TN run, TC run, DN run, DC run, NC run, TDN run, TDC run, TNC run, DNC run, and TDNC run. Each participant can submit up to 3 runs for each language pair regardless of the type of run. Here, the language pair means the combination of topic language and document language(s). Among these run types, D run is mandatory. Each participant has to submit at least a D run for a language pair. Each run has to be associated with a RunID. RunID is an identity for each run. The format for RunID is as follows.

Group's ID-Topic Lang-Doc Lang-Run Type-pp

The "pp" is two digits used to represent the priority of the run. It will be used as a parameter for pooling. The participants have to decide the priority for each submitted run in the basis of each language pair. "01" means the high priority. The participants could use any topic field to carry out the CLIR tasks provided that D run is mandatory.

3. The Test Collection

The test collection used in CLIR task is composed of document set and topic set. The following will give a brief description of each set.

3.1 Document Set

The documents used in CLIR are news articles collected from different news agencies of different countries. Table 6 shows the related information of documents.

Table 6. Composition of Document Set

Japan	Mainichi Newspaper (1998-1999): Japanese	220,078*
	Mainichi Daily (1998-1999): English	12,723
Korea	Korea Economic Daily (1994): Korean	66,146
Taiwan	CIRB011 (1998-1999): Chinese	132,172*
	United Daily News (CIRB020, 1998-1999): Chinese	249,203*
	Taiwan News and China Times English News (EIRB010, 1998-1999): English	10,204

* The number of Mainichi Newspaper news articles shown in CFP is 236,664, that of CIRB011 is 132,173 and that of CIRB020 is 249,508. However, we find some abnormal news articles. Table 6 shows the final statistics. We have also justified this problem while carrying out relevance judgment and performance scoring. The list of abnormal documents could be downloaded from NTCIR ftp site.

The participants have to sign different contracts for using these materials. Each contract has its own requirements. We hope participants could understand the complicated situations of copyright issues in different countries.

The period of permitted use of the Mainichi Newspapers and Mainichi Daily (Document collections from Japan) are from 2001-09-01 to 2003-09-30. For active participants who will submit the results and who affiliated at the organization outside Japan will be able to extend the period up to 2008-09-30. After the permitted period will be terminated, the participants will have to delete all the documents, or will have to purchase the data from Mainichi Newspaper Co., and obtain the permission for research purpose use from the Company.

The participants have to sign a contract with Dr. Kuang-hua Chen for the use of United Daily News (CIRB020). However, the udn.com (the company of United Daily News) reserves the right to reject the contract. Basically, udn.com will prove the contracts in normal situations. The participants also have to sign a contract with Dr. Kuang-hua Chen for the use of CIRB011 (The minor updated version of CIRB010) and EIRB010 (Taiwan News and China Times English News). The participants have to sign contract with Prof. Sung Hyon Myaeng and will be proved the right to use Korea Economic Daily for 2 years plus possible extensions.

The format of each news article is consistent by using a set of tags. These tags are defined by CLIR Executive Committee based on the convention of IR evaluation. The tag set is shown in Table 7 and a sample document is shown in Figure 1.

Table 7. Document Tags used in CLIR Task

Mandatory tags		
<DOC>	</DOC>	The tag for each document
<DOCNO>	</DOCNO>	Document identifier
<LANG>	</LANG>	Language code: CH, EN, JA, KR
<HEADLINE>	</HEADLINE>	Title of this news article
<DATE>	</DATE>	Issue date
<TEXT>	</TEXT>	Text of news article
Optional tags		
<P>	</P>	Paragraph marker
<SECTION>	</SECTION>	Section identifier in original newspapers
<AE>	</AE>	Contain figures or not
<WORDS>	</WORDS>	Number of words in 2 bytes (for Mainichi Newspaper)

```
<DOC>
<DOCNO>ctg_xxx_19990110_0001</DOCNO>
<LANG>EN</LANG>
<HEADLINE> Asia Urged to Move Faster in Shoring Up Shaky Banks </HEADLINE>
<DATE>1999-01-10</DATE>
<TEXT>
<P>HONG KONG, Jan 10 (AFP) - Bank for International Settlements (BIS) general manager Andrew Crockett has urged Asian economies to move faster in reforming their shaky banking sectors, reports said Sunday. Speaking ahead of Monday's meeting at the BIS office here of international central bankers including US Federal Reserve chairman Alan Greenspan, Crockett said he was encouraged by regional banking reforms but "there is still some way to go." Asian banks shake off their burden of bad debt if they were to be able to finance recovery in the crisis-hit region, he said according to the Sunday Morning Post. Crockett added that more stable currency exchange rates and lower interest rates had paved the way for recovery. "Therefore I believe in the financial area, the crisis has in a sense been contained and that now it is possible to look forward to real economic recovery," he was quoted as saying by the Sunday Hong Kong Standard.
</P>
<P>"It would not surprise me, given the interest I know certain governors have, if the subject of hedge funds was discussed during the meeting," Crockett said.
</P>
<P>He reiterated comments by BIS officials here that the central bankers would stay tight-lipped about their meeting, the first to be held at the Hong Kong office of the Swiss-based institution since it opened last July.
</P>
</TEXT>
</DOC>
```

Figure 1. Sample Document

3.2 Topic Set

The topics used in NTCIR Workshop 3 are created by CLEF [3], Japan, Korea, Taiwan, and TREC [4]. As a result, the topics are much more internationalized. All topics have 4 language versions, i.e., Chinese, English, Japanese, and Korean. We use the tag <SLANG> to denote the contributor or the languages used to create the topic. For example, <SLANG>CLEF</SLANG> means this topic is contributed by CLEF; <SLANG>CH</SLANG> means this topic is created by Taiwan; <SLANG>EN</SLANG> created by TREC; <SLANG>JA</SLANG> created by Japan; <SLANG>KR</SLANG> created by Korea. <TLANG> tag denotes the language of topic as you look at it. The following shows a sample topic. It is created using Chinese by Taiwan and translated into English. The details of tags for topic could be referred to Table 8.

```
<TOPIC>
<NUM>013</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>NBA labor dispute</TITLE>
<DESC>
To retrieve the labor dispute between the two parties of the US National Basketball Association at the end of 1998 and the agreement that they reached.
</DESC>
<NARR>
The content of the related documents should include the causes of NBA labor dispute, the relations between the players and the management, main controversial issues of both sides, compromises after negotiation and content of the new agreement, etc. The document will be regarded as irrelevant if it only touched upon the influences of closing the court on each game of the season.
</NARR>
<CONC>
NBA (National Basketball Association), union, team, league, labor dispute, league and union, negotiation, to sign an agreement, salary, lockout, Stern, Bird Regulation.
</CONC>
</TOPIC>
```

Figure 2. Sample Topic

As previous section describes, we have 1998-1999 Chinese, English, and Japanese news articles and 1994 Korean news articles. We have to prepare topics for the two different time-window documents. As a result, we created hundreds of candidate topics and carried out a pre-test search to select the topics for formal run. Considering the number of documents, we decided to select 30 topics and 50 topics for 1994 documents and 1998-1999 documents, respectively. Table 9 shows the distribution of final topics.

Table 8. Topic Tags used in CLIR Task

<TOPIC>	</TOPIC>	The tag for each topic
<NUM>	</NUM>	Topic identifier
<SLANG>	</SLANG>	Source language code: CH, EN, JA, KR
<TLANG>	</TLANG>	Target language code: CH, EN, JA, KR
<TITLE>	</TITLE>	The concise representation of information request, which is composed of noun or noun phrase.
<DESC>	</DESC>	A short description of the topic. The brief description of information need, which is composed of one or two sentences.
<NARR>	</NARR>	A much longer description of topic. The <NARR> has to be detailed, like the further interpretation to the request and proper nouns, the list of relevant or irrelevant items, the specific requirements or limitations of relevant documents, and so on.
<CONC>	</CONC>	The keywords relevant to whole topic.

Table 9. Distribution of Topic Set of CLIR Task

Time Window	1994	1998-1999
CLEF	4	0
Japan	5	15
Korea	13	12
Taiwan	8	13
TREC	0	10
TOTAL	30	50

4. Evaluation

This workshop is the third NTCIR Workshop. In fact, we have experience in relevance judgment and performance evaluation. However, the CLIR task is much more complicated than ever before we have met. The relevance judgment is executed in Japan, Korea, and Taiwan, separately. Taiwan is responsible of compiling the final relevance judgment and executing the performance evaluation. The TREC_EVAL program is used to score the research results submitted by participants. It provides the interpolated recall and precision at 11 points, AvgPre (non-interpolated) over all relevant documents and precision at 5, 10, 15, 20, 30, 100, 200, 500, and 1000 documents. Each participating group has to submit its search results in the designated format. The result file is a list of tuples in the following form:

qid iter docid rank sim runid

giving *docid* (a string extracted from the <DOCNO> </DOCNO> field in the documents, e.g. <DOCNO> ctg_xxx_19990110_0001</DOCNO>) retrieved by topic *qid* (an integer extracted from <NUM> </NUM> field of topic, e.g., <NUM>002</NUM>, the *qid* is 002) with similarity *sim* (a float). The result file is assumed to be sorted numerically by *qid*. *Sim* is assumed to be higher for the documents to be retrieved first. The *iter* and *rank* could be regarded as the dummy filed in tuples. In addition, each field in tuples is separated by inserting 'TAB' (\x0A, \t) character.

The relevance judgment is undertaken by pooling method. The steps of pooling are as follows,

(1) Collect the top X documents for each topic from all of the corresponding runs to form the pools. One document pool was made for 1998-1999 document sets, and another document pool was made for 1994 document set.
(2) Separate the pooled documents into each language set, that is, the pool for 1998-1999 were separated into four sets which corresponded to four document set, Chinese, Japanese, Taiwan-English and Japan-English set[1].
(3) Check whether the number of pooled documents for each language was appropriate for relevance judgment. If not, increase or decrease the number "X" and go to (1)

The number "X" determines the size of pool for each search topic. As for the formal run, the X was adjusted from 80 to 100 for 1998-99 documents, and from 180 to 200 for 1994 documents so that the total number of documents for each topic might be around 2000. The Xs are shown in Table 10 and Table 11.

The reason why we limited the Xs is that the pooled documents must be judged by human assessors within a given period. According to our experiences of the 1st and 2nd NTCIR Workshop, an appropriate number of documents in a language set for a topic is around 2000. Though the Xs were different among topics, but the same number X of documents were pooled from each run.

As for measurement granularity, it is supposed that some distinct definitions of relevance degree should be identified to keep judgment objective. 4 categories of relevance are identified: "Highly Relevant", "Relevant", "Partially relevant", and "Irrelevant." Each kind of relevance is assigned a relevance score. "Highly relevant" is 3, "Relevant" is 2, "Partially relevant" is 1, and "Irrelevant" is 0.

Since TREC_EVAL scoring program adopts binary relevance, we have to decide the thresholds for 4 categories of relevance. The first is the so-called "Rigid Relevance", that is to say, "Highly Relevant" and "Relevant" are regarded as relevant.

[1] For participants, there is only one English document set. For assessors, we have Taiwan-English and Japan-English document set.

The second is the so-called “Relaxed Relevant”, that is to say, “Highly Relevant”, “Relevant” and “Partially Relevant” are regarded as relevant. Therefore, we will create two scoring results for each submitted run. For example, a run LIPS-E-EJ-D-01 will be evaluated to produce two scoring results using TREC_EVAL: one is LIPS-E-EJ-D-01-Relax; the other is LIPS-E-EJ-D-01-Rigid.

Table 10. Pool Size of each Topic for CJE

Topic	Pool Size	Total	JA	JP-EN	TW-EN	CH
001	100	4695	2185	794	246	1470
002	90	5105	2289	320	624	1872
003	100	5563	2269	801	505	1988
004	80	4655	2096	519	624	1416
005	100	5755	2114	744	902	1995
006	100	3602	1205	673	358	1366
007	100	4969	1759	539	522	2149
008	100	4844	1907	888	363	1686
009	100	4248	1677	561	424	1586
010	100	3935	1313	651	505	1466
011	100	4273	1765	872	357	1279
012	100	4491	1588	598	385	1920
013	80	5569	1774	515	1093	2187
014	100	4676	2015	529	473	1659
015	100	3703	1108	928	329	1338
016	100	3983	1355	678	371	1579
017	100	5059	1840	978	269	1972
018	90	5243	1988	833	325	2097
019	100	4101	1256	636	553	1656
020	100	3861	1135	735	274	1717
021	100	4337	1248	651	442	1996
022	100	4680	1570	958	417	1735
023	100	4696	1894	647	455	1700
024	100	5182	1797	1036	384	1965
025	100	5214	1273	943	546	2452
026	100	5520	1694	943	688	2195
027	100	3516	1107	490	654	1265
028	100	4669	1797	748	235	1889
029	100	4605	1228	1302	389	1686
030	100	4361	1623	845	198	1695
031	90	5168	2122	1211	154	1681
032	80	4796	1762	489	425	2120
033	100	5533	2043	625	684	2181
034	100	5508	2131	643	644	2090
035	100	5097	1798	896	283	2120
036	100	4932	2029	625	340	1938
037	100	4754	1258	874	534	2088
038	100	3301	1017	712	431	1141
039	100	5453	1948	952	260	2293
040	90	5188	1988	733	385	2082
041	80	5672	2069	712	616	2275
042	90	5071	2024	547	545	1955
043	90	4905	1718	609	449	2129
044	100	5279	2249	704	407	1919
045	100	5199	1965	834	371	2029
046	100	4928	2020	861	368	1679
047	80	5585	2065	611	689	2220
048	100	5508	2044	952	374	2138
049	80	5237	1966	659	481	2131
050	90	4944	1755	793	276	2120

After executing relevance judgment, we find the number of relevant documents for some topics is small or is zero. For example, the topics created by Japanese member have few relevant Chinese documents. As a result, the members of Executive Committee of CLIR Task have discussed how to screen out the unsuitable topics for each combination of target languages based on a basic idea of keeping as many topics as possible. Here, the source language is query language and the target language(s) is/are document language(s). We adopt the so-called “3-in-S+A” criterion. Please remember that we have 4-level relevance as shown in Table 12. The “3-in-S+A” means that a qualified topic must have at least 3 relevant documents with ‘S’ or ‘A’ score. Based on this criterion, we identify various topic sets for each combination of target language document set no matter what source (query) languages are used. We construct a **NTCIR-3 Formal Test Collection** for the Evaluation of NTCIR-3 CLIR Task which we describe in the following.

Table 11. Pool Size of each Topic for Korean

Topic	Pool Size	KR
001	200	1996
002	190	1932
003	190	1917
004	200	1604
005	200	1742
006	180	1987
007	200	1419
008	200	1469
009	200	1213
010	200	1276
011	200	1698
012	200	1113
013	180	2028
014	200	1394
015	200	1730
016	200	1776
017	200	1072
018	200	1293
019	200	1369
020	200	1338
021	200	1122
022	200	1744
023	200	1610
024	200	1558
025	200	1616
026	180	1931
027	200	1876
028	200	956
029	200	1705
030	200	1247

Table 12. 4-Level Relevance

Meaning	Symbolic Score	Numerical Score
Highly Relevant	S	3
Relevant	A	2
Partially Relevant	B	1
Irrelevant	C	0

1. For Chinese, Japanese, and English Document Set (note that these are 1998-1999 news articles) and the accompanying 1998-1999 Topic Set with 50 topics as we send to you in the beginning of CLIR task, we create the following sub-test collection based on "3-in-S+A" criterion. In the FORMAL Test Collection, each target document set (C, J, E, CJ, CE, JE, CJE, and K) has different set of topics.
   (1) **NTCIR-3 Formal Chinese Test Collection**
   It contains 381,681 Chinese documents and 42 topics in source language of Chinese, Japanese, Korean, and English. The IDs of topics in 1998-1999 Topic Set used in this collection are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 45, 46, 47, 48, 49, and 50.
   (2) **NTCIR-3 Formal Japanese Test Collection**
   It contains 220,078 Japanese documents and 42 topics in source language of Chinese, Japanese, Korean, and English. The IDs of topics in 1998-1999 Topic Set used in this collection are 2, 4, 5, 7, 8, 10, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, and 50.
   (3) **NTCIR-3 Formal English Test Collection**
   It contains 22,927 English documents and 32 topics in source language of Chinese, Japanese, Korean, and English. The IDs of topics in 1998-1999 Topic Set used in this collection are 2, 4, 5, 7, 9, 12, 13, 14, 18, 19, 20, 21, 23, 24, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 42, 43, 45, 46, and 50.
   (4) **NTCIR-3 Formal CJ Test Collection**
   It contains 601,759 Chinese and Japanese documents and 50 topics in source language of Chinese, Japanese, Korean, and English. All of the 50 topics in 1998-1999 Topic Set are used in this collection.
   (5) **NTCIR-3 Formal CE Test Collection**
   It contains 404,608 Chinese and English documents and 46 topics in source language of Chinese, Japanese, Korean, and English. The IDs of topics in 1998-1999 Topic Set used in this collection are 1 ,2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32 ,33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 45, 46, 47, 48, 49, and 50.
   (6) **NTCIR-3 Formal JE Test Collection**
   It contains 243,005 Japanese and English documents and 45 topics in source languages of Chinese, Japanese, Korean, and English. The topics in 1998-1999 Topic Set used in this collection are 2, 4, 5, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, and 50.
   (7) **NTCIR 3 Formal CJE Test Collection**
   It contains 624,686 Chinese, Japanese, and English documents and 50 topics in source language of Chinese, Japanese, Korean, and English. All of the 50 topics in 1998-1999 Topic Set are used in this collection
2. For Korean Document Set (note that these documents are 1994 news articles) and the accompanying 1994 Topic Set with 30 topics as we send to you in the beginning of CLIR task, we create one sub-test collection also based on "3-in-S+A" criterion.
   (8) **NTCIR-3 Formal Korean Test Collection**
   It contains 66,146 Korean documents and 30 topics in Chinese, Japanese, Korean, and English. All of the 30 topics in 1994 Topic Set are used in this collection.

To sum up, we will have a **NTCIR-3 Formal Test Collection** with 8 sub-test collections for all of the possible combination of target language documents except that Korean documents could not be combined with Chinese, Japanese, and English documents.

5. Search Results

We will report the search results in a broad view and analyze some of runs in this section. Before going to the different tracks, we would like to describe the 3 interactive runs among 189 submitted runs. The first interactive run is an E-E run, the second is an E-K run, and the third is a J-C run. They are the top run in the corresponding query-document pair runs except the E-E run. Another interesting point is 12 groups apply probabilistic model and 11 groups apply vector-space model. More groups use probabilistic model than ever before and many top runs are fulfilled by various probabilistic models. The aforementioned 3 interactive runs are also fulfilled by probabilistic model. In addition, most groups apply inverted file, but CRL and SSTUT use suffix array. The TF/IDF-based weighting is used in most of submitted runs. The technique descriptions could be referred to the Appendix, but please examine the corresponding papers in workshop proceedings for detailed information. Since the <DESC> runs are mandatory runs, we will examine them first. SLIR track, BLIR track and MLIR track will go after <DESC> runs.

5.1 <DESC> Runs

Generally speaking, the words in <DESC> are much more similar to the queries users propose. As a result, using the <DESC> field to construct queries is designed to be mandatory for submitted runs of each query-document pair. However, not every group submitted D runs. Table 13 shows the distribution of D runs.

Table 13. Distribution of D Runs

Run Types	#DRuns	#Grps submit D Run	#Runs	#Grps
C-C	14	13	34	15
E-E	13	10	29	15
J-J	18	13	30	13
K-K	9	8	17	8
C-E	1	1	3	1
C-J	2	2	4	2
E-C	6	5	16	6
E-J	7	5	11	5
E-K	4	2	6	2
J-C	3	3	5	3
J-E	1	1	1	1
K-C	0	0	2	1
K-E	0	0	2	1
C-CE	1	1	3	1
C-CJ	1	1	3	1
C-JE	1	1	3	1
C-CJE	2	2	4	2
E-CE	4	2	6	2
E-CJE	2	2	4	2
E-JE	1	1	1	1
J-JE	2	2	2	2
J-JCE	3	1	3	1

Since the population of some D runs is not enough, we would like to examine D runs of C-C, E-E, J-J, K-K, E-C, E-J, E-K, and E-CE retrieval tasks only. Table 14 shows the relaxed AvgPre of respective D runs. Figure 3 to Figure 10 show the 11-point precision curves for C-C runs, E-E runs, J-J runs, E-C runs, E-J runs, E-K runs, and E-CE runs, respectively.

For C-C D runs, APL group uses statistical language model without morphological operations and query expansion. Brkly uses logistic regression model (probabilistic model in some sense) with unigram+bigram character indexing, removing stop words, query expansion and term re-weighting. CRL uses *okapi* model (probabilistic model) and considers term position with all the patterns of strings, morphology, and automatic query expansion. Pircs group uses probabilistic model, unigram+bigram character indexing, stop words list, and dictionary without query expansion.

For E-E runs, TSB submits 2 D runs and performs well. The different between the two runs is query expansion. The first run uses the offer-weight query expansion, but the second one based on chi-square. However, the difference in performance is not significant. Brkly and APL use same techniques as in C-C runs. OKSAT-E-E-D-01 run is an interactive run using probabilistic model. It uses n-gram query indexing and word+phrase in document indexing. It adds synonym into query as query expansion. HUM uses vector space model with query expansion and case normalization for Latin characters.

For J-J runs, CRL, Brkly, TSB, APL, OKSAT, and HUM apply the same techniques as the aforementioned provided that the query indexing and document indexing are almost character-based (bigram or n-gram). IFLAB uses probabilistic model and word-based system without query expansion. Tlrrd group uses inference network model, bi-character query and document indexing without expansion.

Table 14. Relaxed AvgPre of Top D Runs

Run Types	Run ID	AvgPre
C-C	pircs-C-C-D-001	0.3617
	pircs-C-C-D-002	0.3576
	Brkly-C-C-D-01	0.3517
	CRL-C-C-D-02	0.3448
	APL-C-C-D-02	0.3398
E-E	TSB-E-E-D-01	0.4649
	TSB-E-E-D-02	0.4619
	Brkly-E-E-D-03	0.4368
	APL-E-E-D-02	0.3894
	^OKSAT-E-E-D-01	0.3563
	HUM-E-E-D-01	0.3381
J-J	CRL-J-J-D-03	0.3998
	Brkly-J-J-D-01	0.3946
	TSB-J-J-D-03	0.3910
	TSB-J-J-D-02	0.3904
	TSB-J-J-D-01	0.3903
	APL-J-J-D-02	0.3569
	OKSAT-J-J-D-01	0.3508
	IFLAB-J-J-D-01	0.3427
	HUM-J-J-D-01	0.3272
	tlrrd-J-J-D-01	0.3115
K-K	CRL-K-K-D-03	0.3602
	Brkly-K-K-D-01	0.3131
	APL-K-K-D-02	0.2716
	KUNLP-K-K-D-01	0.2693
	POSTECH-K-K-D-02	0.2665
	HUM-K-K-D-01	0.2433
E-C	MSRA-E-C-D-03	0.1921
	Brkly-E-C-D-01	0.1609
	pircs-E-C-D-002	0.1587
	pircs-E-C-D-001	0.1334
E-J	TSB-E-J-D-02	0.3404
	TSB-E-J-D-01	0.3394
	TSB-E-J-D-03	0.3378
	Brkly-E-J-D-01	0.2207
E-K	^KUNLP-E-K-D-03	0.1987
	KUNLP-E-K-D-02	0.1795
	KUNLP-E-K-D-01	0.1422
E-EC	ISCAS-E-CE-D-03	0.1651
	pircs-E-EC-D-001	0.1620
	pircs-E-EC-D-002	0.1577
	pircs-E-EC-D-003	0.1320

^ denotes the interactive run.

For K-K runs, KUNLP group uses word-based system and probabilistic model without query expansion. Other groups are character-based systems.

POSTECH uses the *okapi* model without query expansion. Query expansion is always a good method to improve retrieval performance.

For E-C runs, both Brkly and pircs groups use MT-based method to translate query. MSRA uses dictionary and co-occurrence information to carry out query translation. All three groups are probabilistic models and have word and character indexing in query and document.

For E-J runs, the basic techniques of TSB, Brkly, and APL will be described in the following sections. TSB uses MT-based method to translate query as Brkly and APL do. They all carry out query expansion.

E-K and E-EC runs also will be described in Subsection 5.3 BLIR and Subsection 5.4 MLIR. We will not repeat these techniques information here.

5.2 SLIR Track

Totally, 110 runs are submitted from 22 groups in SLIR track. Table 15 shows the SLIR participants and Table 16 shows the number of submitted runs. In general, the submitted C-C runs, J-J runs, and K-K runs are character-based and E-E runs are word-based. It should be the direct results of recent IR evaluation. The researchers have learned many experiences from the activities of TREC, CLEF and NTCIR. In the following, we will examine each query-document language pair.

5.2.1 C-C Runs

In total, 8 different combinations of topic fields are used in Chinese-Chinese SLIR track. Table 17 shows the distribution and the corresponding MEAN, MAX, and MIN of AvgPre.

Generally speaking, the TDNC runs show good performance than other runs. The highest precision run of C-C runs is a TNDC run, but the lowest is also a TNDC. Figure 11 and Figure 12 show the relaxed top C-C runs and rigid top C-C runs, respectively. We have to investigate the used techniques. In relaxed top C-C runs, only 2 runs use vector-space model. 8 runs use various probabilistic models. Rigid top C-C runs have the same phenomena. The top three runs are TNDC runs. In addition, the most of top runs adopt query expansion mechanism.

5.2.2 E-E Runs

As the C-C runs, 8 different combinations of topic fields appear in E-E runs. Table 18 shows the distribution and the corresponding MEAN, MAX, and MIN of AvgPre. From viewpoint of topic fields, using all topic fields performs well than other combinations of topic fields. The index unit and query unit of all submitted E-E runs are word-based. Since E-E IR has been the focus of IR researches for a long time, the performance of E-E runs are much better than other query-document language pairs. Figure 13 and Figure14 show the relaxed top E-E runs and rigid top E-E runs, respectively.

Table 15. List of SLIR Participants

1	Applied Physics Laboratory Johns Hopkins University, USA
2	University of California at Berkeley, USA
3	Chungnam National University, Korea
4	Language Technologies Institute Carnegie Mellon University, USA
5	Communications Research Laboratory, Japan
6	Dept. of Library and Information Science Fu Jen Catholic University, Taiwan
7	Hong Kong Polytechnic University, H.K.
8	Hummingbird, Canada
9	University of Library and Information Science, Japan
10	Institute of Software Chinese Academy of Sciences, China
11	Kent Ridge Digital Labs, Singapore
12	NLP Lab., Korea University, Korea
13	Microsoft Research China, China
14	Department of Computer Science and Information Engineering National Taiwan University, Taiwan
15	the University of Aizu, Japan
16	Osaka Kyoiku University, Japan
17	Queens College City University of New York, USA
18	Pohang University of Science & Technology Korea
19	Toyohashi University of Technology, Japan
20	R&D Group Thomson Legal & Regulatory, USA
21	Knowledge Media Laboratory Toshiba R&D Center, Japan
22	Computer Science Department University of Waterloo, Canada

Table 16. Number of SLIR Submitted Runs

Run Types	#of Runs	# of Groups	Total # of Runs	Total # of Groups
C-C	34	15	110	22
E-E	29	15		
J-J	30	13		
K-K	17	8		

Table 17. AvgPre of C-C Runs (Relaxed)

Topic Fields	# of Runs	MEAN	MAX	MIN
C	4	0.2605	0.2929	0.2403
D	14	0.2557	0.3617	0.0443
DC	1	0.2413	0.2413	0.2413
T	1	0.2467	0.2467	0.2467
TC	4	0.3109	0.3780	0.2389
TDC	1	0.3086	0.3086	0.3086
TDN	1	0.3499	0.3499	0.3499
TDNC	8	0.3161	0.4165	0.0862
Total	34	0.2806	0.4165	0.0443

Table 18. AvgPre of E-E Runs (Relaxed)

Topic Fields	# of Runs	MEAN	MAX	MIN
C	1	0.4422	0.4422	0.4422
D	13	0.2731	0.4649	0.0351
DC	1	0.0840	0.0840	0.0840
T	1	0.3526	0.3526	0.3526
TC	2	0.3682	0.4198	0.3166
TDC	1	0.3513	0.3513	0.3513
TDN	1	0.4472	0.4472	0.4472
TDNC	9	0.4574	0.5000	0.3229
Total	29	0.3476	0.5000	0.0351

From Figure 13, we find that 9 out of 12 top runs use probabilistic model, 9 out of 12 top runs use query expansion, and only one run use no stemming and stop words. Figure 14 shows that only 1 run out of 6 runs uses vector-space model. In short, stop words, stemming, query expansion, using all topic fields, and applying probabilistic model result in good E-E runs.

5.2.3 J-J Runs

Most participants of J-J runs adopt D and TDNC topic fields to carry out document search. Table 19 shows the distribution and the corresponding MEAN, MAX, and MIN of AvgPre. It is clear that TDNC runs perform well than other runs. Figure 15 and Figure 16 show the relaxed top J-J runs and rigid top J-J runs, respectively.

Table 19. AvgPre of J-J Runs (Relaxed)

Topic Fields	# of Runs	MEAN	MAX	MIN
C	2	0.2786	0.2898	0.2674
D	18	0.2633	0.3998	0.0460
DC	2	0.3454	0.4104	0.2804
T	2	0.3298	0.3562	0.3034
TC	2	0.3195	0.3990	0.2400
TDNC	4	0.4500	0.4932	0.3977
Total	30	0.3029	0.4932	0.0460

Observing Figure 15, we find most top runs use probabilistic model again. The common features of top runs are using morphological operations, query expansion, and TDNC topic fields.

5.2.4 K-K Runs

The D runs and TDNC runs are submitted by most of participants. Table 20 shows the distribution and the corresponding MEAN, MAX, and MIN of AvgPre. It is clear that TDNC runs perform well. Figure 17 and Figure 18 show the relaxed top K-K runs and rigid top K-K runs, respectively.

Figure 17 shows the top 2 runs are better than other top runs. The top 2 runs also use morphological operations, query expansion, and TDNC topic fields. The APL and HUM also show good results. APL uses TDNC fields and applies query expansion, but HUM does not apply query expansion techniques.

Table 20. AvgPre of K-K Runs (Relaxed)

Topic Fields	# of Runs	MEAN	MAX	MIN
C	2	0.2807	0.2938	0.2675
D	9	0.2501	0.3602	0.1256
T	1	0.3317	0.3317	0.3317
TC	1	0.3569	0.3569	0.3569
TDNC	4	0.4356	0.5022	0.3367
Total	17	0.3084	0.5022	0.1256

5.3 BLIR Track

Totally, 50 runs are submitted from 14 groups in SLIR track. Table 21 shows the BLIR participants and Table 22 shows the number of submitted runs. According to the Table 22, C-J, E-C, E-J, E-K, and J-C have at least 2 groups. Therefore, we would like to examine C-J, E-C, E-J, E-K, and J-C runs only. The results and the corresponding techniques of other runs could be referred to the Appendix.

5.3.1 C-J Runs

Only 4 C-J runs are submitted by 2 groups. Figure 19 and Figure 20 show all relaxed C-J runs and rigid C-J runs, respectively. Brkly group submits a C-J run and performs well. It uses query expansion, logistic regression model and character-based indexing. The query translation is a two-stage procedure. The first is to translate Chinese to English using corpus-based method. The second is to translate English to Japanese using MT-based method. The ISCAS submits 3 C-J runs. It uses dictionary-based and MT-based methods to translate query. It is a word-based and vector-space model system without query expansion. The mean AvgPre of C-J runs is 0.0891, but that of J-J runs is 0.3029.

Table 21. List of BLIR Participants

1	Applied Physics Laboratory Johns Hopkins University, USA
2	University of California at Berkeley, USA
3	Chungnam National University, Korea
4	Language Technologies Institute Carnegie Mellon University, USA
5	University of Library and Information Science, Japan
6	Institute of Software Chinese Academy of Sciences, China
7	NLP Lab., Korea University, Korea
8	Microsoft Research China, China
9	Osaka Kyoiku University, Japan
10	Queens College City University of New York, USA
11	Toyohashi University of Technology, Japan
12	Tianjin University
13	R&D Group Thomson Legal & Regulatory, USA
14	Knowledge Media Laboratory Toshiba R&D Center, Japan

Table 22. Number of BLIR Submitted Runs

Run Types	#of Runs	# of Groups	Total # of Runs	Total # of Groups
C-E	3	1	50	14
C-J	4	2		
E-C	16	6		
E-J	11	5		
E-K	6	2		
J-C	5	3		
J-E	1	1		
K-C	2	1		
K-E	2	1		

5.3.2 E-C Runs

16 E-C runs are submitted from 6 groups. Most runs are D runs and TDNC runs. TDNC runs perform well than other runs. Table 23 shows the distribution and the corresponding MEAN, MAX, and MIN of AvgPre. Figure 21 and Figure 22 show the relaxed top E-C runs and rigid top E-C runs, respectively. MSRA group submits 3 E-C runs. The first 2 of MSRA's runs are the top 2. The 3 runs use the same techniques except the third run uses no query expansion. In addition, only one top run uses vector space model. With comparison to 0.2806, the mean AvgPre of C-C runs, the mean AvgPre of E-C runs is 0.1443.

Table 23. AvgPre of E-C Runs (Relaxed)

Topic Fields	# of Runs	MEAN	MAX	MIN
D	6	0.1144	0.1921	0.0098
T	1	0.0092	0.0092	0.0092
TC	1	0.0426	0.0426	0.0426
TDC	2	0.2273	0.2322	0.2223
TDNC	6	0.1861	0.2781	0.0291
Total	16	0.1443	0.2781	0.0092

5.3.3 E-J Runs

11 E-J runs are submitted from 5 groups. Most runs are D runs and TDNC runs. 4 out of 5 groups use probabilistic model. Most runs are D runs and all top runs are D runs. Table 24 shows the distribution and the corresponding MEAN, MAX, and MIN of AvgPre. Figure 23 and Figure 24 show the relaxed top E-J runs and rigid top E-J runs, respectively.

Table 24. AvgPre of E-J Runs (Relaxed)

Topic Fields	# of Runs	MEAN	MAX	MIN
D	7	0.2015	0.3404	0.0000
DC	2	0.0479	0.0957	0.0001
T	1	0.1024	0.1024	0.1024
TDNC	1	0.1531	0.1531	0.1531
Total	11	0.1602	0.3004	0.0000

The top three runs are submitted from TSB group. TSB uses probabilistic model, morphological analysis, query expansion, and MT-based query translation. Be aware that the mean AvgPre of J-J run is 0.3029 and that of E-J runs is 0.1602.

5.3.4 E-K Runs

6 E-J runs are submitted from 2 groups. All top runs are submitted from KUNLP group. Figure 25 and Figure 26 show the relaxed top E-K runs and rigid top E-K runs, respectively. The top one run is an interactive run. KUNLP use probabilistic model, word-based indexing, and dictionary-based and select-top-2 query translation. The top one run is interactive and query expanded. The top two run is automatic and query expanded. The top three run is automatic without query expansion. Be aware that the mean AvgPre of E-K runs 0.1134 and that of K-K run is 0.3084.

5.3.5 J-C Runs

5 J-C runs are submitted from 3 groups. Figure 27 and Figure 28 show the relaxed top J-C runs and rigid top J-C runs, respectively. Each group has one run in the top runs. One top run is an interactive run and it is the top one run. OKSAT group submit the interactive J-C run (OKSAT-J-C-D-01). This run use probabilistic model, character-based indexing for Japanese and word-base indexing for Chinese, dictionary-based query translation, and query expansion. Be aware that the mean AvgPre of J-C runs 0.1010 and that of C-C run is 0.2806.

Before moving to MLIR Track, we would like to show the relaxed mean AvgPre of the available retrieval tasks of SLIR and BLIR in Table 25. It is much clearer for us to know how far we have gone in CLIR researches in this way. As aforementioned, some retrieval tasks have few submitted runs. In order not to mislead the interested readers, the number of submitted runs and that of groups is also shown under corresponding retrieval tasks. N/A means no groups submit the query-document pair runs.

Table 25. Relaxed AvgPre of Retrieval Tasks

Topic \ Doc	C (#run/#grp)	E (#run/#grp)	J (#run/#grp)	K (#run/#grp)
C	0.2086 (34/15)	0.1055 (3/1)	0.0891 (4/2)	N/A
E	0.1443 (16/6)	0.3476 (29/15)	0.1602 (11/5)	0.1134 (6/2)
J	0.1010 (5/3)	0.2149 (1/1)	0.3029 (30/13)	N/A
K	0.0258* (2/1)	0.3148* (2/1)	N/A	0.3084 (17/8)

* K-E and K-C runs are submitted from member of CLIR Executive Committee.

5.4 MLIR

Totally, 29 runs are submitted from 7 groups in MLIR track. Table 26 shows the MLIR participants and Table 27 shows the number of submitted runs. Generally speaking, MLIR is the most difficult task in CLIR. Few groups participate this task. According to Table 27, C-CJE, E-CE, E-CJE, and J-JE have at least 2 groups, we would like to examine them only.

Table 26. List of MLIR Participants

1	Applied Physics Laboratory Johns Hopkins University, USA
2	University of California at Berkeley, USA
3	University of Library and Information Science, Japan
4	Institute of Software Chinese Academy of Sciences, China
5	Department of Computer Science and Information Engineering National Taiwan University, Taiwan
6	Queens College City University of New York, USA
7	Toyohashi University of Technology, Japan

Table 27. Number of MLIR Submitted Runs

Run Types	#of Runs	# of Groups	Total # of Runs	Total # of Groups
C-CE	3	1	29	7
C-CJ	3	1		
C-JE	3	1		
C-CJE	4	2		
E-CE	6	2		
E-CJE	4	2		
E-JE	1	1		
J-JE	2	2		
J-CJE	3	1		

5.4.1 C-CJE Runs

4 runs are submitted from 2 groups. Figure 29 and Figure 30 show the relaxed C-CJE runs and rigid C-CJE runs, respectively. Table 28 shows the AvgPre of these runs. The MLIR track is more complicated. The performance is, of course, lower than that of BLIR and SLIR. Brkly-C-CJE-D-01 is the best run in C-CJE runs, which is the result of direct merge of C-C, C-E, and C-J. Brkly group combines these runs and re-ranks the retrieved documents by using the estimated probability of relevance. The final result consists of the top-ranked 1000 documents per topic.

Table 28. AvgPre of C-CJE Runs

Run ID	Topic Fields	Relaxed	Rigid
Brkly-C-CJE-D-01	D	0.1835	0.1462
ISCAS-C-CJE-D-03	D	0.0920	0.0644
ISCAS-C-CJE-TC-01	TC	0.1263	0.0927
ISCAS-C-CJE-TC-02	TC	0.1264	0.0925

ISCAS group uses vector space model, bigram for Chinese, word for Japanese and English, and dictionary and MT-based query translation. The ISCAS runs are also the merged lists of C-C, C-J, and C-E with different selection criteria. The details could be referred the ISCAS's paper.

5.4.2 E-CE Runs

6 runs are submitted from 2 groups. Figure 31 and Figure 32 show the relaxed E-CE runs and rigid E-CE runs, respectively. Table 29 shows the performance of E-CE runs.

Table 29. AvgPre of E-CE Runs

Run ID	Topic Fields	Relaxed	Rigid
ISCAS-E-CE-D-03	D	0.1651	0.1203
ISCAS-E-CE-TC-01	TC	0.0713	0.0623
ISCAS-E-CE-TDNC-02	TDNC	0.0999	0.0839
pircs-E-EC-D-001	D	0.1620	0.1198
pircs-E-EC-D-002	D	0.1577	0.1158
pircs-E-EC-D-003	D	0.1320	0.0986

ISCAS group and pircs group use different criteria to merge the results of corresponding SLIR and BLIR to form the final results of E-EC runs. Therefore, the performance of E-CE runs depends on the E-C and E-E runs and the selection criteria. ISCAS uses the predefined number of documents for E-E and E-C runs. Pircs group manages to find a ratio for the composition of E-CE run from E-E and E-C runs. Please refer to the respective papers for detailed information.

5.4.3 E-CJE Runs

4 runs are submitted from 2 groups. Figure 33 and Figure 34 show the relaxed E-CJE runs and rigid E-CJE runs, respectively. Table 30 shows the performance of E-CE runs.

Table 30. AvgPre of E-CJE Runs

Run ID	Topic Fields	Relaxed	Rigid
APL-E-CEJ-D-02	D	0.0620	0.0547
APL-E-CEJ-T-03	T	0.0622	0.0492
APL-E-CEJ-TDNC-01	TDNC	0.0759	0.0681
Brkly-E-CJE-D-01	D	0.1535	0.1287

Brkly-E-CJE-D-01 shows good performance. This run is the direct merging of Brkly-E-C-D-01, Brkly-E-J-D-01, and Brkly-E-E-D-01. The same method mentioned in C-CJE runs is also used to create the final E-CJE run.

5.4.4 J-JE Runs

2 runs are submitted from 2 groups. Only one run is shown in Figure 35 and Figure 36. IFLAB-J-JE-D-01 performs much well than the other run. IFLAB uses probabilistic model, word-based indexing for query and document, morphology and stemming, and corpus- and dictionary-based translation with select-top-1.

6. Conclusions

The third NTCIR Workshop is the first joint efforts of Japan, Korea and Taiwan in providing an evaluation mechanism for real multilingual CLIR task not only in queries but also in documents. We hope CLIR task could encourage the CLIR researches in Eastern Asia, promote the concept of IR evaluation, provide an opportunity to share the research ideas and results, investigate the useful

techniques for CLIR researches, and enhance the effectiveness of CLIR systems.

After the preliminary analyses on the submitted runs, some findings are shown in the following. The detailed information of particular CLIR systems could be referred to the respective reports in this workshop.

- More groups use probabilistic model than ever before and show better performance.
- Most groups apply inverted file approach for index structure.
- Many groups adopt tf/idf-based weighting scheme.
- Query expansion is a cutting-edge technique for good retrieval performance
- Stopword list is a good resource for enhancing system performance.
- Among 15 groups which have submitted BLIR and MLIR runs, 7 adopts dictionary-based translation, 3 MT-based translation, 2 corpus-based translation, 2 MT+dict-based translation, and 1 MT+corpus-based translation.

Since we have 16 E-C runs and 11 E-J runs, it is more appropriate to examine the effectiveness of translation techniques in these two categories.

- MSRA group adopts dictionary-based translation in E-C runs and shows good performance.
- TSB group adopts MT-based translation in E-J runs and shows good performance.

It is hard to say which approach is good in this situation. We have another observation on the two groups. MSAR is familiar with Chinese and TSB group is familiar with Japanese. Maybe they hold the edge in languages.

The C-C runs in the second NTCIR workshop show very good performance [5]. We have discussed this point in the report of the second NTCIR workshop. The <concepts> of topics of the second NTCIR workshop are created very carefully using positive and negative examples in deep natural language analysis. The <CONC> field of this year's topic is created in more user-oriented way. Therefore, the performance of C-C runs in this year looks more "normal" and is comparable to J-J and K-K runs.

We would like to say again that these findings are drawn from the submitted runs of the third NTCIR workshop. Although, after several years devoting in evaluation of information retrieval, some common views have been in form, each test collection has its own characteristics and each language also has its own idiosyncrasy. The readers have to examine each paper carefully and do not reach a conclusion directly from the results. We have to carry out more detailed analyses using other test collections.

We have described the refined test collection for evaluation in NTCIR workshop in Section 4. However, for the uses of the test collection after NTCIR workshop, we will create another so-called **NTCIR-3 Reference Test Collection** in order to eliminate the complicated situation, which maybe confuse the non-participants of NTCIR workshop. This test collection contains 2 sub- test collections which are created based on "2-in-S+A" criterion. The Reference Test Collection has only one common Topic Set for CJE target languages. These topics have at least 2 "S+A" documents in each of three target document sets (C, J and E). **NTCIR-3 Reference Test Collection** also has one Topic Set for Korean target language.

1. **NTCIR-3 CJE Reference Test Collection** (Remember that this collection is 1998-1999-based.)
   It contains Chinese DocSet, English DocSet and Japanese DocSet. The users of this test collection could freely combine different target language DocSets. The Chinese DocSet contains 381,681 Chinese documents. The English DocSet contains 22,927 English documents. The Japanese DocSet contains 220,078 Japanese documents. Only one topic set with 31 topics is prepared for this test collection. The IDs of these topics selected from 1998-1999 Topic Set are 2, 4, 5, 7, 10, 12, 14, 15, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 32, 33, 34, 35, 36, 37, 38, 39, 42, 43, 45, 46, and 50.
2. **NTCIR-3 Korean Reference Test Collection**
   In fact, this is the same as the NTCIR-3 Korean Formal Test Collection. It contains 66,146 Korean documents and 30 topics in Chinese, Japanese, Korean, and English. All of the 30 topics in 1994 Topic Set are used in this collection.

Of course, we will prepare the relevance judgment (Answer Set) of each sub-test collection for Formal Test Collection and Reference Test Collection. On the one hand, if the users of our test collection do not feel the NTCIR-3 Formal Test Collection complicated, they would have more topics to be used for their IR researches. On the other hand, if they prefer to a simpler situation, use NTCIR-3 Reference Test Collection. The details could be referred to the NTCIR Website (http://research.nii.ac.jp/ntcir/).

Acknowledgments

We would like to thank China Times, Commercial Times, China Times Express, Central Daily News, China Daily News, United Daily News, Taiwan News, Mainichi Newspaper, Mainichi Daily, and Korea Economic Daily for their kindly providing news articles. We are grateful to CLEF and TREC for their helps in creating topics. We would like to thank the participants for their contributions and the assessors for their hard working in relevance judgment. The last but not the least, we would like to thank Emi Ishida for her hard working for CLIR task.

References

[1] Kando, Noriko. Overview of the Second NTCIR Workshop. In *Proceedings of the Second NTCIR Workshop on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, pages 51-72, Tokyo, 2001.

[2] NTCIR. http://research.nii.ac.jp/ntcir/

[3] CLEF. http://clef.iei.pi.cnr.it:2002/

[4] TREC. http://trec.nist.gov/

[5] Chen, Kuang-hua and Chen, Hsin-Hsi. The Chinese Text Retrieval tasks of NTCIR Workshop 2. In *Proceedings of the Second NTCIR Workshop on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, pages 51-72, Tokyo, 2001.

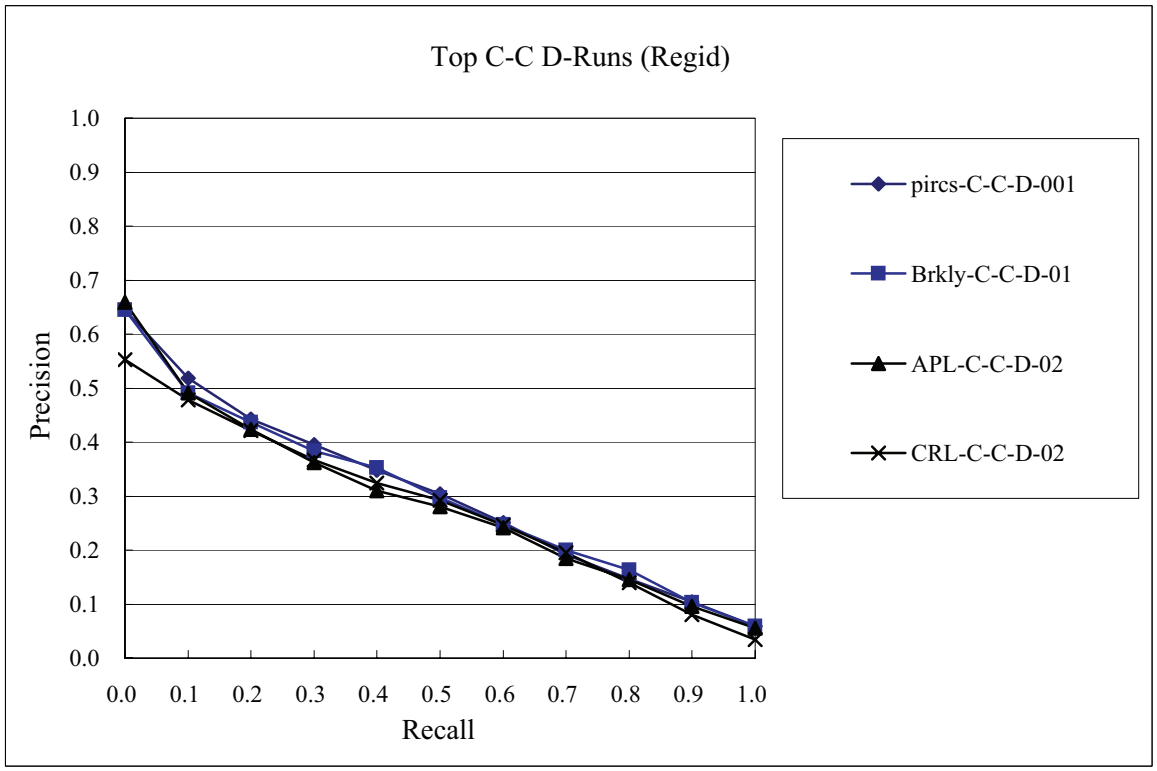

Figure 3(a). CLIR Task: Top C-C D-Runs (Rigid)

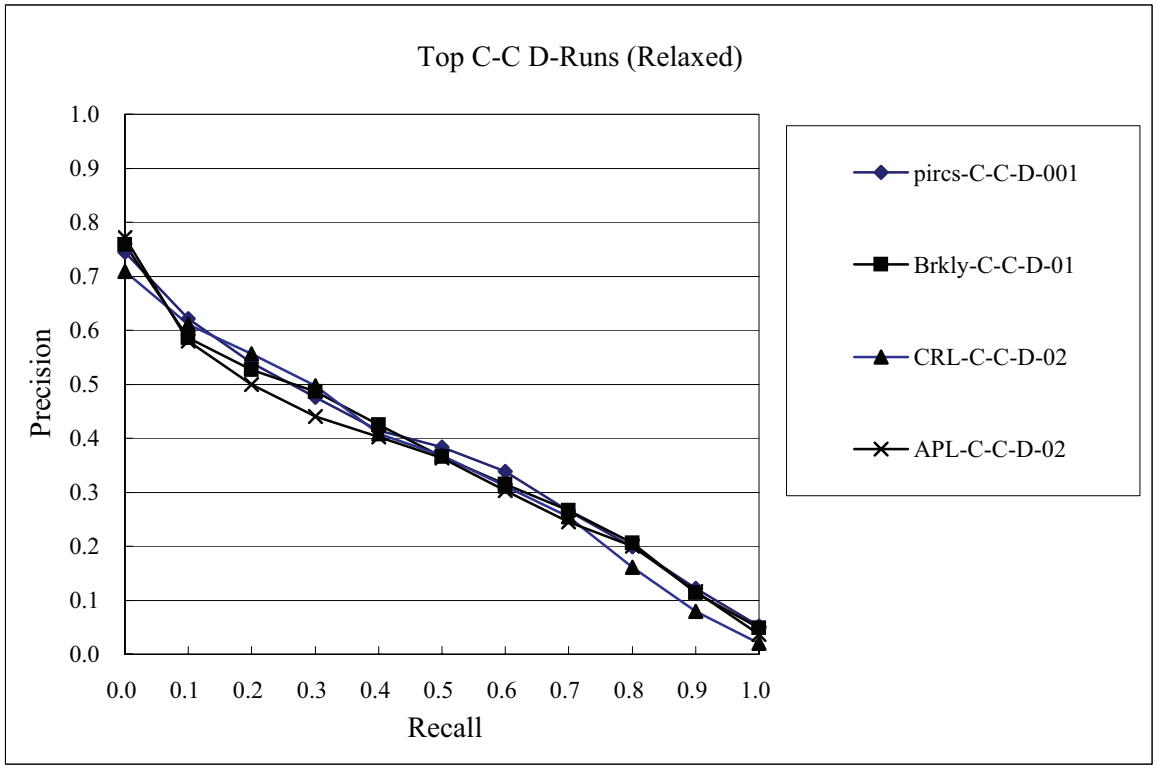

Figure 3(b). CLIR Task: Top C-C D-Runs (Relaxed)

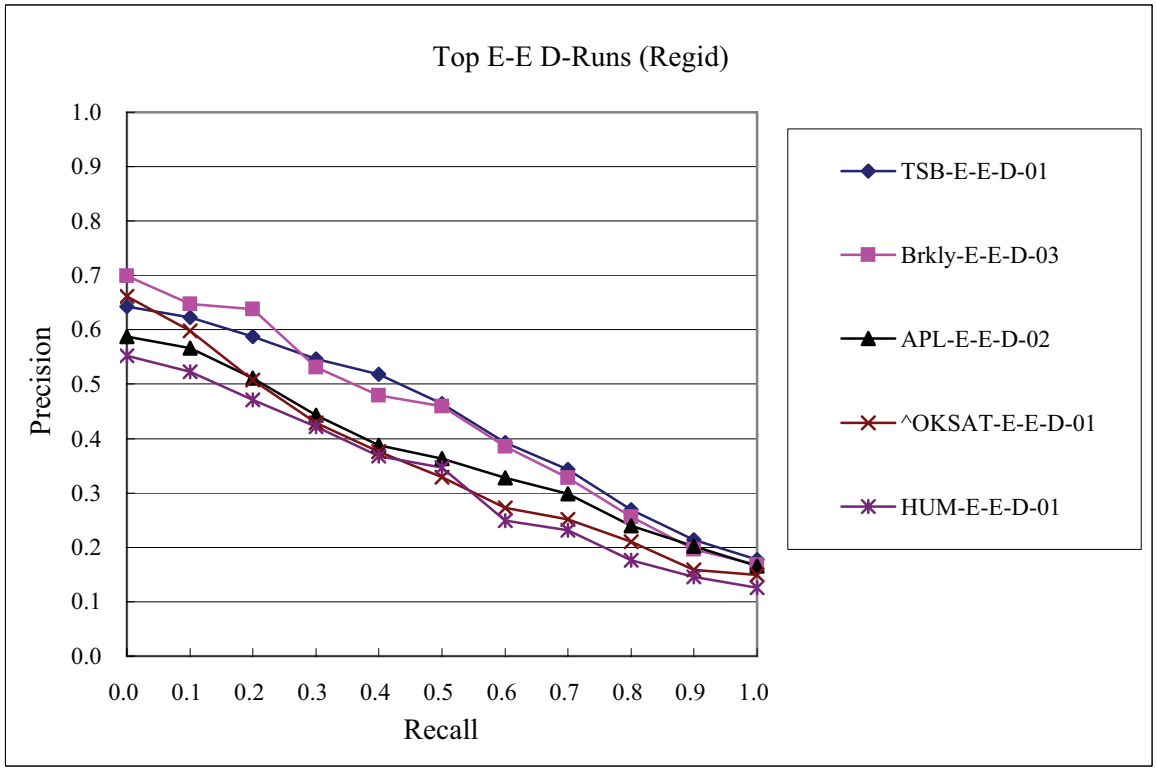

Figure 4(a). CLIR Task: Top E-E D-Runs (Rigid)

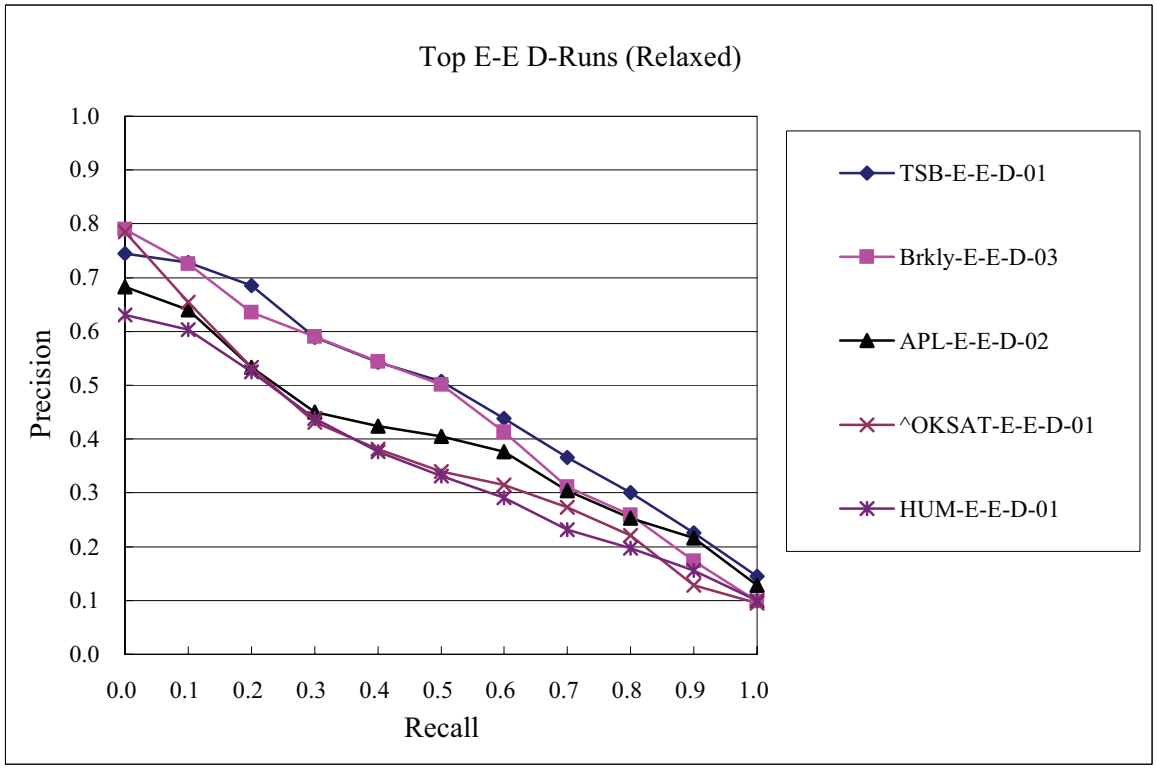

Figure 4(b). CLIR Task: Top E-E D-Runs (Relaxed)

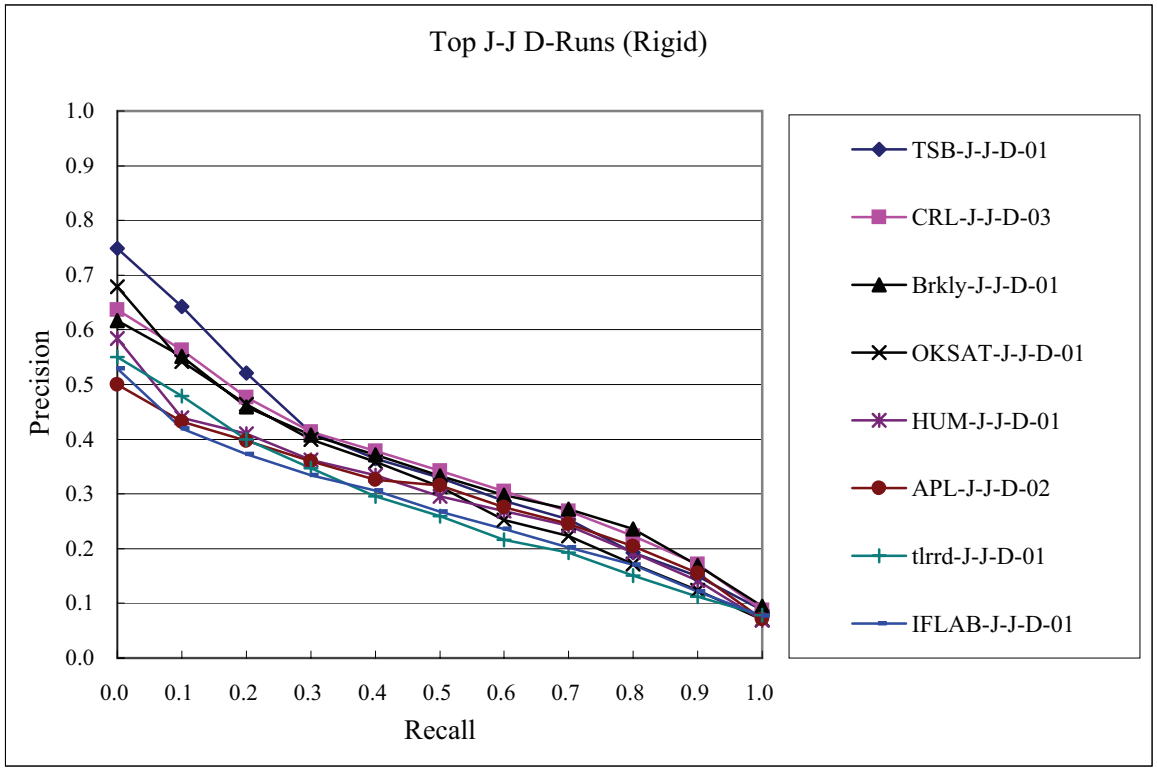

Figure 5(a). CLIR Task: Top J-J D-Runs (Rigid)

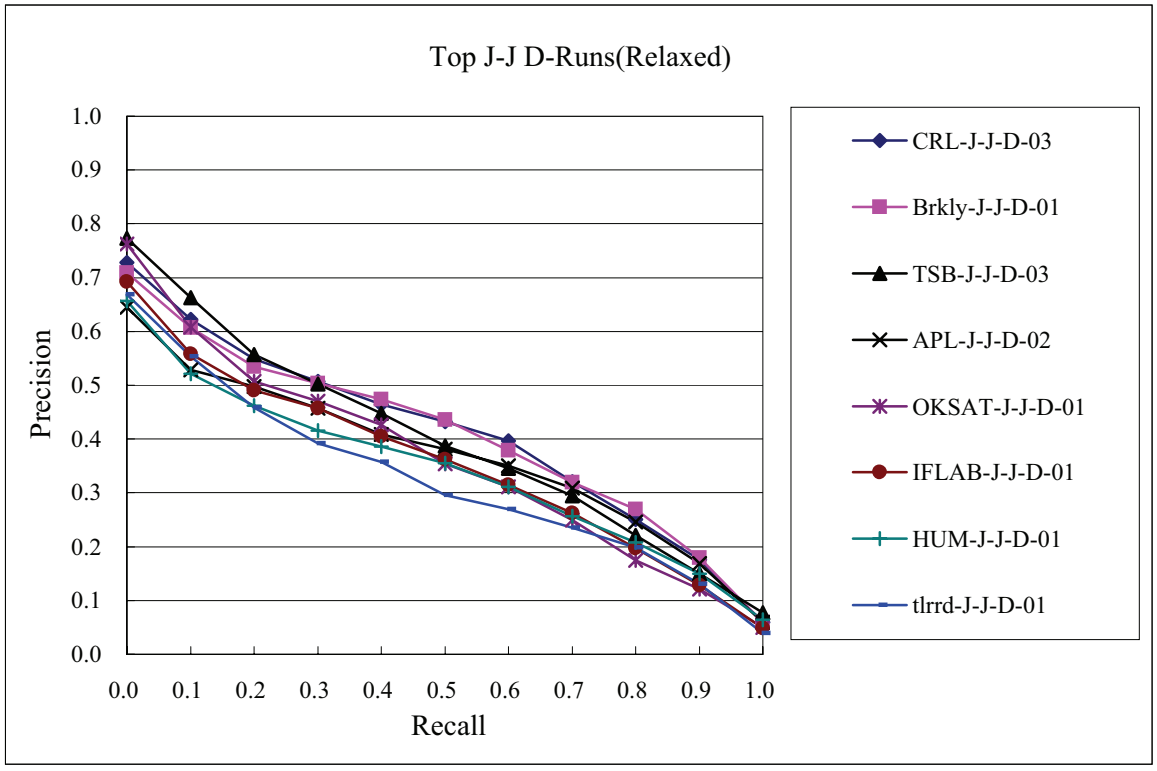

Figure 5(b). CLIR Task: Top J-J D-Runs (Relaxed)

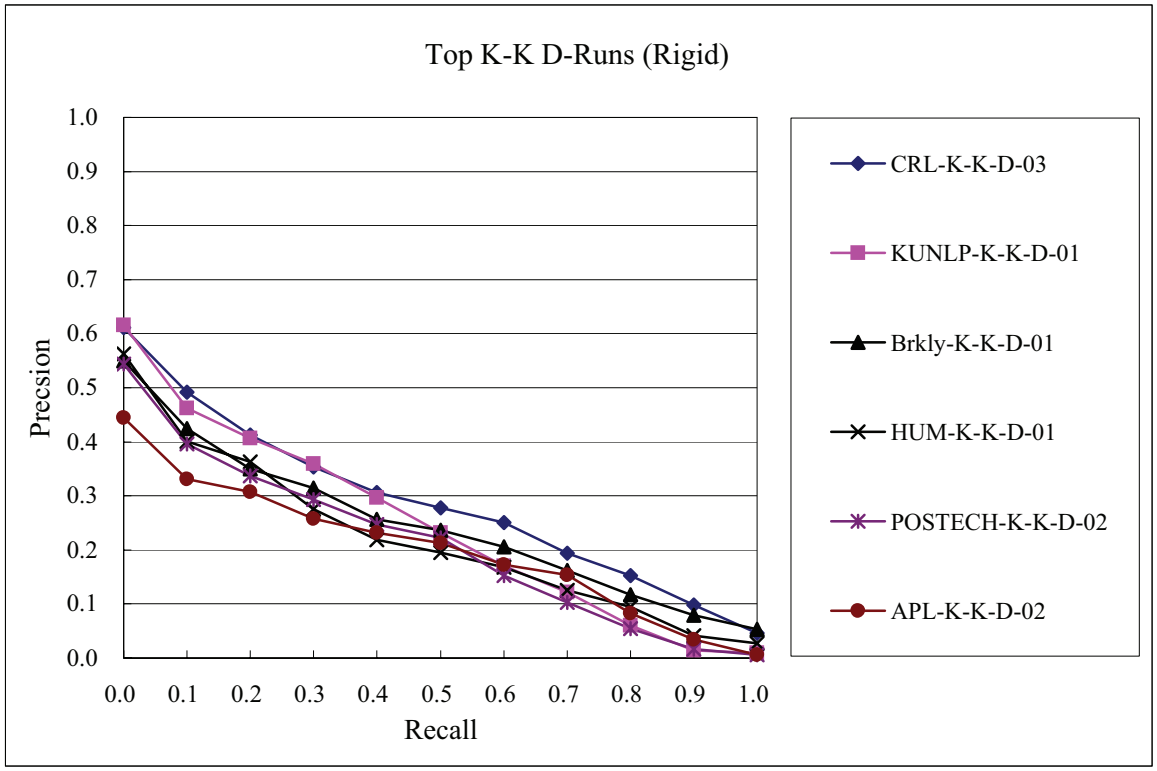

Figure 6(a). CLIR Task: Top K-K D-Runs (Rigid)

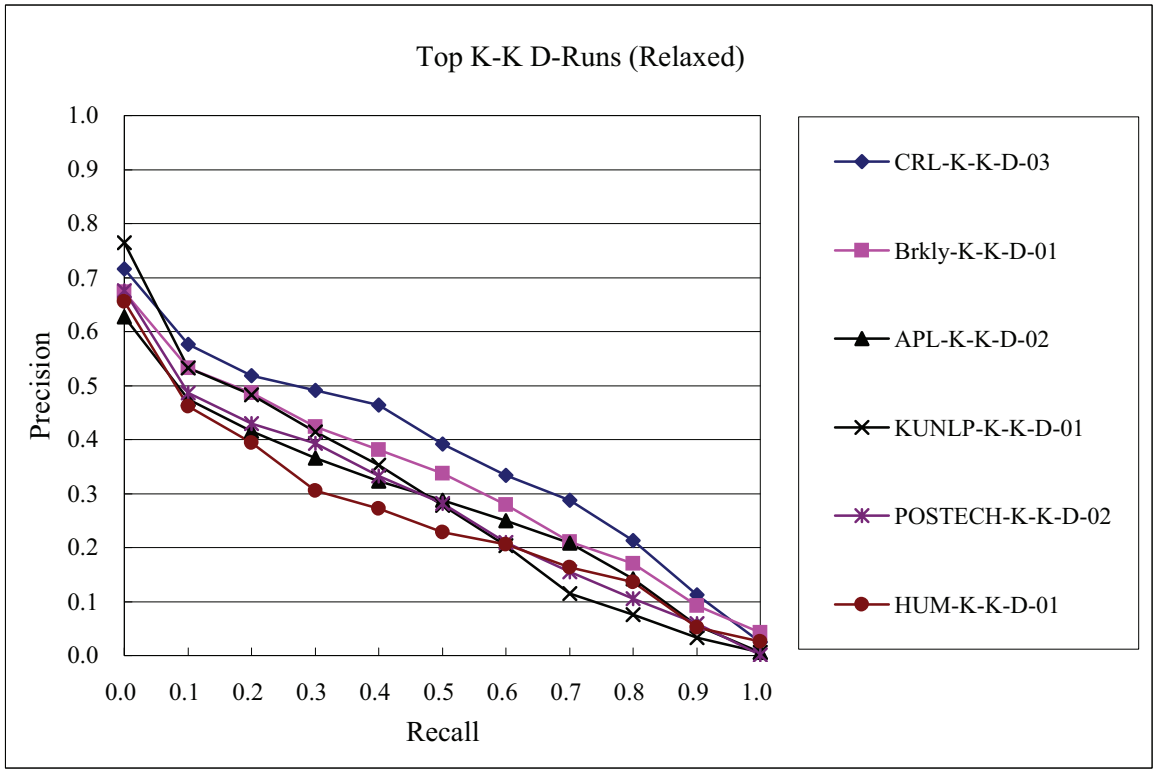

Figure 6(b). CLIR Task: Top K-K D-Runs (Relaxed)

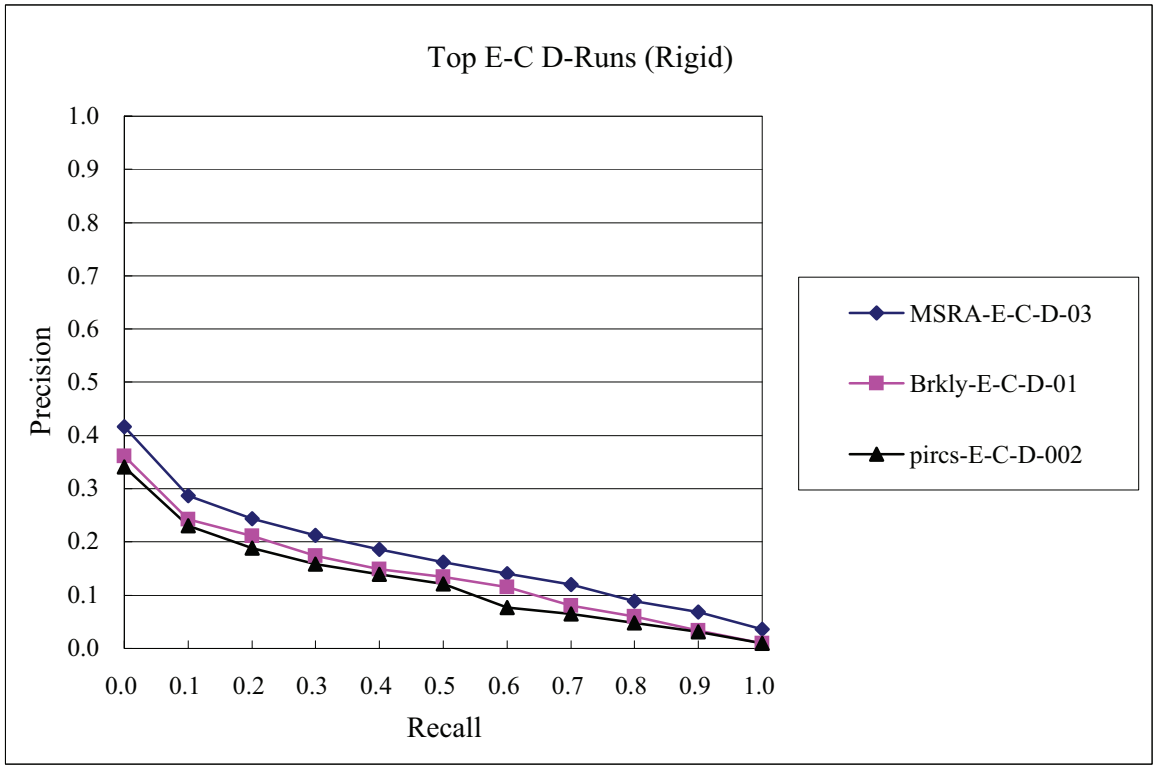

Figure 7(a). CLIR Task: Top E-C D-Runs (Rigid)

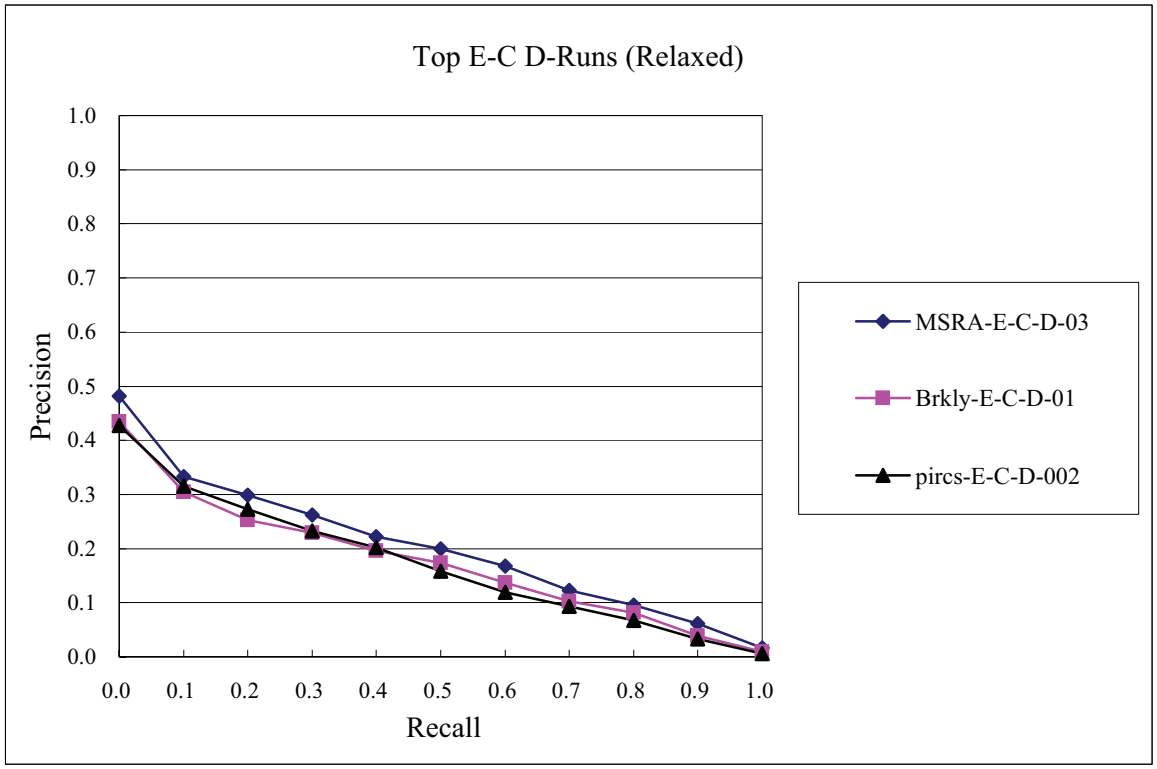

Figure 7(b). CLIR Task: Top E-C D-Runs (Relaxed)

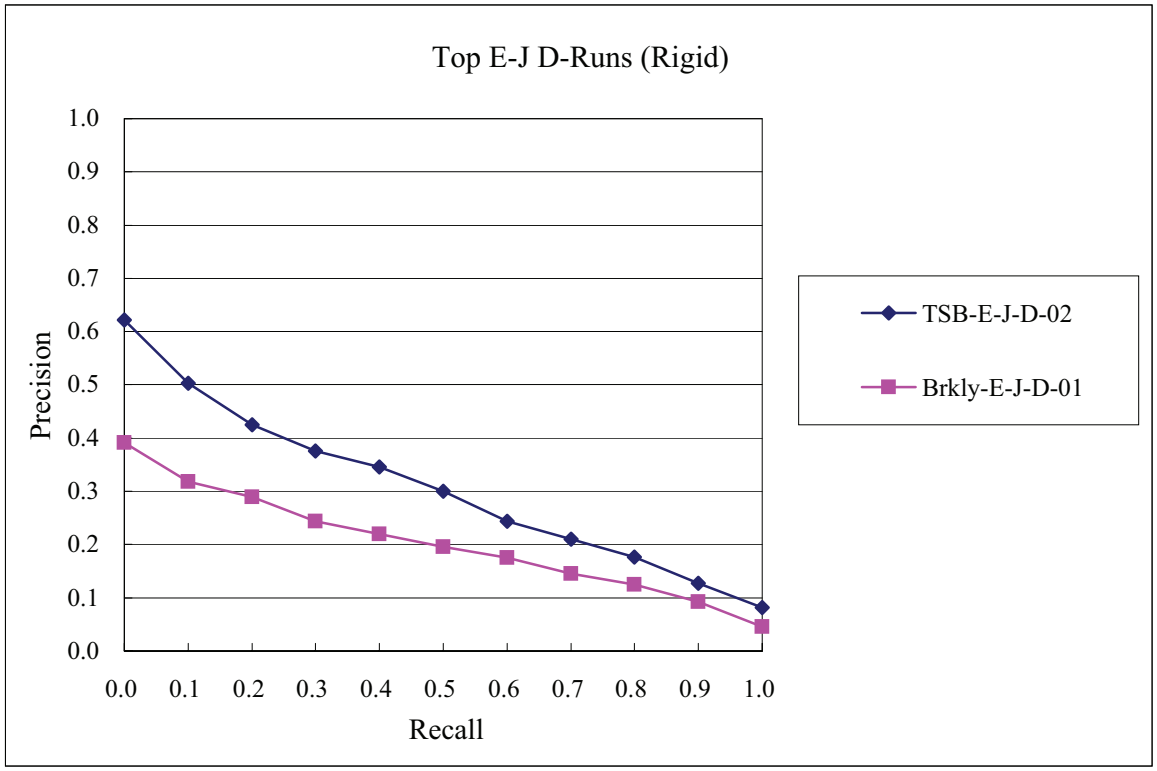

Figure 8(a). CLIR Task: Top E-J D-Runs (Regid)

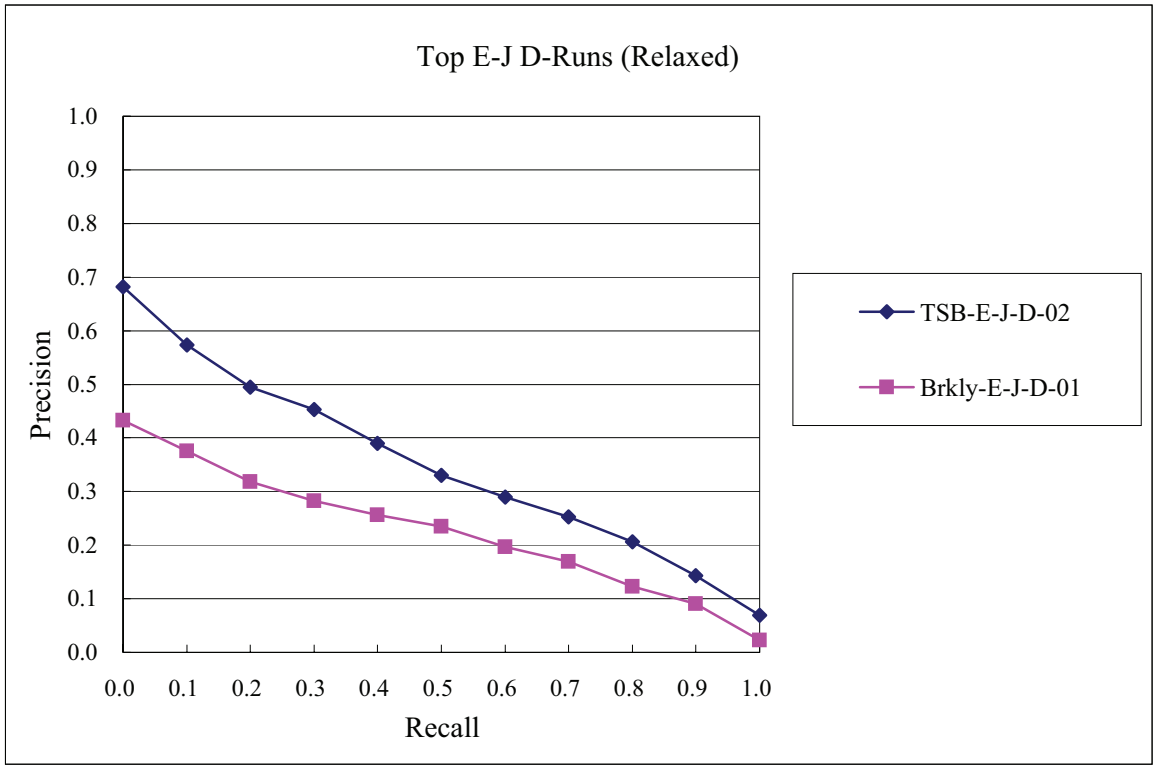

Figure 8(b). CLIR Task: Top E-J D-Runs (Relaxed)

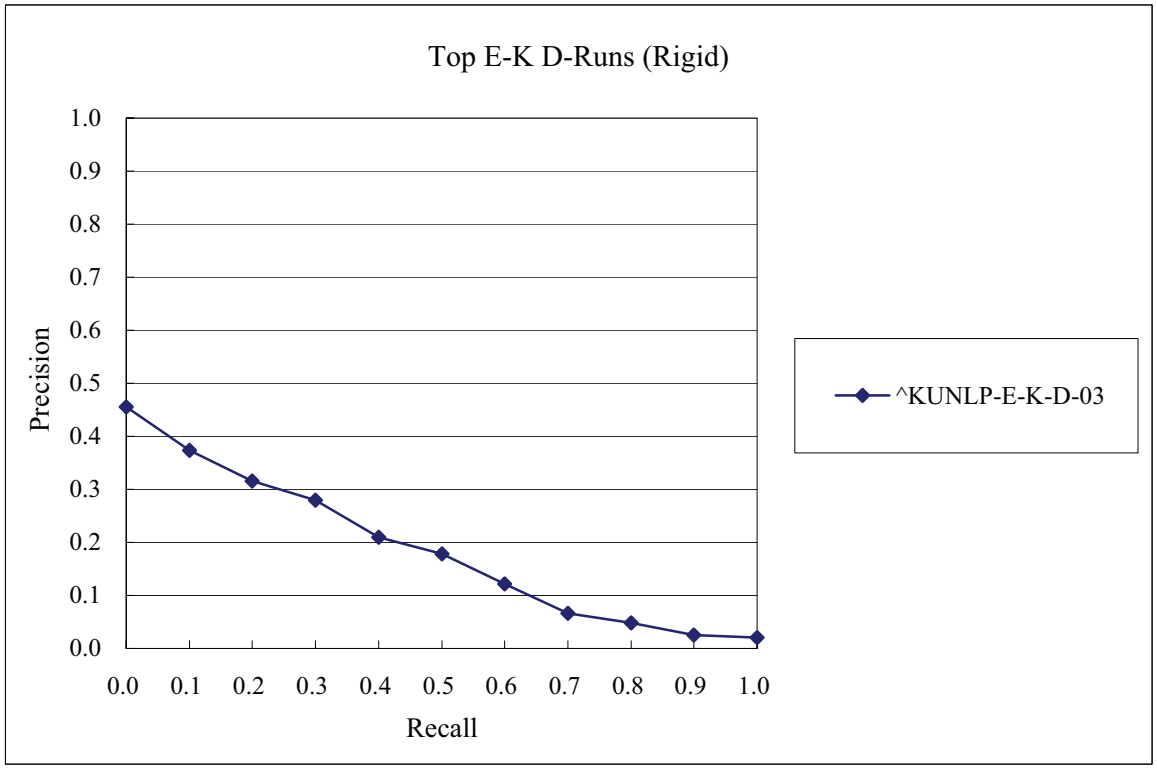

Figure 9(a). CLIR Task: Top E-K D-Runs (Rigid)

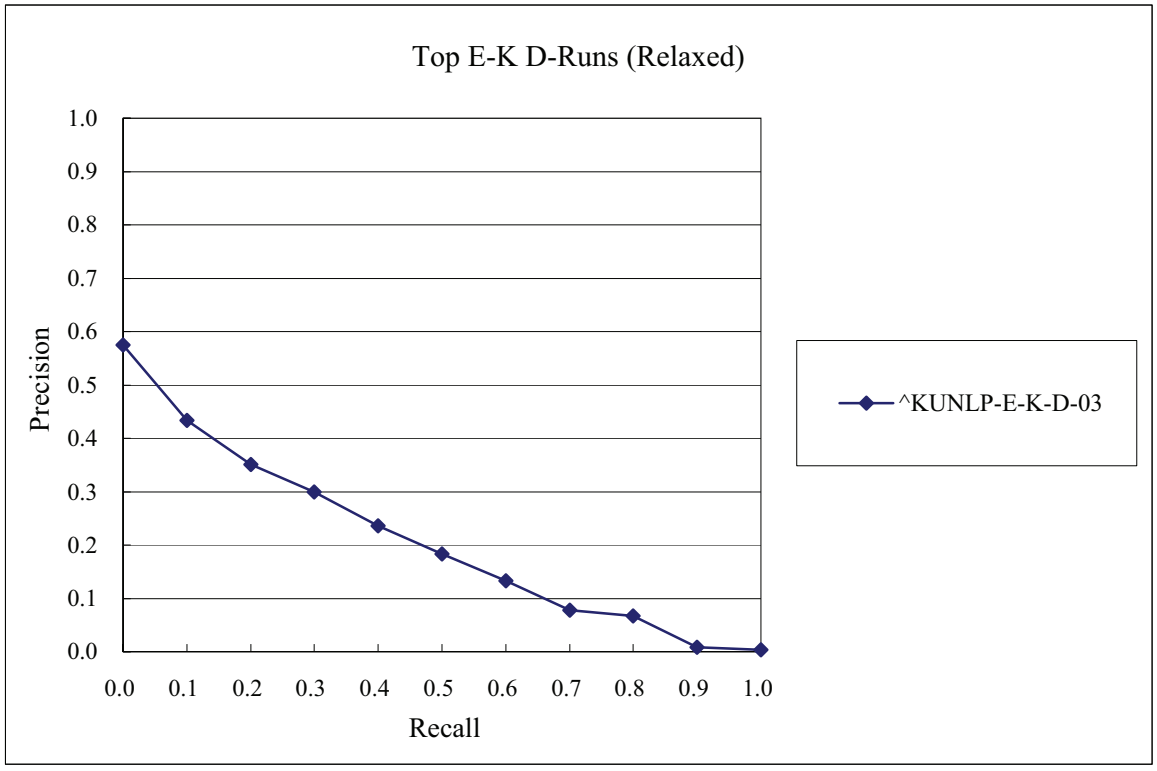

Figure 9(b). CLIR Task: Top E-K D-Runs (Relaxed)

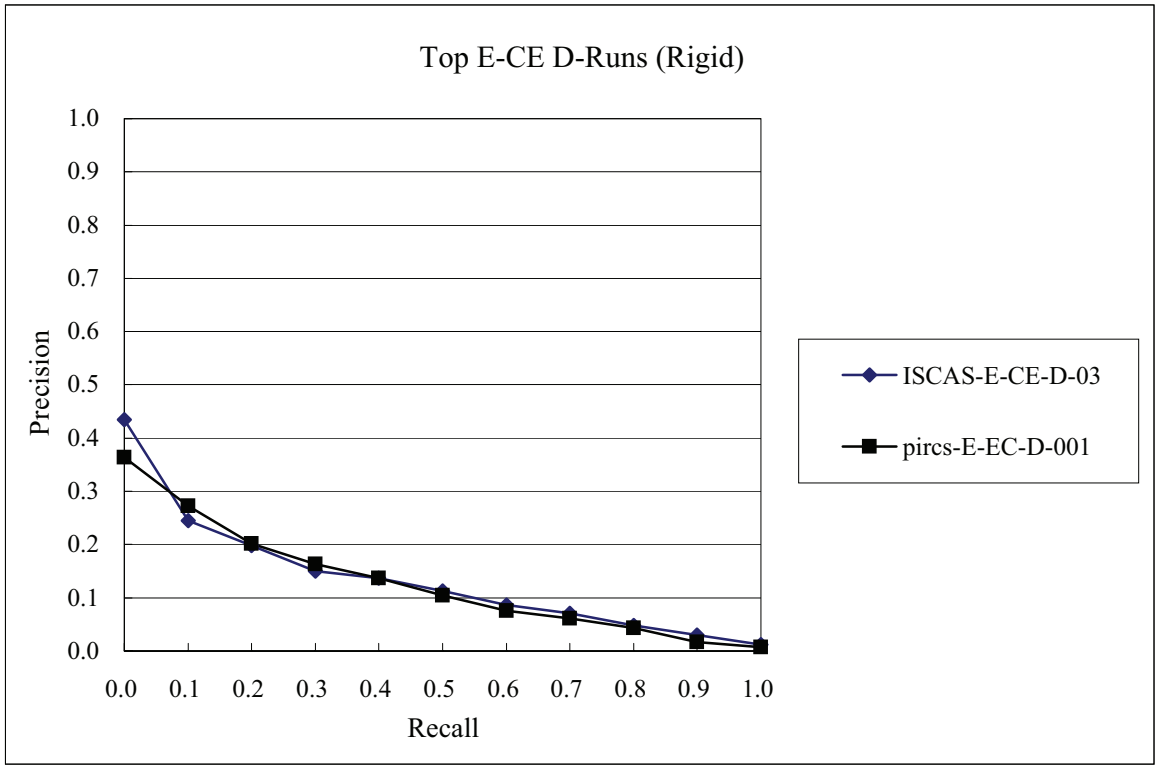

Figure 10(a). CLIR Task: Top E-CE D-Runs (Rigid)

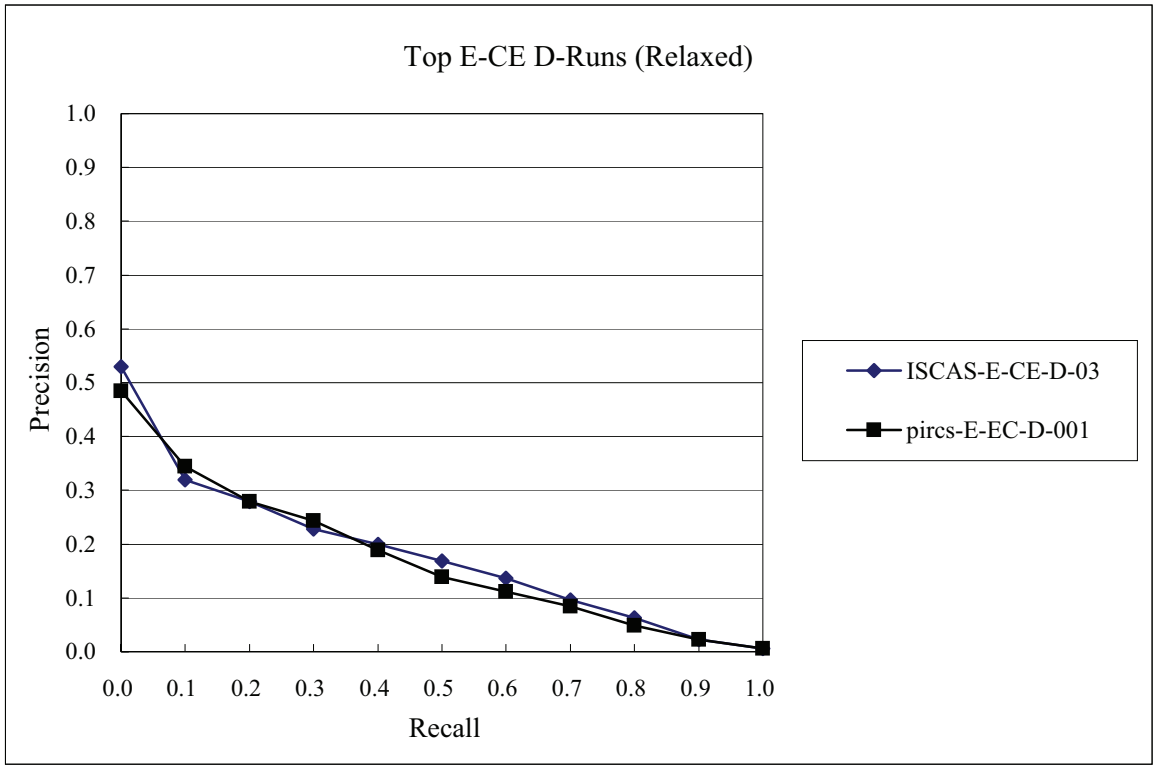

Figure 10(b). CLIR Task: Top E-CE D-Runs (Relaxed)

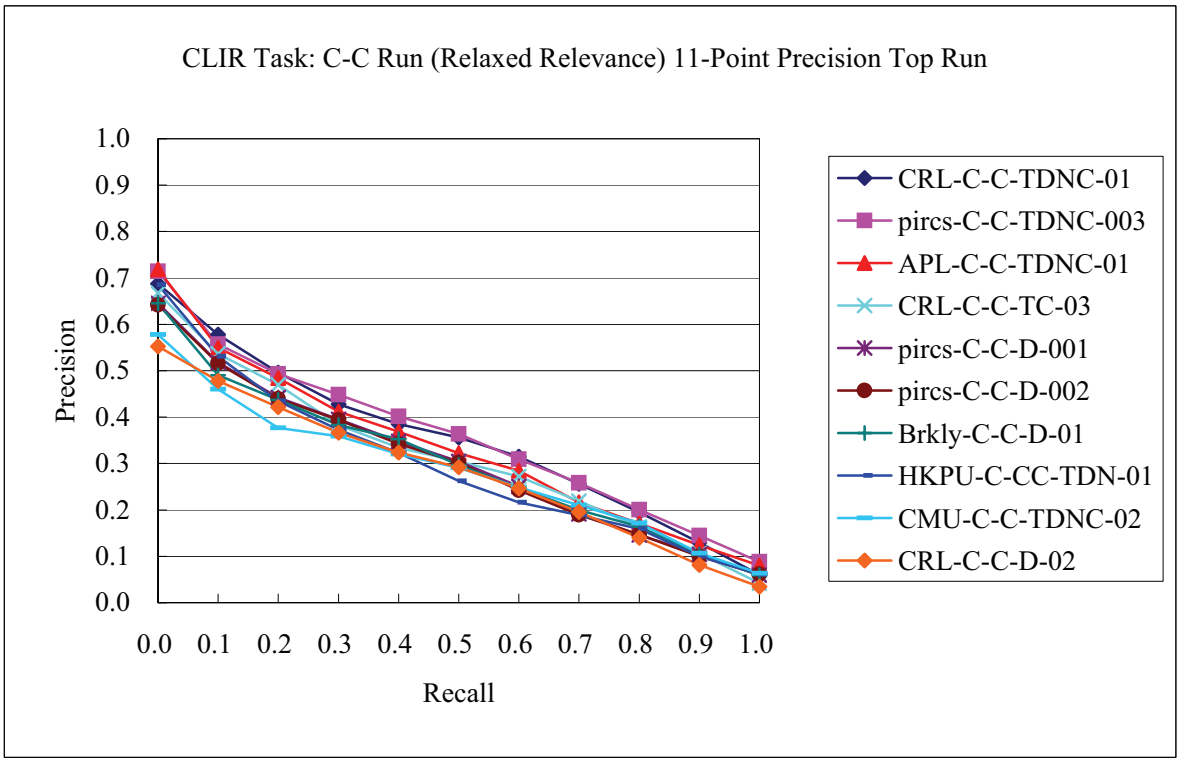

Figure 11. CLIR Task: C-C Run 11-Point Precision (Relaxed Relevance)

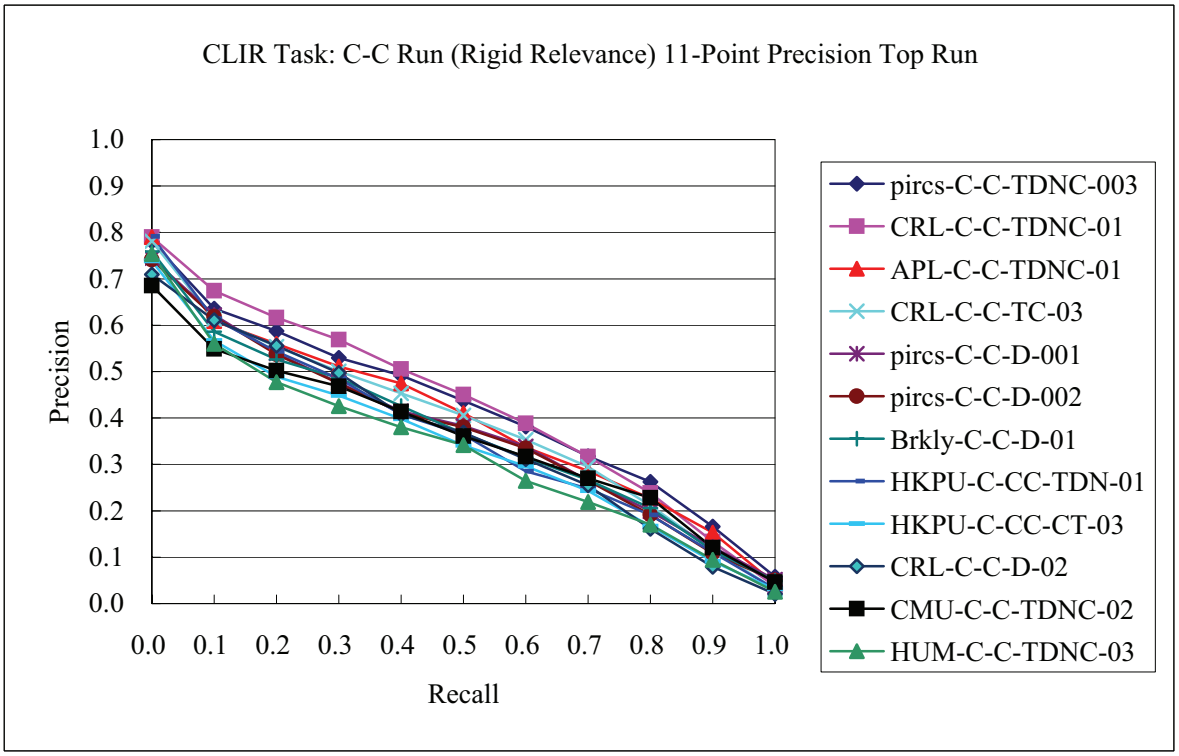

Figure 12. CLIR Task: C-C Run 11-Point Precision (Rigid Relevance)

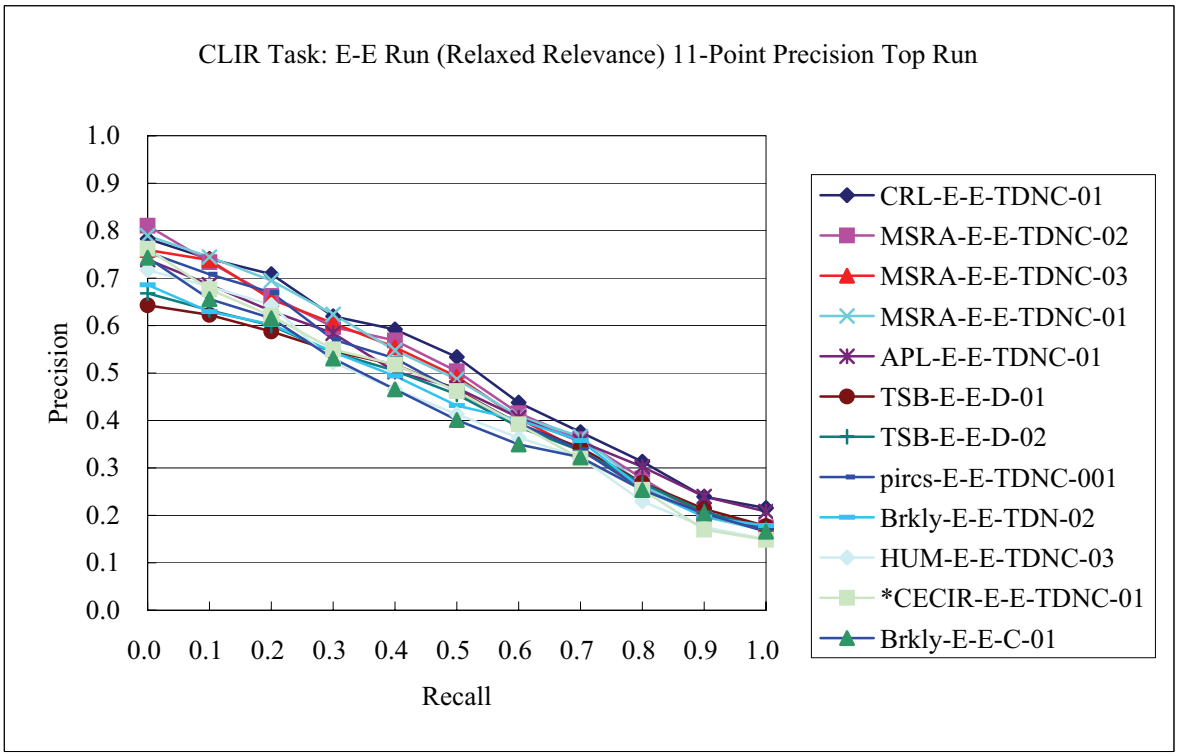

Figure 13. CLIR Task: E-E Run 11-Point Precision (Relaxed Relevance)

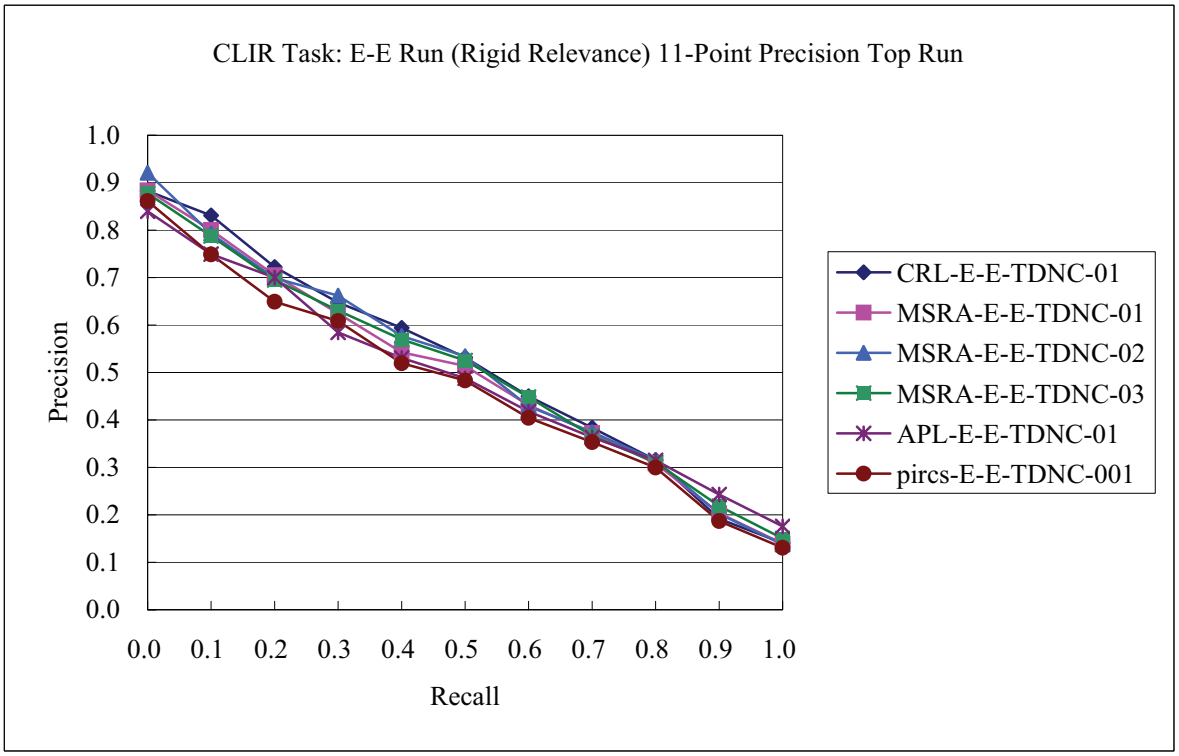

Figure 14. CLIR Task: E-E Run 11-Point Precision (Rigid Relevance)

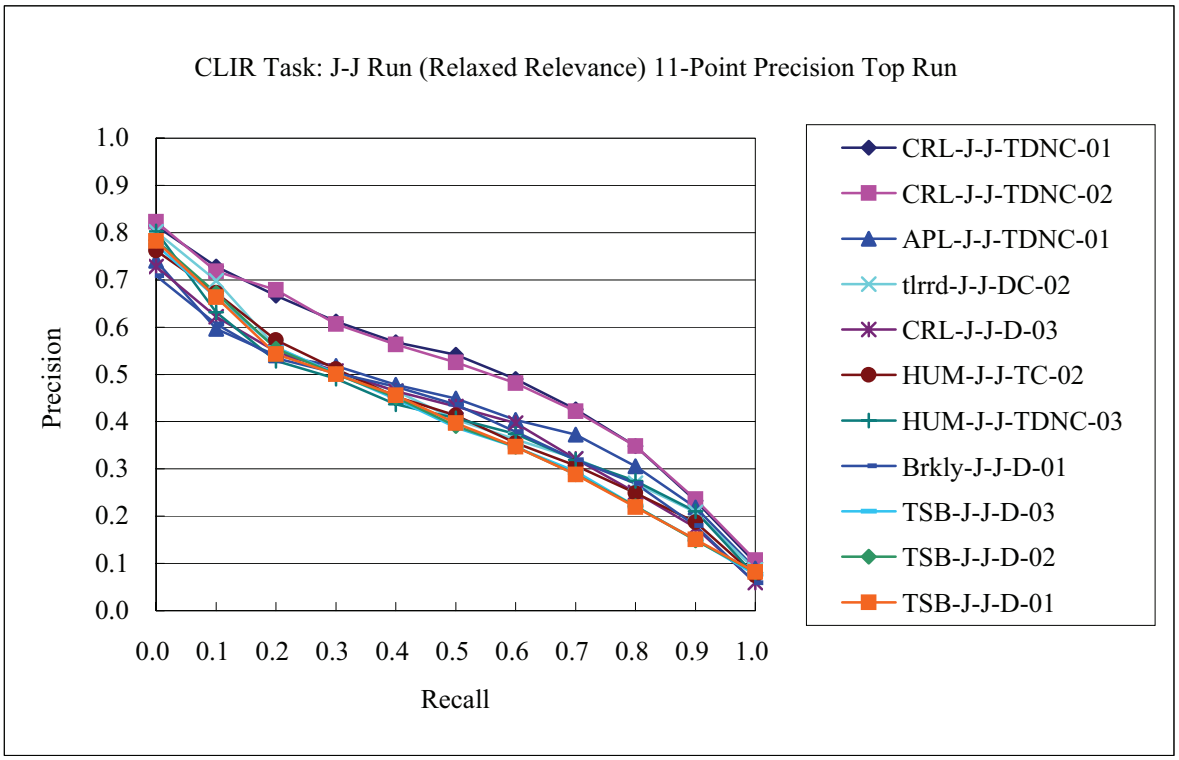

Figure 15. CLIR Task: J-J Run 11-Point Precision (Relaxed Relevance)

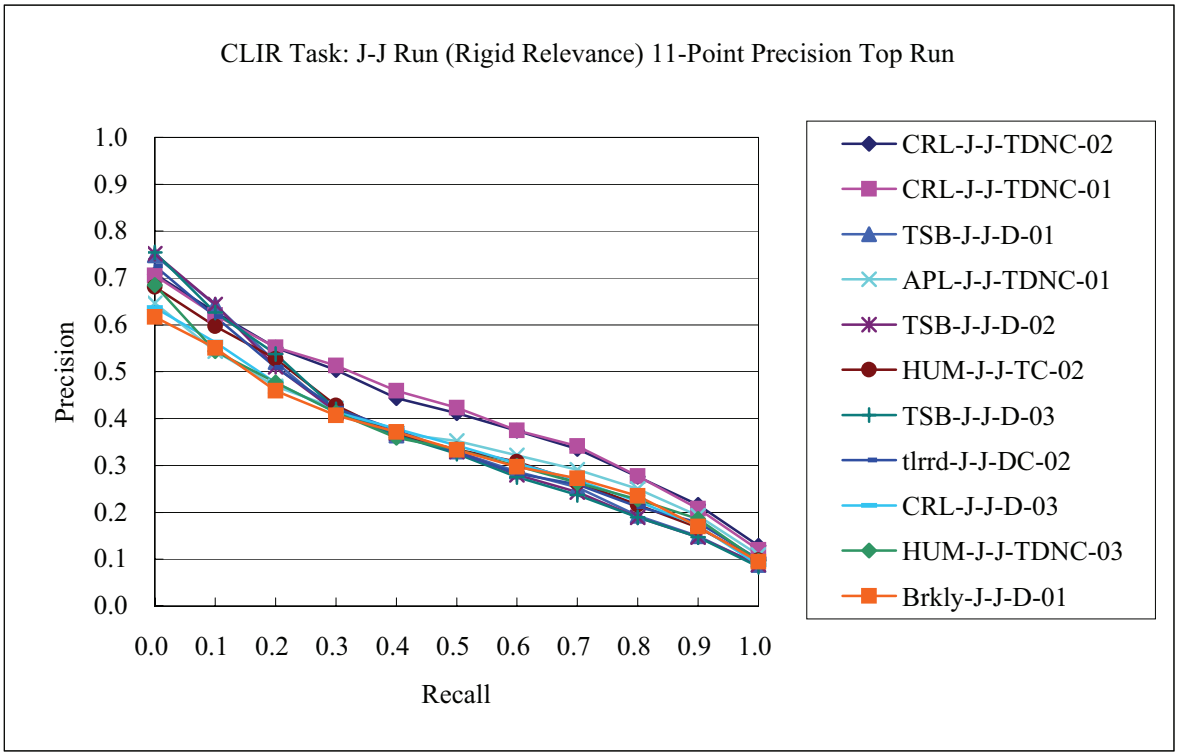

Figure 16. CLIR Task: J-J Run 11-Point Precision (Rigid Relevance)

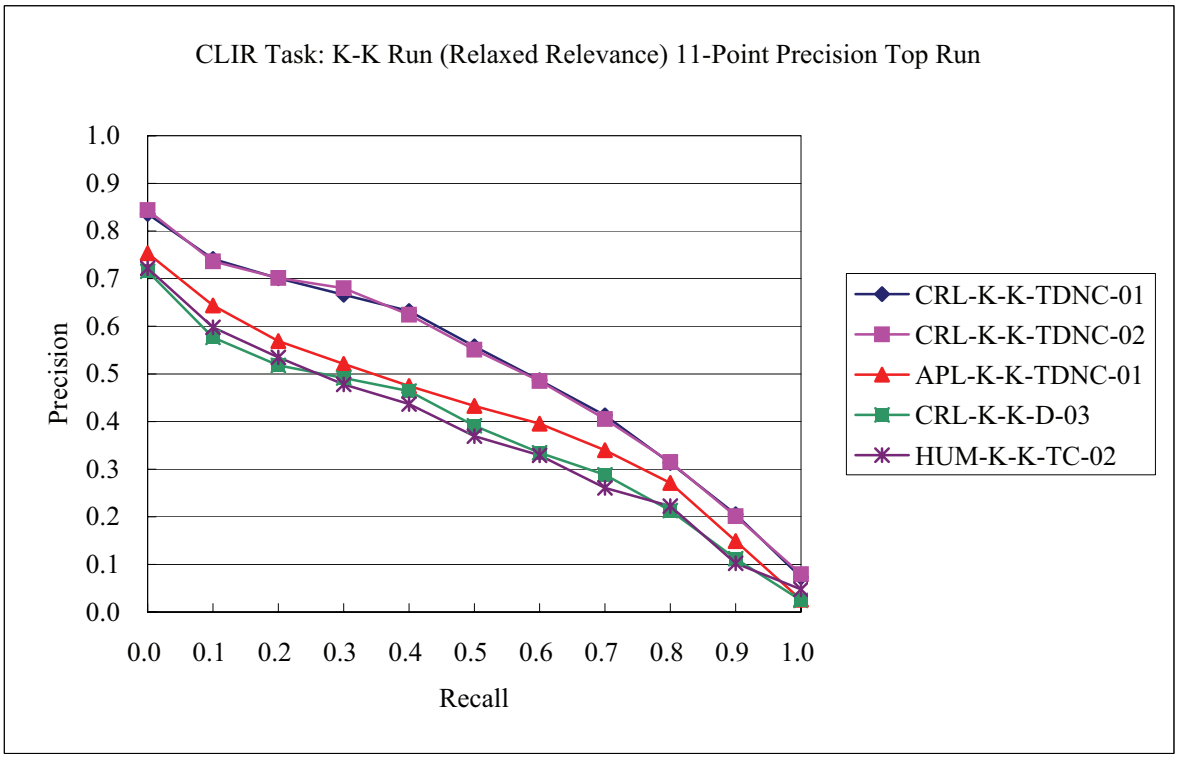

Figure 17. CLIR Task: K-K Run 11-Point Precision (Relaxed Relevance)

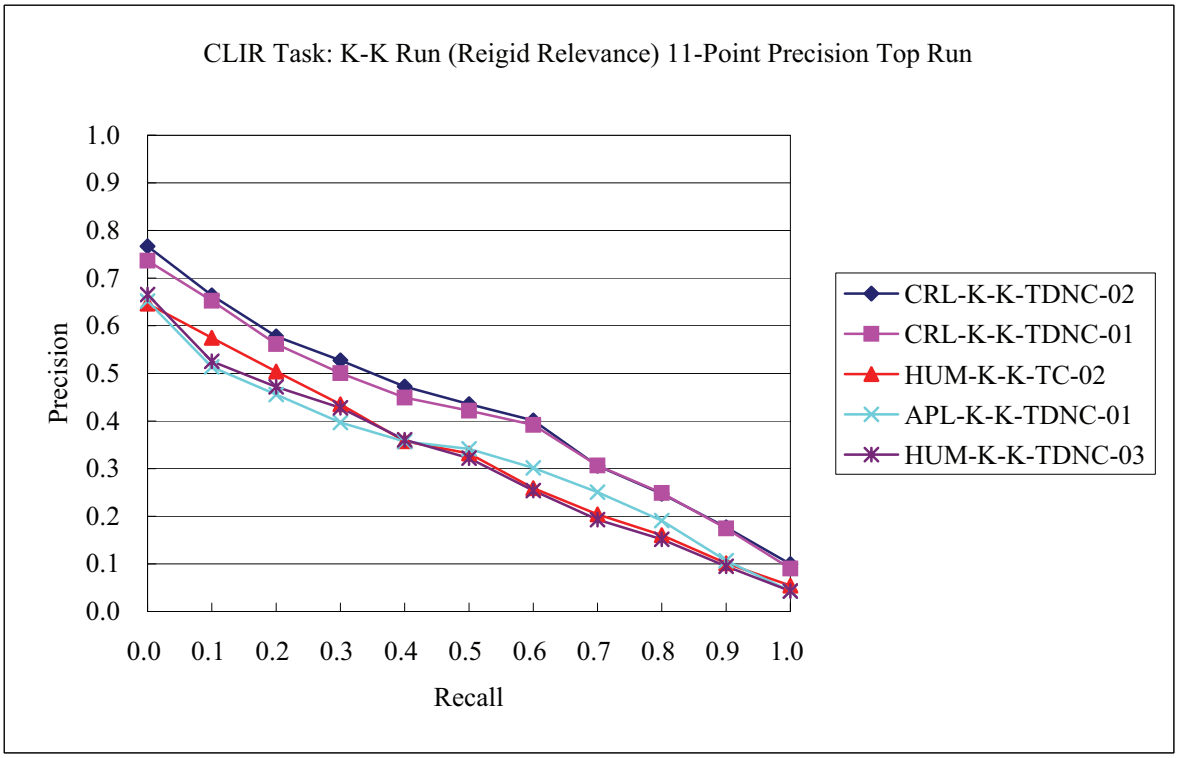

Figure 18. CLIR Task: K-K Run 11-Point Precision (Rigid Relevance)

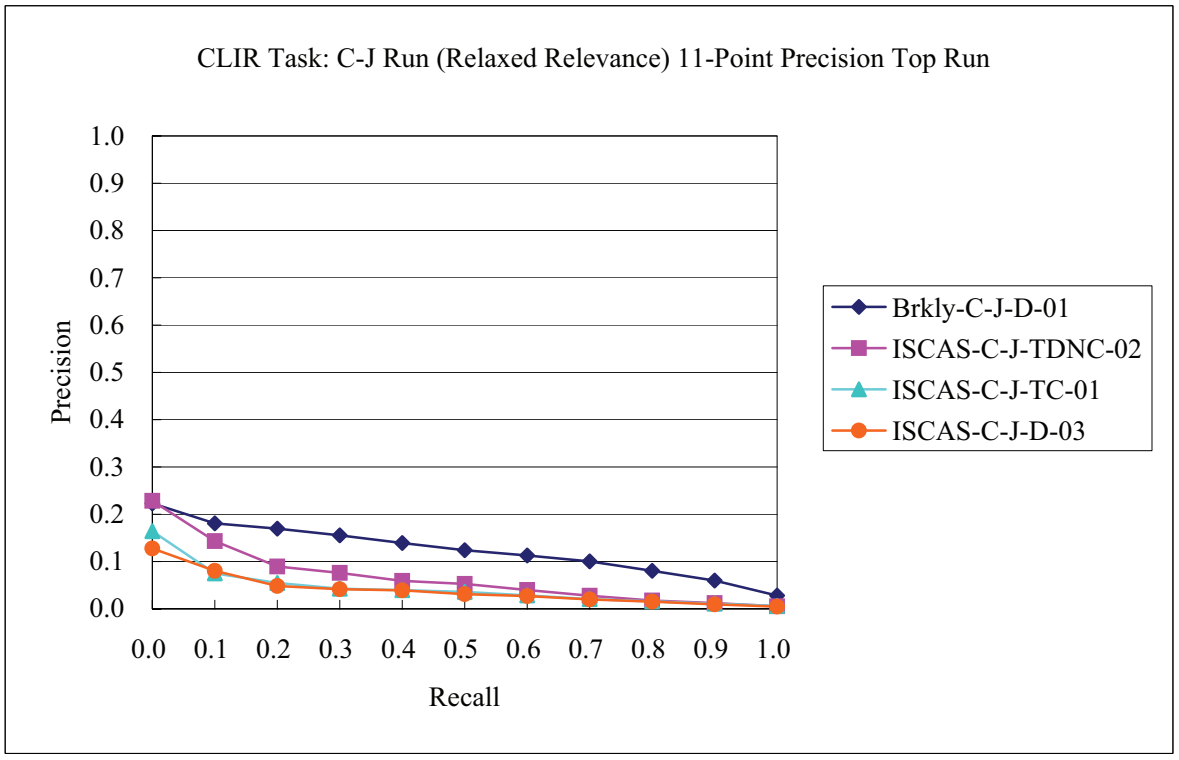

Figure 19. CLIR Task: C-J Run 11-Point Precision (Relaxed Relevance)

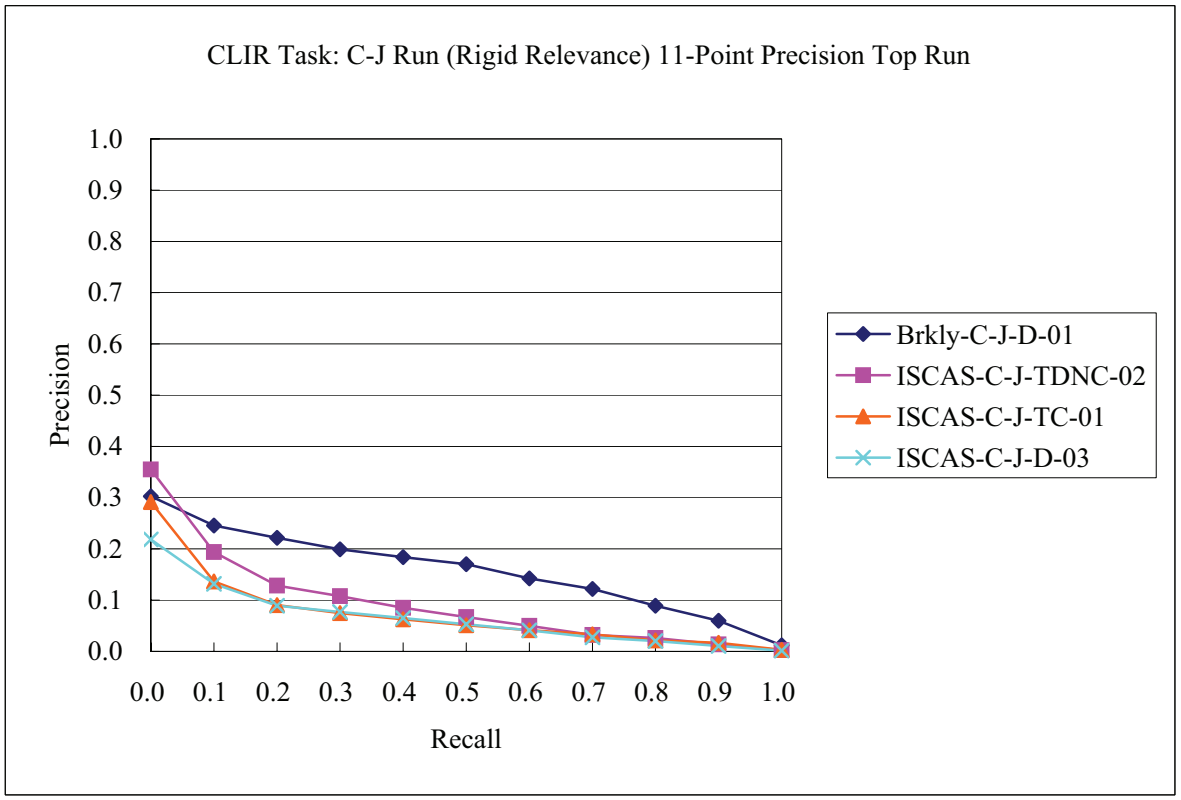

Figure 20. CLIR Task: C-J Run 11-Point Precision (Rigid Relevance)

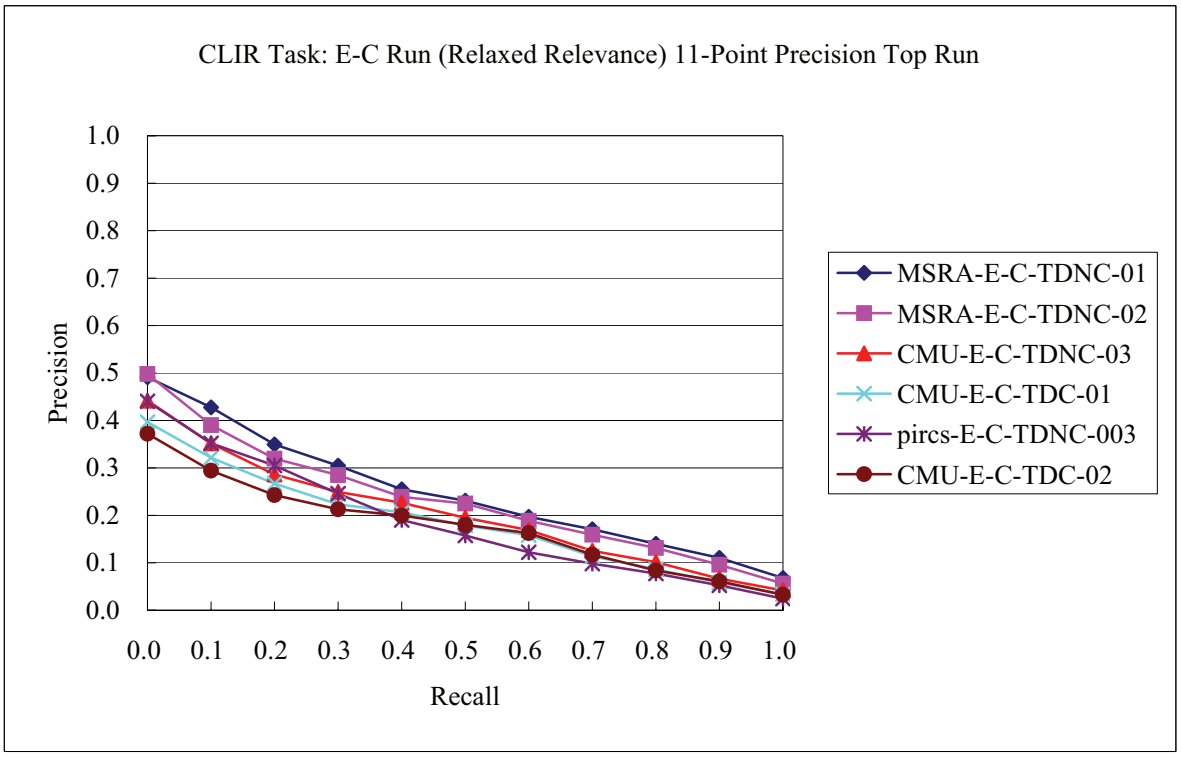

Figure 21. CLIR Task: E-C Run 11-Point Precision (Relaxed Relevance)

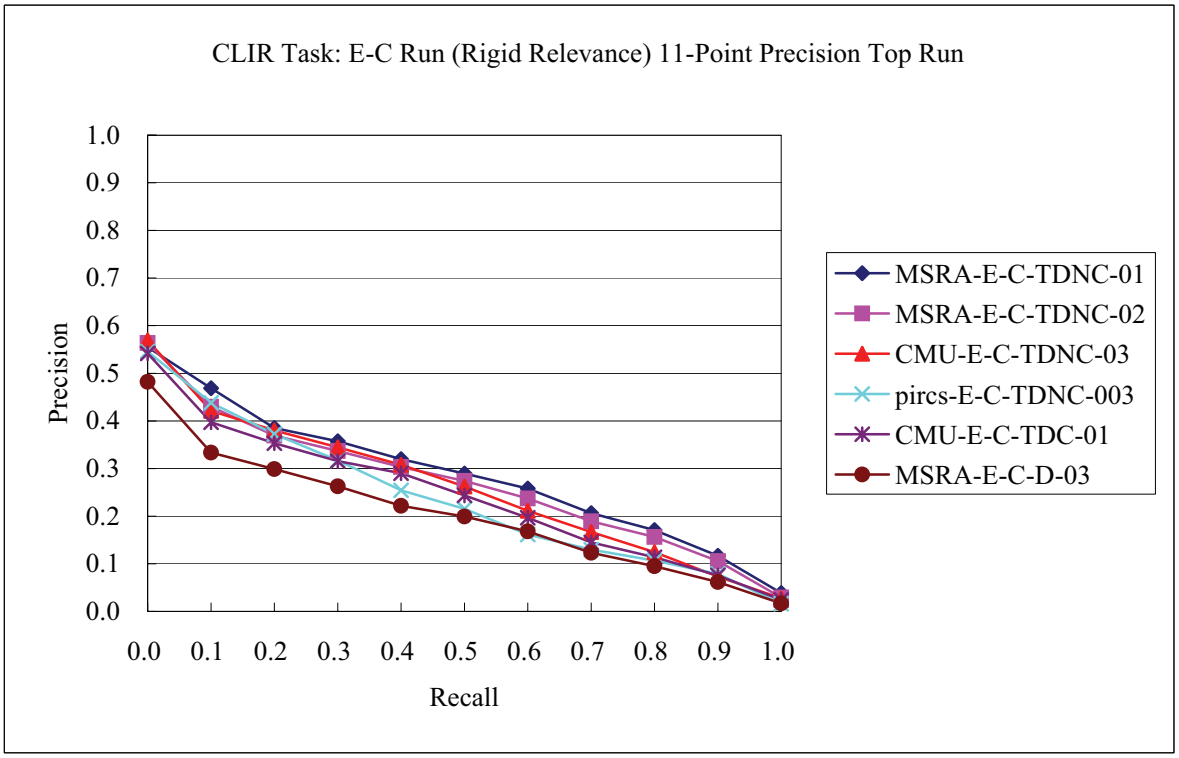

Figure 22. CLIR Task: E-C Run 11-Point Precision (Rigid Relevance)

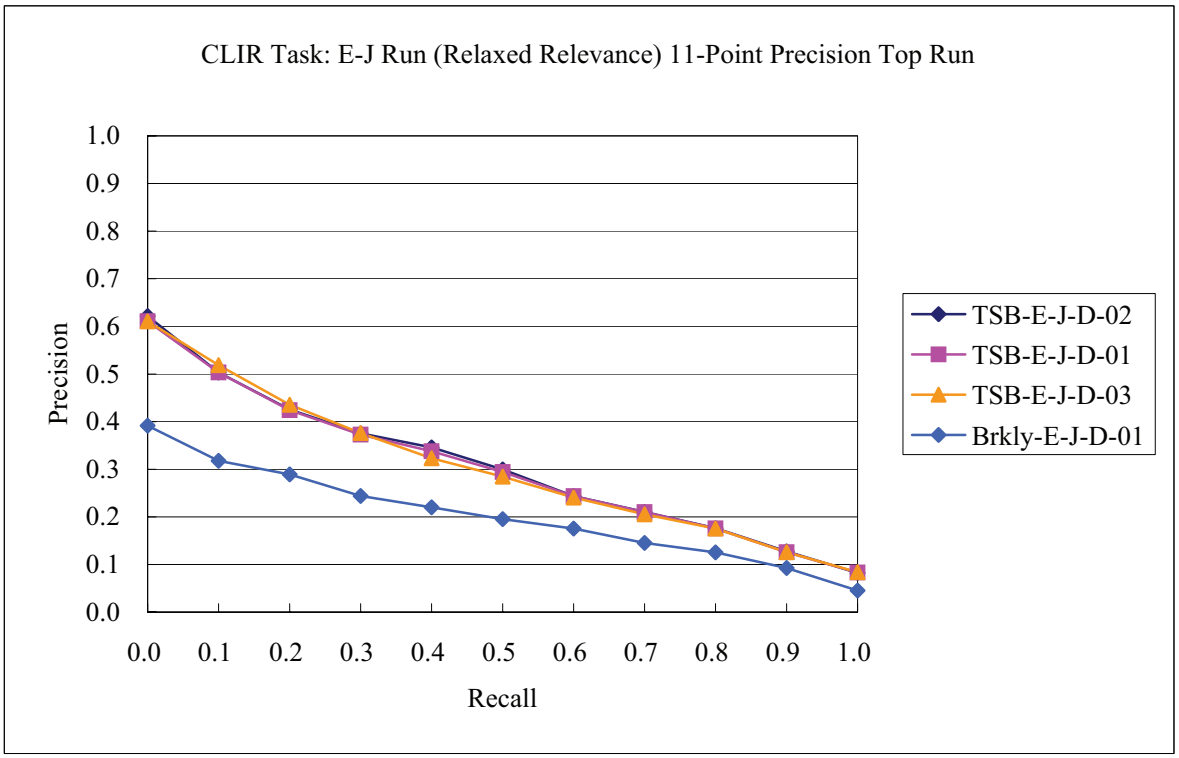

Figure 23. CLIR Task: E-J Run 11-Point Precision (Relaxed Relevance)

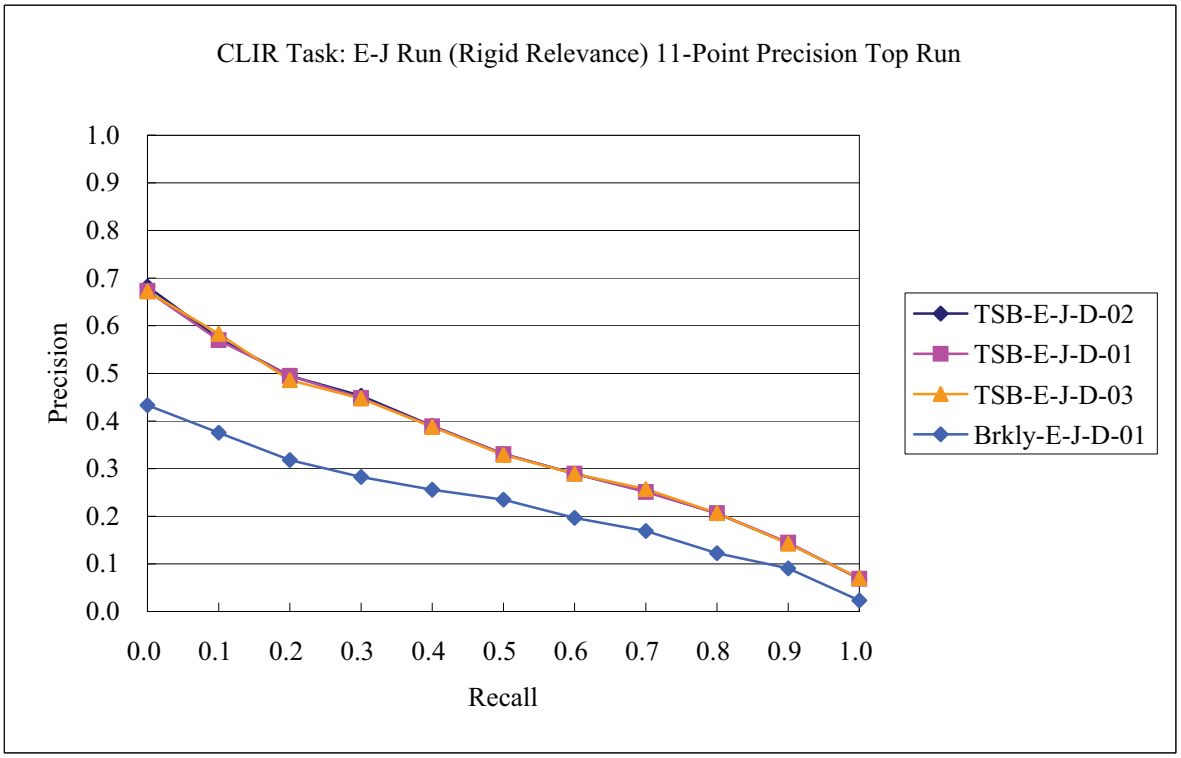

Figure 24. CLIR Task: E-J Run 11-Point Precision (Rigid Relevance)

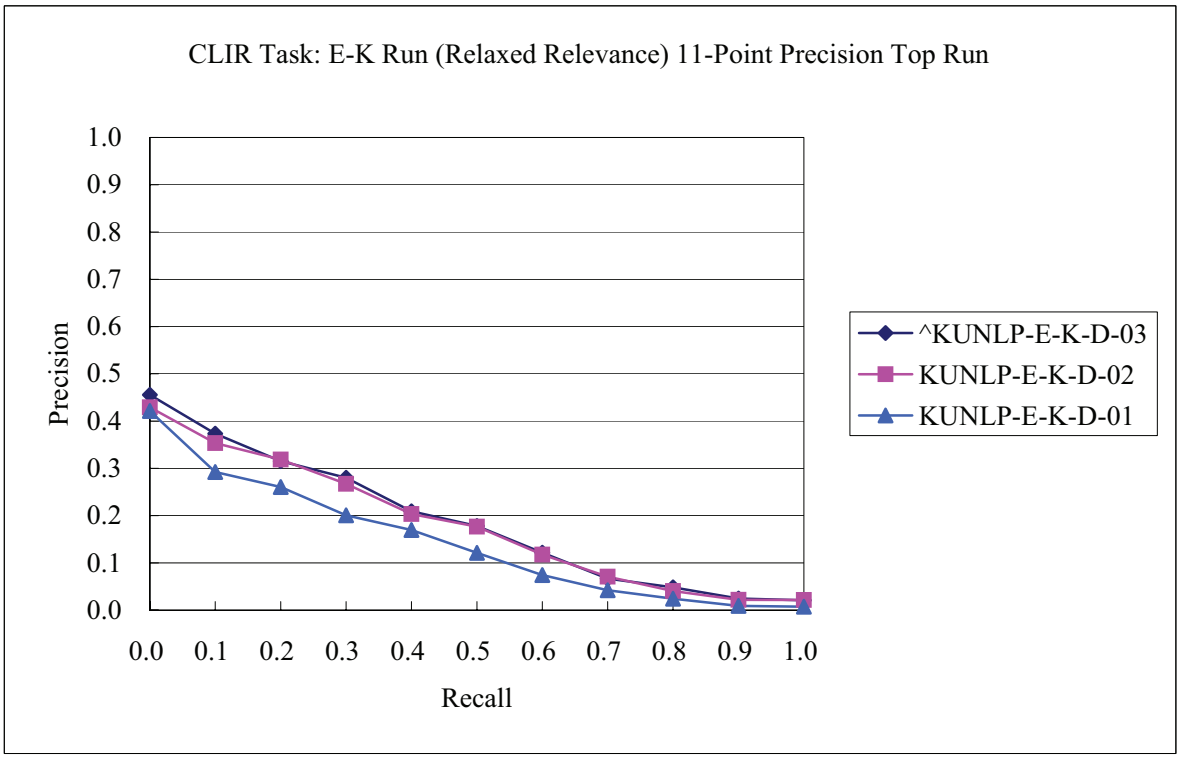

Figure 25. CLIR Task: E-K Run 11-Point Precision (Relaxed Relevance)

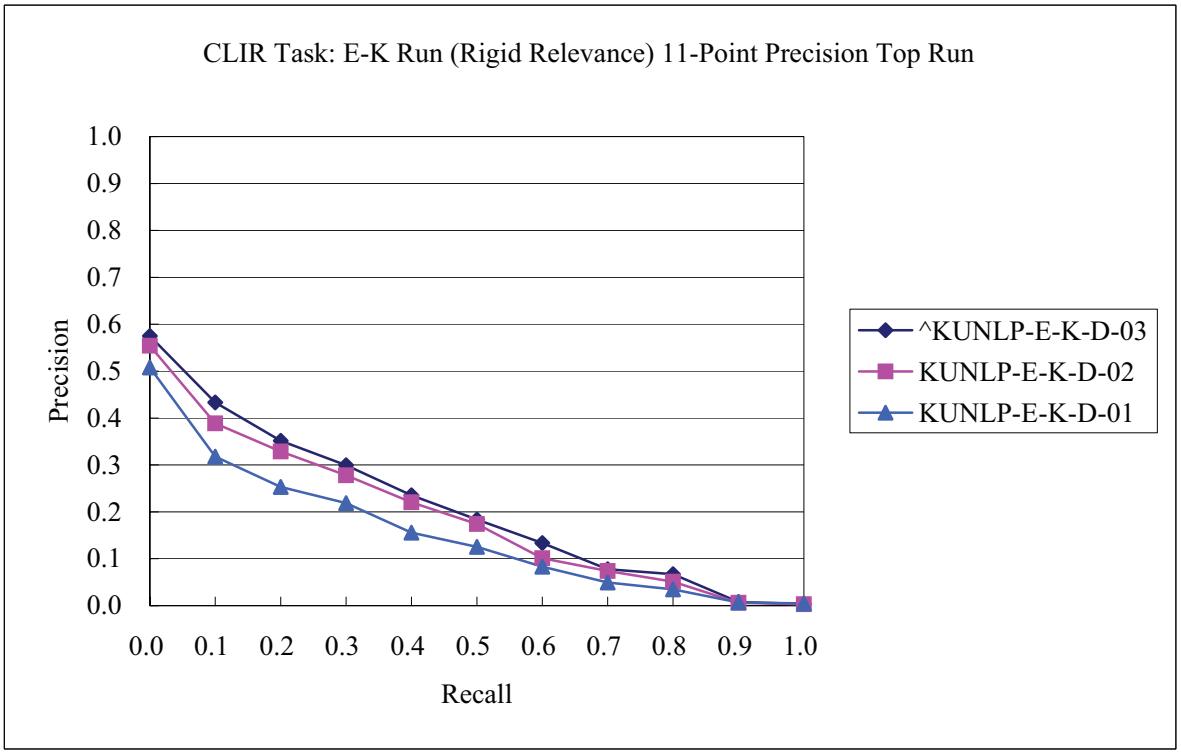

Figure 26. CLIR Task: E-K Run 11-Point Precision (Rigid Relevance)

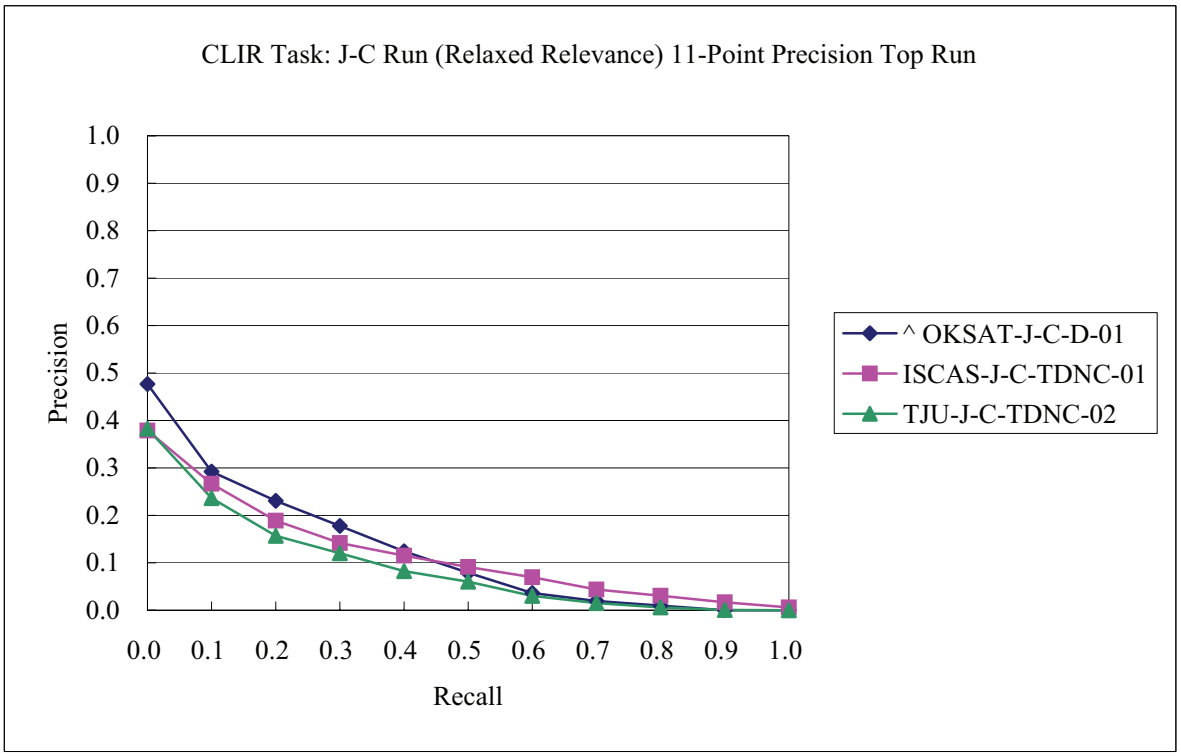

Figure 27. CLIR Task: J-C Run 11-Point Precision (Relaxed Relevance)

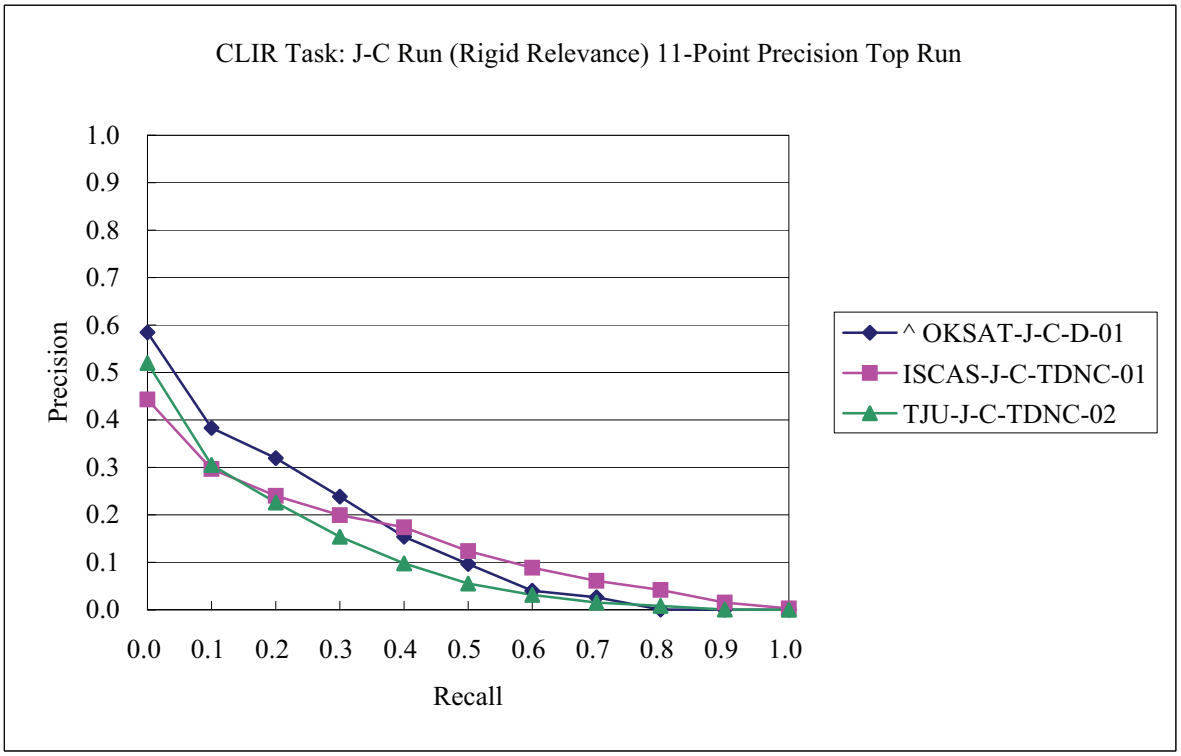

Figure 28. CLIR Task: J-C Run 11-Point Precision (Rigid Relevance)

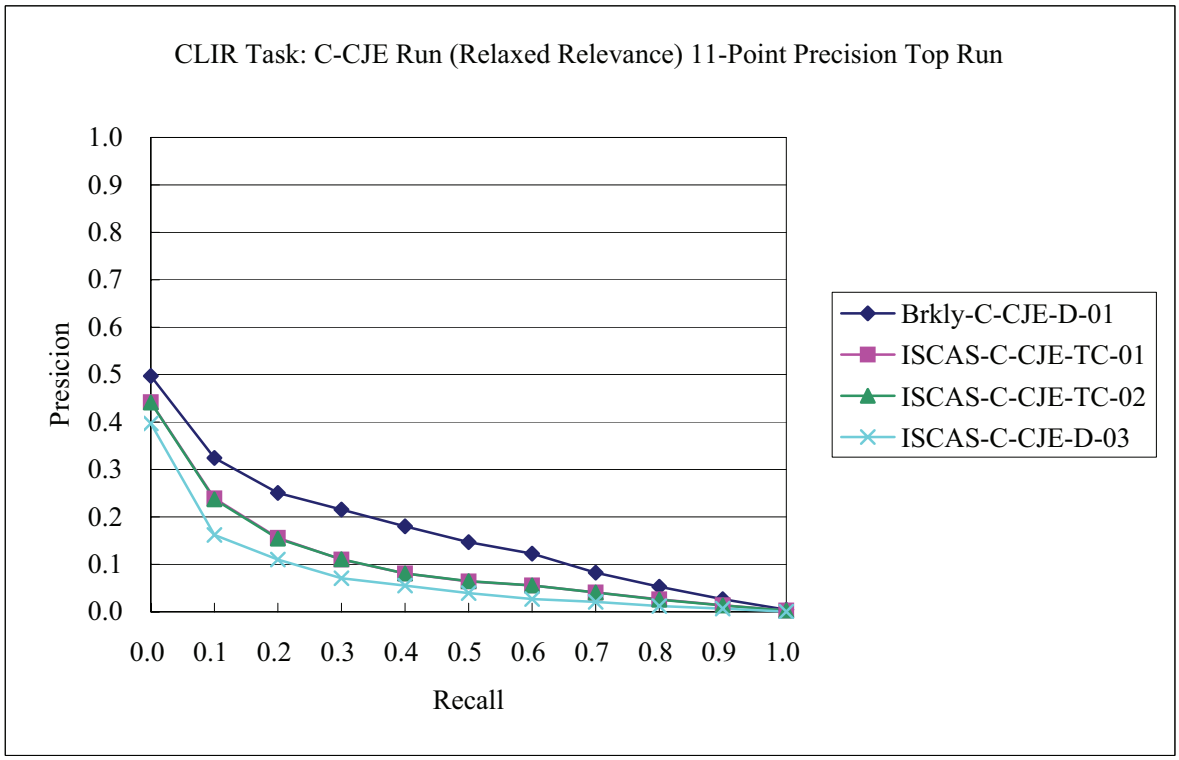

Figure 29. CLIR Task: C-CJE Run 11-Point Precision (Relaxed Relevance)

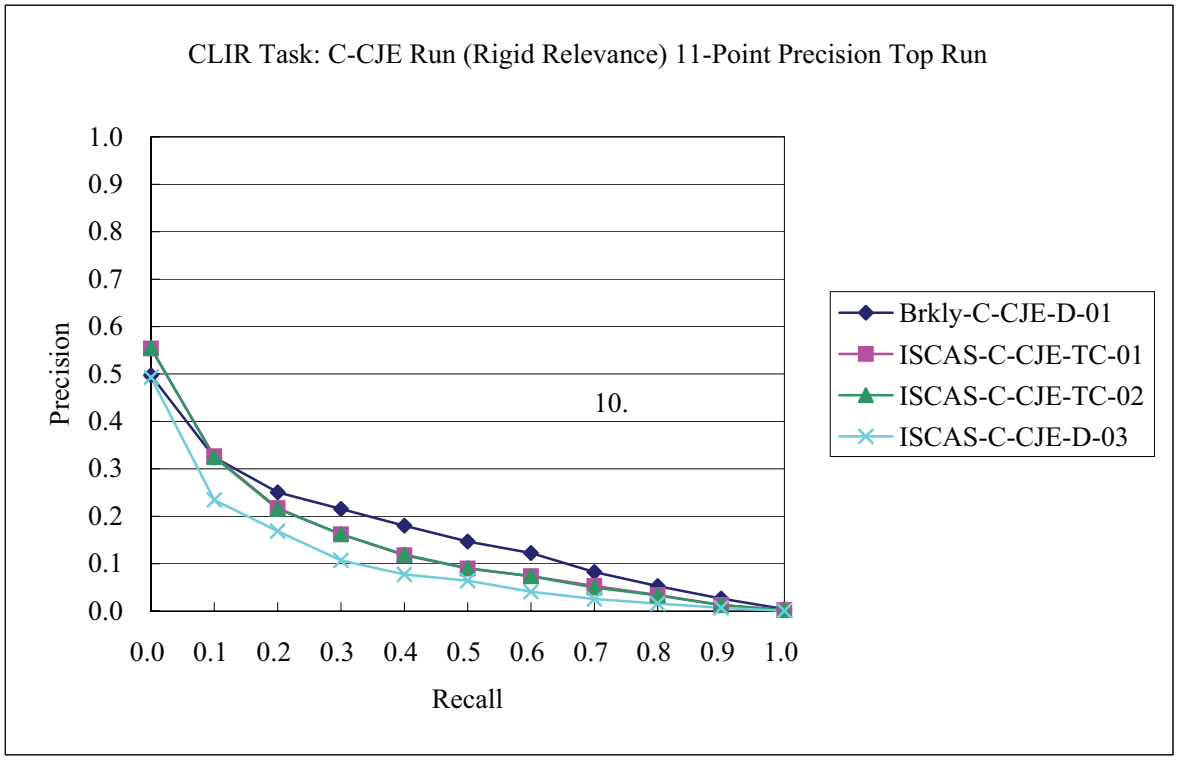

Figure 30. CLIR Task: C-CJE Run 11-Point Precision (Rigid Relevance)

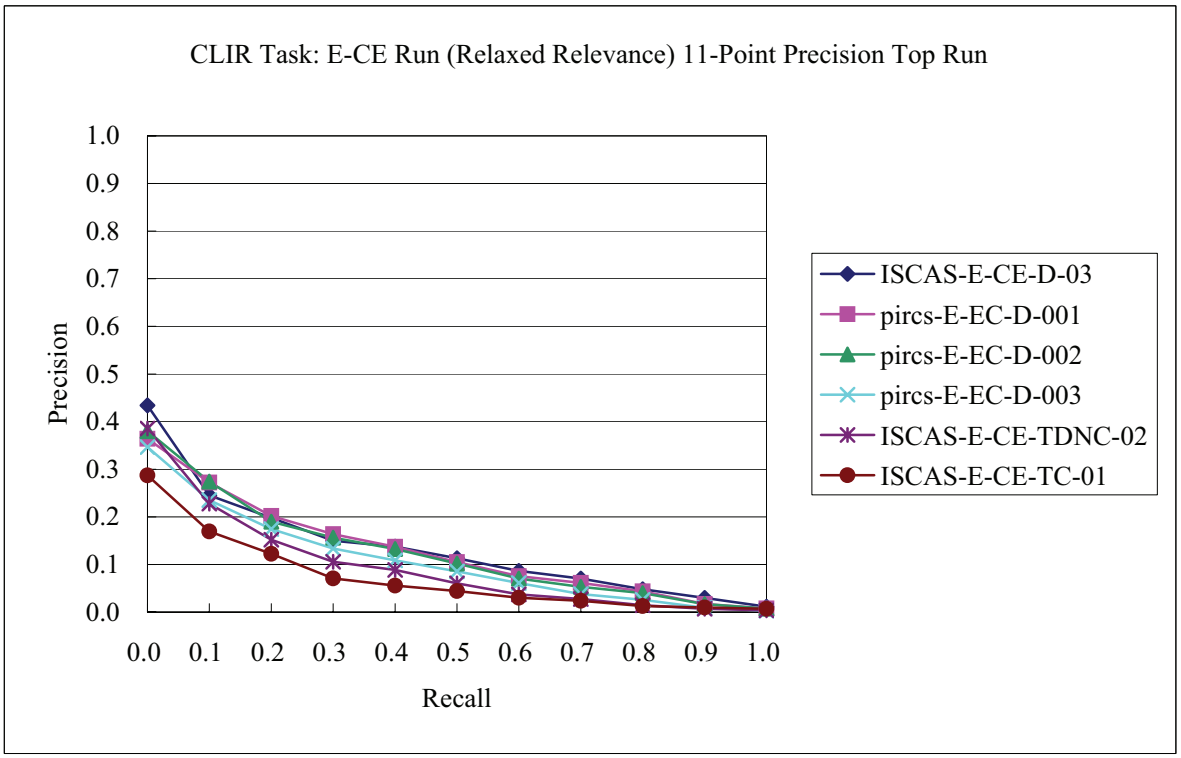

Figure 31. CLIR Task: E-CE Run 11-Point Precision (Relaxed Relevance)

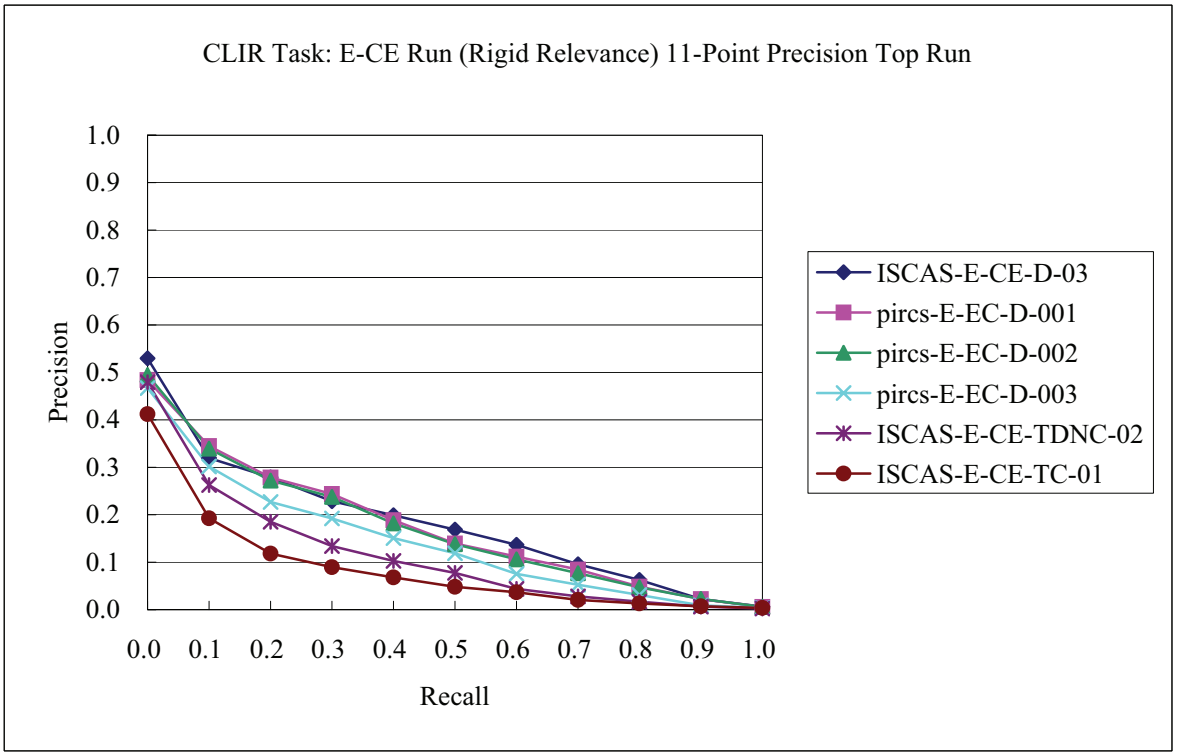

Figure 32. CLIR Task: E-CE Run 11-Point Precision (Rigid Relevance)

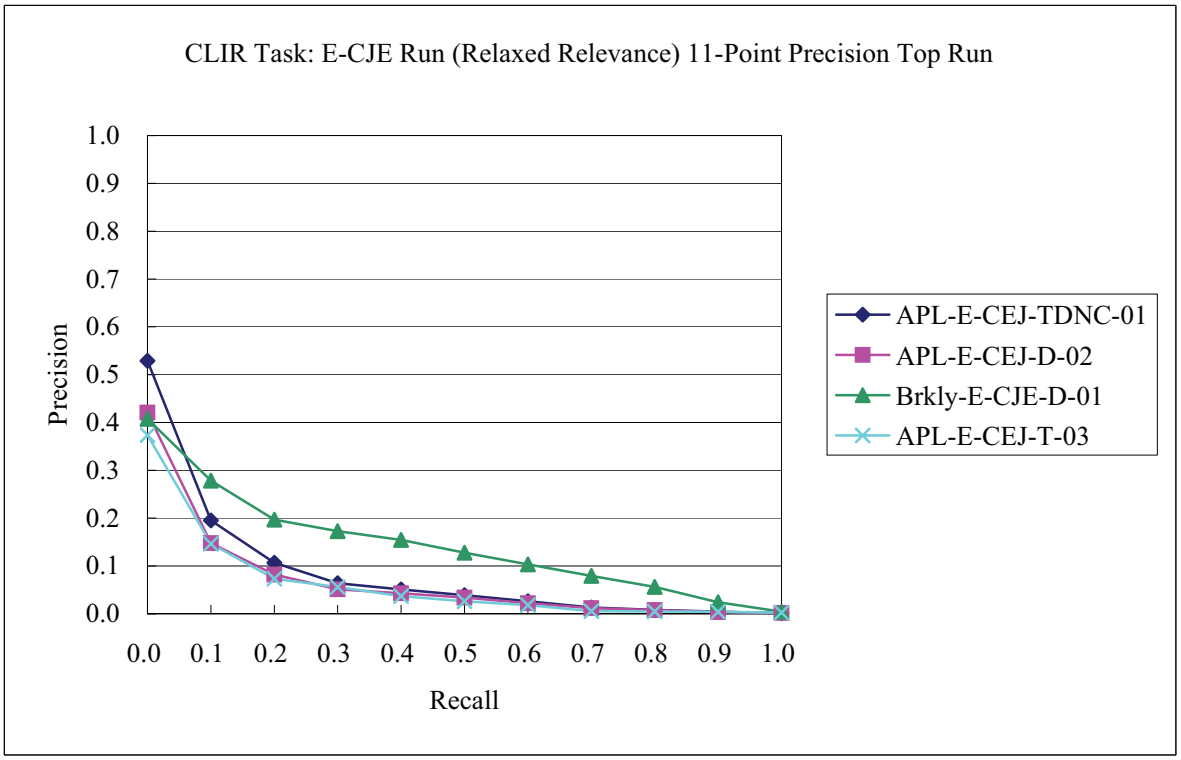

Figure 33. CLIR Task: E-CJE Run 11-Point Precision (Relaxed Relevance)

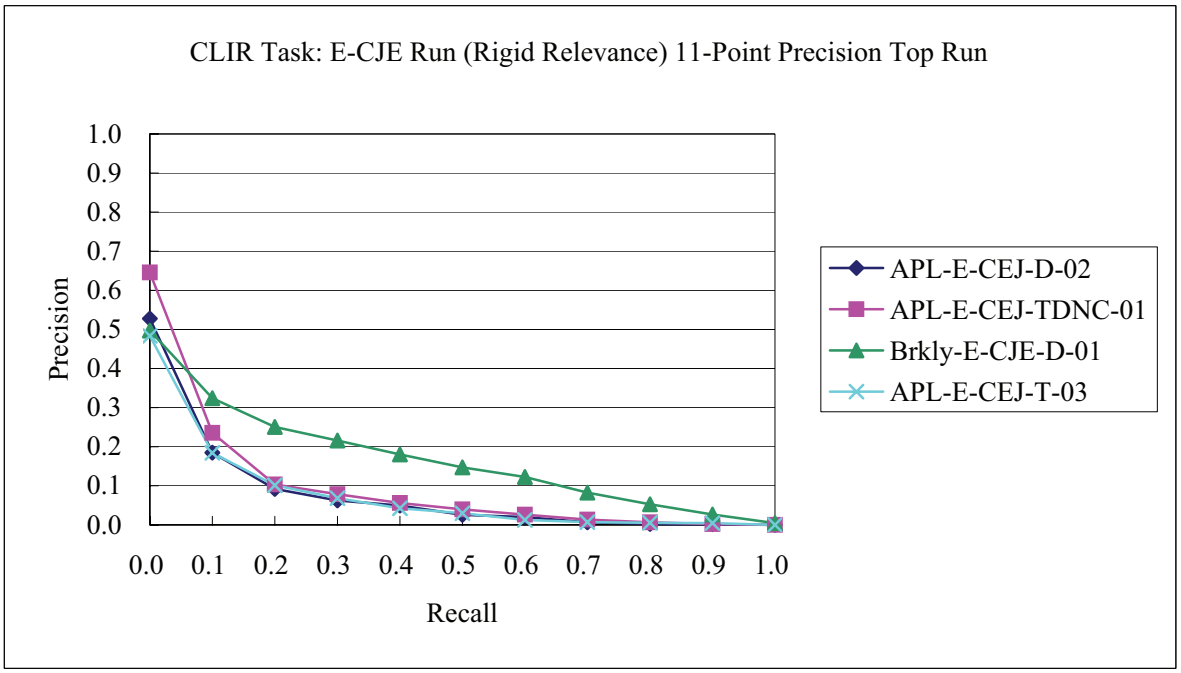

Figure 34. CLIR Task: E-CJE Run 11-Point Precision (Rigid Relevance)

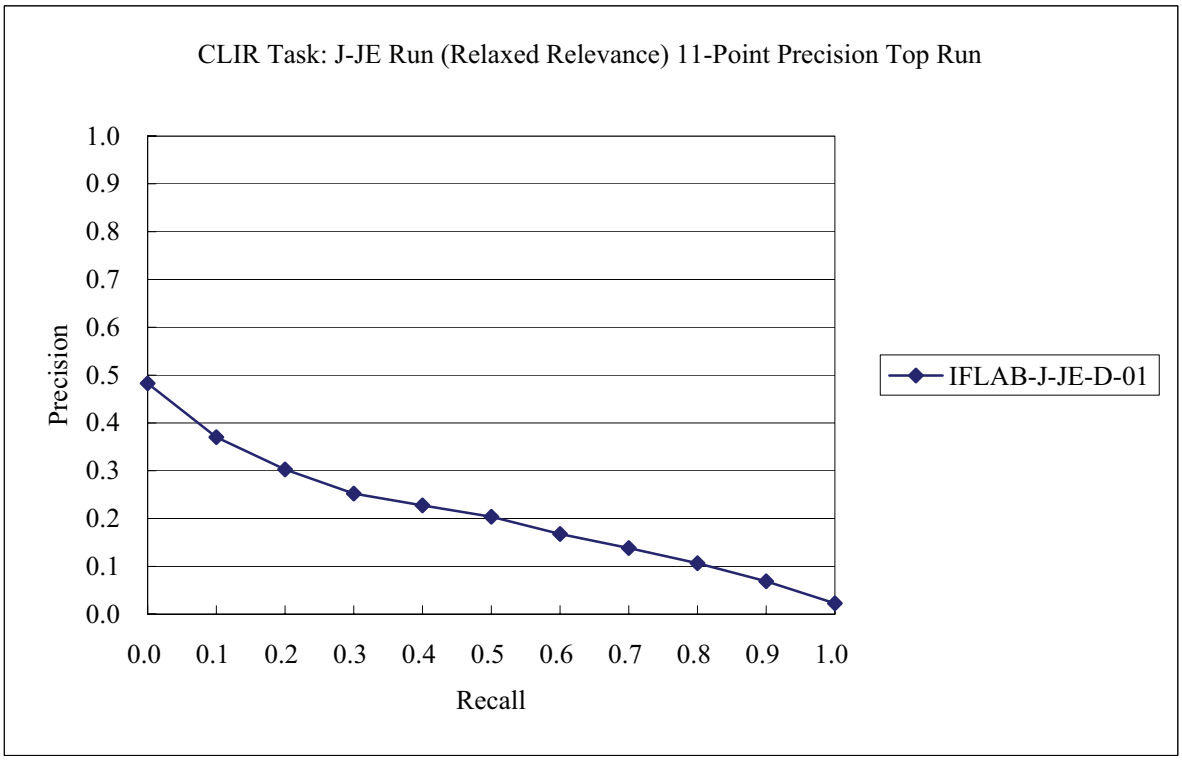

Figure 35. CLIR Task: J-JE Run 11-Point Precision (Relaxed Relevance)

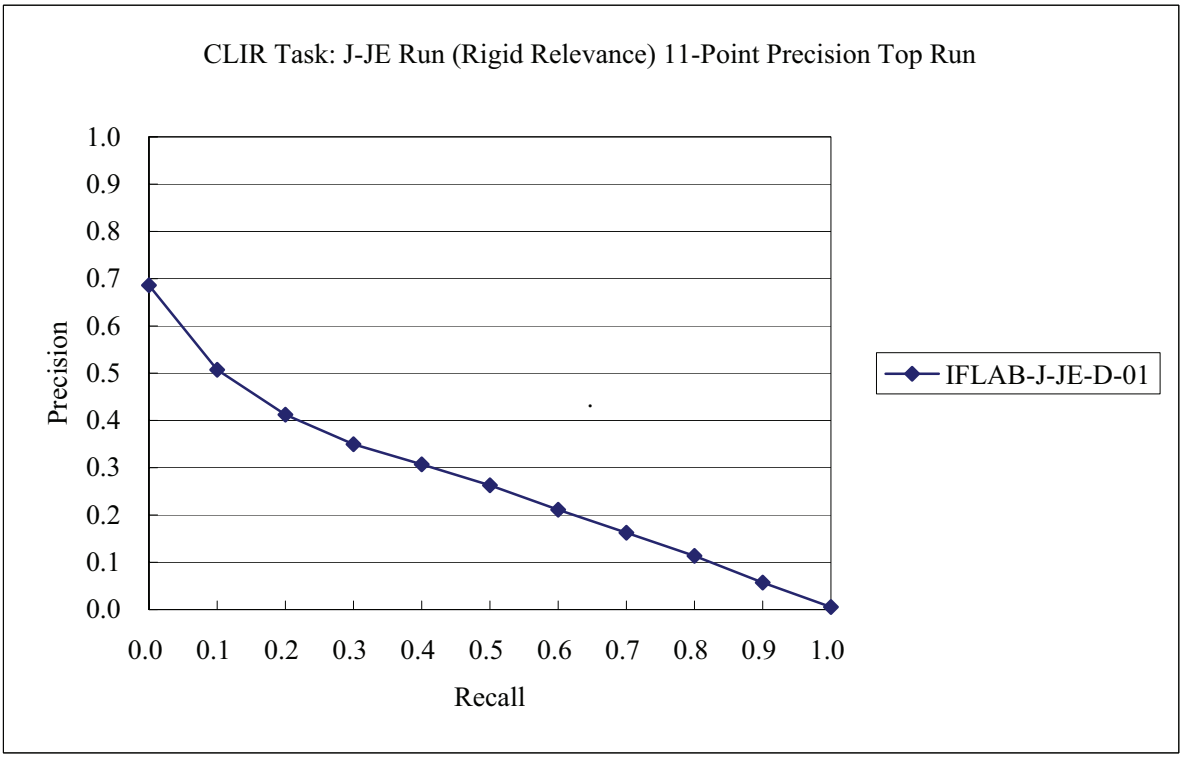

Figure 36. CLIR Task: J-JE Run 11-Point Precision (Rigid Relevance)

NTCIR Workshop 3
CLIR Task: Techniques Description

Run ID	Index Unit	Index Tech	Index Struc	Query UnIt	Query Method	IR Model	Ranking	Query Expan	Trans Tech	Training Corpus
APL-C-C-TDNC-01	character n-grams of lengths 1 and 2	no morphological operations	Inverted file and corresponding dual file	n-grams generated from input text	automatic	statistical language model	TF*IDF	*1	None	None
APL-C-C-D-02	character n-grams of lengths 1 and 2	no morphological operations	Inverted file and corresponding dual file	n-grams generated from input text	automatic	statistical language model	TF*IDF	*1	None	None
APL-E-C-TDNC-01	character n-grams of lengths 1 and 2	no morphological operations	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*2	Single term translations were obtained by the Babelfish MT translator	*5
APL-E-C-D-02	character n-grams of lengths 1 and 2	no morphological operations	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*2	Single term translations were obtained by the Babelfish MT translator	*5
APL-E-C-T-03	character n-grams of lengths 1 and 2	no morphological operations	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*2	Single term translations were obtained by the Babelfish MT translator	*5
APL-E-CEJ-TDNC-01	words (E); character n-grams of lengths 1, 2, and 3 (J), n-grams of lengths 1 and 2 (C)	no morphological operations	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*2	Single term translations were obtained by the Babelfish MT translator.	*6
APL-E-CEJ-D-02	words (E); character n-grams of lengths 1, 2, and 3 (J), n-grams of lengths 1 and 2 (C)	no morphological operations	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*2	Single term translations were obtained by the Babelfish MT translator.	*6
APL-E-CEJ-T-03	words (E); character n-grams of lengths 1, 2, and 3 (J), n-grams of lengths 1 and 2 (C)	no morphological operations	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*2	Single term translations were obtained by the Babelfish MT translator.	*6
APL-E-E-TDNC-01	words	no morphological operations; whitespace-delimited, lower-cased tokens	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*1	None	None

APL-E-E-D-02	words	no morphological operations; whitespace-delimited, lower-cased tokens	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*1	None	None
APL-E-E-T-03	words	no morphological operations; whitespace-delimited, lower-cased tokens	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*1	None	None
APL-E-J-TDNC-01	character n-grams of lengths 1, 2, and 3	no morphological operations	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*2	Single term translations were obtained by the Babelfish MT translator	*4
APL-E-J-D-02	character n-grams of lengths 1, 2, and 3	no morphological operations	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*2	Single term translations were obtained by the Babelfish MT translator	*4
APL-E-J-T-03	character n-grams of lengths 1, 2, and 3	no morphological operations	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*2	Single term translations were obtained by the Babelfish MT translator	*4
APL-E-K-TDNC-01	character n-grams of lengths 1, 2, and 3	no morphological operations	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*2	Single term translations were obtained by the Babelfish MT translator	*3
APL-E-K-D-02	character n-grams of lengths 1, 2, and 3	no morphological operations	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*2	Single term translations were obtained by the Babelfish MT translator	*3
APL-E-K-T-03	character n-grams of lengths 1, 2, and 3	no morphological operations	Inverted file and corresponding dual file	words	automatic	statistical language model	TF*IDF	*2	Single term translations were obtained by the Babelfish MT translator	*3
APL-J-J-TDNC-01	character n-grams of lengths 1, 2, and 3	no morphological operations	Inverted file and corresponding dual file	n-grams generated from input text	automatic	statistical language model	TF*IDF	*1	None	None
APL-J-J-D-02	character n-grams of lengths 1, 2, and 3	no morphological operations	Inverted file and corresponding dual file	n-grams generated from input text	automatic	statistical language model	TF*IDF	*1	None	None
APL-J-J-T-03	character n-grams of lengths 1, 2, and 3	no morphological operations	Inverted file and corresponding dual file	n-grams generated from input text	automatic	statistical language model	TF*IDF	*1	None	None
APL-K-K-TDNC-01	character n-grams of lengths 1, 2, and 3	no morphological operations	Inverted file and corresponding dual file	n-grams generated from input text	automatic	statistical language model	TF*IDF	*1	None	None
APL-K-K-D-02	character n-grams of lengths 1, 2, and 3	no morphological operations	Inverted file and corresponding dual file	n-grams generated from input text	automatic	statistical language model	TF*IDF	*1	None	None

APL-K-K-T-03	character n-grams of lengths 1, 2, and 3	no morphological operations	Inverted file and corresponding dual file	n-grams generated from input text	automatic	statistical language model	TF*IDF	*1	None	None
Brkly-C-C-D-01	character + bi-character	stopwords removed	inverted file	character + bi-character	automatic	logistic regression	*7	50 terms selected from 20 top-ranked documents after the initial retrieval.	None	None
Brkly-C-CJE-D-01		Direct merging from Brkly-C-C-D-01, Brkly-C-J-D-01, and Brkly-C-E-D-01.								
Brkly-C-J-D-01	character + bi-character	Hiragana discarded	inverted file	character + bi-character	automatic	logistic regression	*7	post-translation, 50 terms selected from 20 top-ranked documents after the initial retrieval.	Corpus-based for translating Chinese to English; MT-based for translating English to Japanese.	Hong Kong News, FBIS corpus.
Brkly-E-C-D-01	character + bi-character	stopwords removed	inverted file	character + bi-character	automatic	logistic regression	*7	50 terms selected from 20 top-ranked documents after the initial retrieval.	MT. Word not translated by MT were looked up in a bilingual dictionary created from parallel corpora.	Hong Kong News, FBIS corpus.
Brkly-E-CJE-D-01		Direct merging from Brkly-E-C-D-01, Brkly-E-J-D-01, and Brkly-E-E-D-04.								
Brkly-E-E-C-01	words	stopwords removed, content-words stemmed.	inverted file	words	automatic	logistic regression	*7	30 terms selected from 20 top-ranked documents after the initial retrieval.	None	None
Brkly-E-E-TDN-02	words	stopwords removed, content-words stemmed.	inverted file	words	automatic	logistic regression	*7	30 terms selected from 20 top-ranked documents after the initial retrieval.	None	None
Brkly-E-E-D-03	words	stopwords removed, content-words stemmed.	inverted file	words	automatic	logistic regression	*7	30 terms selected from 20 top-ranked documents after the initial retrieval.	None	None
Brkly-E-J-D-01	character + bi-character	Hiragana discarded	inverted file	character + bi-character	automatic	logistic regression	*7	50 terms selected from 20 top-ranked documents after the initial retrieval.	MT.	None
Brkly-J-J-D-01	character + bi-character	Hiragana discarded	inverted file	character + bi-character	automatic	logistic regression	*7	50 terms selected from 20 top-ranked documents after the initial retrieval.	None	None
Brkly-K-K-D-01	character + bi-character	stopwords removed	inverted file	character + bi-character	automatic	logistic regression	*7	50 terms selected from 20 top-ranked documents after the initial retrieval.	None	None
*CECIR-C-C-TDNC-01	bigram	None	inverted index file	bigram	automatic	vector space	2-poisson	None	None	None
*CECIR-E-E-TDNC-01	word	stemming and reducing stop words	inverted index file	word	automatic	vector space	2-poisson	None	None	None

*CECIR-K-CC-TDNC-01	bigram	None	inverted index file	bigram	automatic	vector space	2-poisson	None	Dictionary based translation with manual disambiguation	None
*CECIR-K-CC-TDNC-02	bigram	None	inverted index file	bigram	automatic	vector space	2-poisson	None	Dictionary based translation with disambiguation by using semantic hierarchy	None
*CECIR-K-EE-TDNC-06	word	stemming and stop word elimination	inverted index file	words with stemming and stop word elimination	automatic	vector space	2-poisson	None	Dictionary based translation with manual disambiguation	None
*CECIR-K-EE-TDNC-07	word	stemming and stop word elimination	inverted index file	words with stemming and stop word elimination	automatic	vector space	2-poisson	None	Dictionary based translation with automatic disambiguation and weighting based on statistical information	None
CMU-C-C-TDC-01	phrase	segmentation	inverted file	word	automatic	VSM	tf+idf	PRF	None	None
CMU-C-C-TDNC-02	phrase	segmentation	inverted file	word	automatic	VSM	tf+idf	PRF	None	None
CMU-C-C-D-03	phrase	segmentation	inverted file	word	automatic	VSM	tf+idf	PRF	None	None
CMU-E-C-TDC-01	phrase	segmentation	inverted file	word	automatic	VSM	tf+idf	PRF	Machine translation, dictionary	None
CMU-E-C-TDC-02	phrase	segmentation	inverted file	word	automatic	VSM	tf+idf	PRF	Machine translation, dictionary	None
CMU-E-C-TDNC-03	phrase	segmentation	inverted file	word	automatic	VSM	tf+idf	PRF	Machine translation, dictionary	None
CMU-E-E-TDC-01	word	Stop words, stemming	inverted file	word	automatic	VSM	tf+idf	PRF	None	None
CRL-C-C-TDNC-01	all the patterns of strings	morphology	suffix array	word	automatic	okapi model (probabilistic model)	tf/idf, doc. Leng., term position	automatic feedback	None	IREX, NTCIR1, NTCIR2
CRL-C-C-D-02	all the patterns of strings	morphology	suffix array	word	automatic	okapi model (probabilistic model)	tf/idf, doc. Leng., term position	automatic feedback	None	IREX, NTCIR1, NTCIR2
CRL-C-C-TC-03	all the patterns of strings	morphology	suffix array	word	automatic	okapi model (probabilistic model)	tf/idf, doc. Leng., term position	automatic feedback	None	IREX, NTCIR1, NTCIR2
CRL-E-E-TDNC-01	all the words	morphology	suffix array	word	automatic	okapi model (probabilistic model)	tf/idf, doc. Leng., term position	automatic feedback	None	IREX, NTCIR1, NTCIR2
CRL-J-J-TDNC-01	all the patterns of strings	morphology	suffix array	word	automatic	okapi model (probabilistic model)	tf/idf, doc. Leng., term position	automatic feedback	None	IREX, NTCIR1, NTCIR2
CRL-J-J-TDNC-02	all the patterns of strings	morphology	suffix array	word	automatic	okapi model (probabilistic model)	tf/idf, doc. Leng.	automatic feedback	None	IREX, NTCIR1, NTCIR2
CRL-J-J-D-03	all the patterns of strings	morphology	suffix array	word	automatic	okapi model (probabilistic model)	tf/idf, doc. Leng., term position	automatic feedback	None	IREX, NTCIR1, NTCIR2
CRL-K-K-TDNC-01	all the patterns of strings	morphology	suffix array	word	automatic	okapi model (probabilistic model)	tf/idf, doc. Leng., term position	automatic feedback	None	IREX, NTCIR1, NTCIR2

CRL-K-K-TDNC-02	all the patterns of strings	morphology	suffix array	word	automatic	okapi model (probabilistic model)	tf/idf, doc. Leng.	automatic feedback	None	IREX, NTCIR1, NTCIR2
CRL-K-K-D-03	all the patterns of strings	morphology	suffix array	word	automatic	okapi model (probabilistic model)	tf/idf, doc. Leng., term position	automatic feedback	None	IREX, NTCIR1, NTCIR2
FJUIR-C-C-C-01	*8	*9	inverted file.	*8	automatic	vector space model	tf/idf, word leng., doc. Leng.	None	None	None
FJUIR-C-C-D-02	*8	*9	inverted file.	*8	automatic	vector space model	t tf/idf, word leng., doc. Leng.	None	None	None
FJUIR-J-J-C-01	*8	*9	inverted file.	*8	automatic	vector space model	t tf/idf, word leng., doc. Leng.	None	None	None
FJUIR-J-J-D-02	*8	*9	inverted file.	*8	automatic	vector space model	t tf/idf, word leng., doc. Leng.	None	None	None
FJUIR-K-K-C-01	*8	*9	inverted file.	*8	automatic	vector space model	t tf/idf, word leng., doc. Leng.	None	None	None
FJUIR-K-K-D-02	*8	*9	inverted file.	*8	automatic	vector space model	t tf/idf, word leng., doc. Leng.	None	None	None
HKPU-C-CC-TDN-01	Short Word + Bigram	Hybrid	Extensible Inverted File	Short Word + Bigram	automatic	Probabilistic	2-Poisson	None	None	None
HKPU-C-CC-C-02	Short Word + Bigram	Hybrid	Extensible Inverted File	Short Word + Bigram	automatic	Probabilistic	2-Poisson	None	None	None
HKPU-C-CC-CT-03	Short Word + Bigram	Hybrid	Extensible Inverted File	Short Word + Bigram	automatic	Probabilistic	2-Poisson	None	None	None
HUM-C-C-D-01	*10	case normalization of Latin, old Kanji normalized to new	inverted file	overlapping n-grams for CJK, words for Latin	automatic	vector space	tf/idf, tf dampened, idf squared, document length importance 50%	None	None	None
HUM-C-C-TC-02	*10	case normalization of Latin, old Kanji normalized to new	inverted file	overlapping n-grams for CJK, words for Latin	automatic	vector space	tf/idf, tf dampened, document length importance 0%	None	None	None
HUM-C-C-TDNC-03	*10	case normalization of Latin, old Kanji normalized to new	inverted file	overlapping n-grams for CJK, words for Latin	automatic	vector space	tf/idf, tf dampened, document length importance 100%	None	None	None
HUM-E-E-D-01	words	case normalization of Latin	inverted file	words	automatic	vector space	tf/idf, tf dampened, idf squared, document length importance 50%	None	None	None
HUM-E-E-TC-02	words	case normalization of Latin	inverted file	words	automatic	vector space	tf/idf, tf dampened, document length importance 0%	None	None	None
HUM-E-E-TDNC-03	words	case normalization of Latin	inverted file	words	automatic	vector space	tf/idf, tf dampened, document length importance 100%	None	None	None

HUM-J-J-D-01	*10	case normalization of Hiragana, Katakana and Latin, old Kanji normalized to new	inverted file	overlapping n-grams for CJK, words for Latin	automatic	vector space	tf/idf, tf dampened, idf squared, document length importance 50%	None	None	None
HUM-J-J-TC-02	*10	case normalization of Latin, old Kanji normalized to new	inverted file	overlapping n-grams for CJK, words for Latin	automatic	vector space	tf/idf, tf dampened, document length importance 0%	None	None	None
HUM-J-J-TDNC-03	*10	case normalization of Latin, old Kanji normalized to new	inverted file	overlapping n-grams for CJK, words for Latin	automatic	vector space	tf/idf, tf dampened, document length importance 100%	None	None	None
HUM-K-K-D-01	*10	Han mapped to Hangul, case normalization of Latin, old Kanji normalized to new	inverted file	overlapping n-grams for CJK, words for Latin	automatic	vector space	tf/idf, tf dampened, idf squared, document length importance 50%	None	None	None
HUM-K-K-TC-02	*10	Han mapped to Hangul, case normalization of Latin, old Kanji normalized to new	inverted file	overlapping n-grams for CJK, words for Latin	automatic	vector space	tf/idf, tf dampened, document length importance 0%	None	None	None
HUM-K-K-TDNC-03	*10	Han mapped to Hangul, case normalization of Latin, old Kanji normalized to new	inverted file	overlapping n-grams for CJK, words for Latin	automatic	vector space	tf/idf, tf dampened, document length importance 100%	None	None	None
IFLAB-E-E-D-01	word	morphology, stemming, POS	inverted file	word	automatic	probabilistic model	the Okapi method	None	None	None
IFLAB-J-E-D-01	word	morphology, stemming, POS	inverted file	word	automatic	probabilistic model	the Okapi method	None	dictionary/corpus-based, select the top translation	the target documents were used to produce language models
IFLAB-J-J-D-01	word	morphology, stemming, POS	inverted file	word	automatic	probabilistic model	the Okapi method	None	None	None
IFLAB-J-JE-D-01	word	morphology, stemming, POS	inverted file	word	automatic	probabilistic model	the Okapi method	None	dictionary/corpus-based, select the top translation	the target documents were used to produce language models
ISCAS-C-C-TC-01	2-char	2-char	Inverted File	2-char	automatic	VSM	Tf/idf	None	None	None
ISCAS-C-C-TDNC-02	2-char	2-char	Inverted File	2-char	automatic	VSM	Tf/idf	None	None	None
ISCAS-C-C-D-03	2-char	2-char	Inverted File	2-char	automatic	VSM	Tf/idf	None	None	None
ISCAS-C-CE-TC-01	C/2-char E/word	E/ stemming	Inverted File	C/2-char E/word	automatic	VSM	Tf/idf	Yes	Dictionary+MT	Pre-translation query expansion
ISCAS-C-CE-TDNC-02	C/2-char E/word	E/ stemming	Inverted File	C/2-char E/word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-CE-D-03	C/2-char E/word	E/ stemming	Inverted File	C/2-char E/word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-CJE-TC-01	C/2-char J/word E/word	JE/ stemming	Inverted File	C/2-char J/word E/word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-CJE-TC-02	C/2-char J/word E/word	JE/ stemming	Inverted File	C/2-char J/word E/word	automatic	VSM	Tf/idf	None	Dictionary+MT	None

ISCAS-C-CJE-D-03	C/2-char JE/word	JE/ stemming	Inverted File	C/2-char JE/word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-CJ-TC-01	C/2-char J/word	J/ stemming	Inverted File	C/2-char J/word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-CJ-TDNC-02	C/2-char J/word	J/ stemming	Inverted File	C/2-char J/word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-CJ-D-03	C/2-char J/word	J/ stemming	Inverted File	C/2-char J/word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-E-TC-01	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-E-TDNC-02	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-E-D-03	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-JE-TC-01	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-JE-TDNC-02	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-JE-D-03	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-J-TC-01	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-J-TDNC-02	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-C-J-D-03	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-E-CE-TC-01	C/2-char E/word	E/ stemming	Inverted File	C/2-cha E/word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-E-CE-TDNC-02	C/2-char E/word	E/ stemming	Inverted File	C/2-char E/word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-E-CE-D-03	C/2-char E/word	E/ stemming	Inverted File	C/2-char E/word	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-E-C-TC-01	2-char	2-char	Inverted File	2-char	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-E-C-TDNC-02	2-char	2-char	Inverted File	2-char	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-E-C-D-03	2-char	2-char	Inverted File	2-char	automatic	VSM	Tf/idf	None	Dictionary+MT	None
ISCAS-E-E-TC-01	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	None	None
ISCAS-E-E-TDNC-02	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	None	None
ISCAS-E-E-D-03	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	None	None
ISCAS-J-C-TDNC-01	2-char	2-char	Inverted File	2-char	automatic	VSM	Tf/idf	None	MT	None
ISCAS-J-C-D-02	2-char	2-char	Inverted File	2-char	automatic	VSM	Tf/idf	None	MT	None
ISCAS-J-J-TC-01	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	None	None
ISCAS-J-J-D-02	Word	Stemming	Inverted File	Word	automatic	VSM	Tf/idf	None	None	None
KRDLNLU-C-C-D-01	Word and term	term extraction	inverted file	word and term	automatic	VSM	Tf/idf	None	None	None
KUNLP-E-K-D-01	word	POS (Nouns)	Inverted File	word	automatic	Okapi-like Probabilistic Model	tf/idf, document length	None	dictionary based-query translation which selects top 2 Korean words of each English word	the Financial Times, 1991-1994 at TREC CD4 is used to pre-query expansion :Korea Economic Daily (1994) is used to construct translation model.

KUNLP-E-K-D-02	word	POS (Nouns)	Inverted File	word	automatic	Okapi-like Probabilistic Model	tf/idf, document length	pre-query expansion using Okapi query expansion model	dictionary based-query translation which selects top 2 Korean words of each English word	the Financial Times, 1991-1994 at TREC CD4 is used to pre-query expansion :Korea Economic Daily (1994) is used to construct translation model.
^KUNLP-E-K-D-03	word	POS (Nouns)	Inverted File	word	interactive	Okapi-like Probabilistic Model	tf/idf, document length	pre-query expansion using Okapi query expansion model	dictionary based-query translation which selects top 2 Korean words of each English word.	The Financial Times, 1991-1994 at TREC CD4 is used to pre-query expansion. Korea Economic Daily (1994) is used to construct translation model.
KUNLP-K-K-D-01	word	POS (Nouns)	Inverted File	word	automatic	Okapi-like Probabilistic Model	tf/idf, document length	None	None	None
MSRA-E-C-TDNC-01	character, word	stopword list	inverted file	character, word	automatic	probabilistic +model	bm2500	pseudo relevance feedback	dictionary based, decaying co-occurrence model, syntactic dependence relations	CIRB010, CIRB020, Wall Street Journal(1987-1992), People's Daily(1980-1998)
MSRA-E-C-TDNC-02	character, word	stopword list	inverted file	character, word	automatic	probabilistic model	bm2500	pseudo relevance feedback	dictionary based, decaying co-occurrence model, syntactic dependence relations	CIRB010, CIRB020, Wall Street Journal(1987-1992), People's Daily(1980-1998)
MSRA-E-C-D-03	character, word	stopword list	inverted file	character, word	automatic	probabilistic model	bm2500	None	dictionary based, decaying co-occurrence model, syntactic dependence relations	CIRB010, CIRB020, Wall Street Journal(1987-1992), People's Daily(1980-1998)
MSRA-E-E-TDNC-01	Word	using stoplist and stemming	inverted file	word	automatic	probabilistic model	bm2500	no query expansion	None	None
MSRA-E-E-TDNC-02	Word	using stoplist and stemming	inverted file	word	automatic	probabilistic model	bm2500	no query expansion	Although we didn't do query expansion, but we used the information we got from the first retrieval result	None
MSRA-E-E-TDNC-03	Word	using stoplist and stemming	inverted file	word	automatic	probabilistic model	bm2500	no query expansion	Using query expansion	None
*NTU-E-E-D-01	word	None	inverted file	word	automatic	Vector space model	tf*idf	None	None	TREC6 text collection
*NTU-J-CJE-D-01	C: haracter bigram, J, E: word	J: morphology	inverted file	C: haracter bigram, J, E: word	automatic	Vector space model	tf*idf	None	*11	ASBC corpus, TREC6 text collection
*NTU-J-CJE-D-02	C: haracter bigram, J, E: word	J: morphology	inverted file	C: haracter bigram, J, E: word	automatic	Vector space model	tf*idf	None	*11	ASBC corpus, TREC6 text collection

*NTU-J-CJE-D-03	C: haracter bigram, J, E: word	J: morphology	inverted file	C: haracter bigram, J, E: word	automatic	Vector space model	tf*idf	None	*11	ASBC corpus, TREC6 text collection
!OASIS-E-E-D-01	word	stop word were discarded.	Inverted index	word	automatic	vector space model	tf/idf	None	None	ntc-e02-mai98.txt
!OASIS-E-E-D-02	word	stop word were discarded	Inverted index	word	automatic	vector space model	tf/idf	*12	None	ntc-e02-mai98.txt
!OASIS-E-E-D-03	word	stop word were discarded	Inverted index	word	automatic	vector space model	tf/idf	None	None	ntc-e02-mai99.txt
#!OASIS-E-E-D-04	word	stop word were discarded	Inverted index	word	automatic	vector space model	tf/idf	*12	None	ntc-e02-mai99.txt
!OASIS-J-J-D-01	combination bi-word and phrases	*13	Inverted index	combination bi-word and phrases	automatic	vector space model	tf/idf	*12	None	ntc-j-mai-99.txt
!OASIS-J-J-D-02	combination bi-word and phrases	*13	Inverted index	combination bi-word and phrases	automatic	vector space model	tf/idf	*12	None	ntc-j-mai-98.txt
!OASIS-J-J-D-03	combination bi-word and phrases	*13	Inverted index	combination bi-word and phrases	automatic	vector space model	tf/idf	None	None	ntc-j-mai-98.txt
#!OASIS-J-J-D-04	combination bi-word and phrases	*13	Inverted index	combination bi-word and phrases	automatic	vector space model	tf/idf	None	None	ntc-j-mai-99.txt
^OKSAT-E-E-D-01	n-gram	n-gram	inverted -gram index in Tree.	word + phrase	interactive	probabilistic model	tf/idf, essential keywords	synonym	None	None
^OKSAT-J-C-D-01	n-gram	n-gram	inverted -gram index	word + phrase	interactive	probabilistic model	tf/idf	Expand query	Dictionary-based	Post-translation
OKSAT-J-J-D-01	n-gram	n-gram	inverted -gram index in Tree.	word + phrase	automatic	probabilistic model	tf/idf, essential keywords	synonym	None	None
pircs-C-C-D-001	2-gram + 1-gram	5 stopwrd ch;	invrt-file network	2-gram + 1-gram	automatic	probabilistic, activation-spreading	activation: tf, ictf, doc + qry length	Based on term freq, doc length	None	None
pircs-C-C-D-002	2-gram + 1-gram; short wd	5 stopwrd ch; dictionary	invrt-file network	2-gram + 1-gram; short wd	automatic	probabilistic, activation-spreading	activation: tf, ictf, doc + qry length + combine 2 retrievals	Based on term freq, doc length	None	None
pircs-C-C-TDNC-003	2-gram + 1-gram; short wd	5 stopwrd ch; dictionary	invrt-file network	2-gram + 1-gram; short wd	automatic	probabilistic, activation-spreading	activation: tf, ictf, doc + qry length + combine 2 retrievals	Based on term freq, doc length	None	None
pircs-E-C-D-001	short wrd + char	dictionary	invrt-file network	short wd + char	automatic	probabilistic, activation-spreading	activation: tf, ictf, doc + qry length	post trans expansion	biling. dict + MT software	None
pircs-E-C-D-002	short wrd + char	dictionary	invrt-file network	short wd + char	automatic	probabilistic, activation-spreading	activation: tf, ictf, doc + qry length	pre & post trans expan	biling. dict + MT software	None
pircs-E-C-TDNC-003	short wrd + char	dictionary	invrt-file network	short wd + char	automatic	probabilistic, activation-spreading	activation: tf, ictf, doc + qry length	post trans expansion	biling. dict + MT software	None
pircs-E-EC-D-001	stems + 2-wd phr; short wd	Porter stemming, stopwords; dictionary	invrt-file network	stems; short wd	automatic	probabilistic, activation-spreading	activation: tf, ictf, doc + qry length + English RSV directly vs. E-C-D-002	pre & post trans expan	biling. dict + MT software	None

pircs-E-EC-D-002	stems + 2-wd phr; short wd	Porter stemming, stopwords; dictionary	invrt-file network	stems; short wd	automatic	probabilistic, activation-spreading	English RSV adjusted vs. E-C-D-002	pre & post trans expan	biling. dict + MT software	None
pircs-E-EC-D-003	stems + 2-wd phr; short wd	Porter stemming, stopwords; dictionary	invrt-file network	stems; short wd	automatic	probabilistic, activation-spreading	English RSV adjusted vs. E-C-D-001	post trans expansion	biling. dict + MT software	None
pircs-E-E-TDNC-001	stems + 2-wd phr	Porter stemming, stopwords	invrt-file network	stems + 2-wd phr	automatic	probabilistic, activation-spreading	activation: tf, ictf, doc + qry length	Based on term freq, doc length	None	None
POSTECH-C-C-D-01	morpheme	POS tagging	inverted file	morpheme	automatic	probabilistic model	modified Okapi BM25	no query expansion	None	None
POSTECH-C-C-C-02	morpheme	POS tagging	inverted file	morpheme	automatic	probabilistic model	modified Okapi BM25	no query expansion	None	None
POSTECH-C-C-T-03	morpheme	POS tagging	inverted file	morpheme	automatic	probabilistic model	modified Okapi BM25	no query expansion	None	None
POSTECH-J-J-C-01	morpheme	POS tagging	inverted file	morpheme	automatic	probabilistic model	modified Okapi BM25	no query expansion	None	None
POSTECH-J-J-D-02	morpheme	POS tagging	inverted file	morpheme	automatic	probabilistic model	modified Okapi BM25	no query expansion	None	None
POSTECH-J-J-T-03	morpheme	POS tagging	inverted file	morpheme	automatic	probabilistic model	modified Okapi BM25	no query expansion	None	None
POSTECH-K-K-D-01	morpheme	POS tagging	inverted file	morpheme	automatic	probabilistic model	modified Okapi BM25	no query expansion	None	None
POSTECH-K-K-D-02	morpheme	POS tagging	inverted file	morpheme	automatic	probabilistic model	modified Okapi BM25	no query expansion	None	None
POSTECH-K-K-C-03	morpheme	POS tagging	inverted file	morpheme	automatic	probabilistic model	modified Okapi BM25	no query expansion	None	None
SSTUT-C-C-D-01	character	all	suffix array	bigram	automatic	probabilistic model	empirical term weighting	no query expansion	automatic extraction of translation word from parallel corpus	NTCIR1/NTCIR2 Corpus
SSTUT-E-J-D-01	character	all	suffix array	bigram	automatic	probabilistic model	empirical term weighting	no query expansion	automatic extraction of translation word from parallel corpus	NTCIR1/NTCIR2 Corpus
SSTUT-E-JE-D-01	character	all	suffix array	bigram	automatic	probabilistic model	empirical term weighting	no query expansion	automatic extraction of translation word from parallel corpus	NTCIR1/NTCIR2 Corpus
SSTUT-J-J-D-01	character	all	suffix array	bigram	automatic	probabilistic model	empirical term weighting	no query expansion	automatic extraction of translation word from parallel corpus	NTCIR1/NTCIR2 Corpus
SSTUT-J-JE-D-01	character	all	suffix array	bigram	automatic	probabilistic model	empirical term weighting	no query expansion	automatic extraction of translation word from parallel corpus	NTCIR1/NTCIR2 Corpus
SSTUT-K-K-D-01	character	all	suffix array	bigram	automatic	probabilistic model	empirical term weighting	no query expansion	automatic extraction of translation word from parallel corpus	NTCIR1/NTCIR2 Corpus
TJU-J-C-D-01	bi-charater	stemming	inverted file	bi-charater	automatic	vector space model	tf*idf	no query expansion	dictionary-based	None
TJU-J-C-TDNC-02	bi-charater	stemming	inverted file	bi-charater	automatic	vector space model	tf*idf	no query expansion	dictionary-based	None
tlrrd-C-C-DC-01	bi-character	None	inverted file	bi-character	automatic	inference networks, probabilistic model	None	None	None	None
tlrrd-C-C-D-02	bi-character	None	inverted file	bi-character	automatic	inference networks, probabilistic model	None	None	None	None
tlrrd-E-E-DC-01	word	stemming	inverted file	word	automatic	inference networks, probabilistic model	None	None	None	None

tlrrd-E-J-D-01	bi-character	None	inverted file	word for English, bi-character for Japanese	automatic	inference networks, probabilistic model	None	None	dictionary-based, select all	None
tlrrd-E-J-D-02	word	morphology, stemming	inverted file	word	automatic	inference networks, probabilistic model	None	None	dictionary-based, select all	None
tlrrd-E-J-DC-03	bi-character	None	inverted file	word for English, bi-character for Japanese	automatic	inference networks, probabilistic model	None	None	dictionary-based, select all	None
tlrrd-J-J-D-01	bi-character	None	inverted file	bi-character	automatic	inference networks, probabilistic model	None	None	None	None
tlrrd-J-J-D-02	bi-character	morphology, stemming	inverted file	bi-character	automatic	inference networks, probabilistic model	None	None	None	None
tlrrd-J-J-DC-03	bi-character	None	inverted file	bi-character	automatic	inference networks, probabilistic model	None	None	None	None
TSB-E-E-D-01	word	stemming	inverted file	word	automatic	probabilistic	BM25	expansion based on the offer weight	None	target docs only
TSB-E-E-D-02	word	stemming	inverted file	word	automatic	probabilistic	BM25	expansion based on chi-square	None	target docs only
TSB-E-J-D-01	morphemes	morphological analysis	inverted files	morphemes	automatic	probabilistic	tf-idf / tf-relevance weight	expansion based on document scores (combination of 2 runs)	MT	mainichi1998-1999 only
TSB-E-J-D-02	morphemes	morphological analysis	inverted files	morphemes	automatic	probabilistic	tf-idf / tf-relevance weight	expansion based on document scores/chi-square (combination of 3 runs)	MT	mainichi1998-2000
TSB-E-J-D-03	morphemes	morphological analysis	inverted files	morphemes	automatic	probabilistic	tf-idf / tf-relevance weight	expansion based on document scores/chi-square (combination of 3 runs)	MT	mainichi1998-2000
TSB-J-J-D-01	morphemes	morphological analysis	inverted files	morphemes	automatic	probabilistic	tf-idf / tf-relevance weight	expansion based on document scores (combination of 2 runs)	None	mainichi1998-1999
TSB-J-J-D-02	morphemes	morphological analysis	inverted files	morphemes	automatic	probabilistic	tf-idf / tf-relevance weight	expansion based on document scores/chi-square (combination of 3 runs)	None	mainichi1998-2000
TSB-J-J-D-03	morphemes	morphological analysis	inverted files	morphemes	automatic	probabilistic	tf-idf / tf-relevance weight	expansion based on document scores/chi-square (combination of 3 runs)	None	mainichi1998-2000
WATERLOO-C-C-TDNC-01	Word	Self-supervised word segmentation	Inverted files	word	automatic	Vector space model	tf, qtf, doc. leng.	None	None	None
WATERLOO-C-C-C-02	character	None	Inverted files	word	automatic	Vector space model	tf, qtf, doc. leng.	None	None	None

WATERLOO-C-C-D-03	word	Self-supervised word segmentation	Inverted files	word	automatic	Vector space model	tf, qtf, doc. leng.	None	None	None

CECIR- and *NTU-* are runs submitted from the members of Executive Committee of CLIR Task, NTCIR Workshop 3.
#OASIS-E-E-D-04 and #OASIS-E-E-D-04 are additional runs (just for reference).
^KUNLP-E-K-D-03, ^OKSAT-E-E-D-01, and ^OKSAT-J-C-D-01 are interactive runs.
!OASIS-E-E-D-01, !OASIS-E-E-D-02, !OASIS-E-E-D-03, !OASIS-E-E-D-04, !OASIS-J-J-D-01, !OASIS-J-J-D-02, !OASIS-J-J-D-03, and !OASIS-J-J-D-04 uses part of document set. Their results could not be used to compare to other runs directly. The details could be referred to the corresponding paper in workshop proceedings

*1: Query was expanded to 60 weighted terms. Additional terms were weighted and selected based on their frequencies in the query, in the top 20 documents, in the bottom 75 (of 1000) documents, and the overall frequency in the collection.
*2: Query was expanded to 60 weighted words prior to translation. The additional words were weighted and selected based on their frequencies in the query, in the top 20 ranked English documents, in the bottom 75 (of 1000) English documents, and the overall frequency in the English collection. No post-translation expansion was applied.
*3: The English subcollection of the NTCIR-3 CLIR document collection was used to assist with pre-translation expansion. When an English term could not be translated, it waseft in the query. Non word-spanning character n-grams were produced from the resultant query after translation; these n-grams were the query used to search against the Korean collection.
*4: The English subcollection of the NTCIR-3 CLIR document collection was used to assist with pre-translation expansion. When an English term could not be translated, it waseft in the query. Non word-spanning character n-grams were produced from the resultant query after translation; these n-grams were the query used to search against the Japan collection.
*5: The English subcollection of the NTCIR-3 CLIR document collection was used to assist with pre-translation expansion. When an English term could not be translated, it waseft in the query. Non word-spanning character n-grams were produced from the resultant query after translation; these n-grams were the query used to search against the Chinese collection.
*6: The English subcollection of the NTCIR-3 CLIR document collection was used to assist with pre-translation expansion for the Chinese and Japanese queries. When an English term could not be translated, it was left in the query. Non word-spanning character n-grams were produced from the resultant query after translation; these n-grams were the query used to search against the Chinese/Japanese collection. Pre-translation expansion alone was used when searching the Chinese and Japanese collections without additional blind relevance feedback, but relevancefeedback was used for the search against the English collection.
*7: qtf (within-query term fequency), dtf (within-document term frequency), ctf (within-collection term frequency), ql (query length), dl (document length), cl (collection length)
*8: 1-gram, 2-gram, dictionary-based indexing, and key-phrase formulation. Thus the index terms include single characters, all 2-grams, dictionary words, and repeated words/phrases unregistered in the dictionary.
*9: a phrase formulation technique based on repeated sequence extraction to combine unsegmented characters or words into new words or phrases.
*10: n-grams for CJK (after conversion to canonical Unicode (UTF-16)), words for Latin
*11: We used CO Model to translate queries. This model uses word co-occurrence, information to select best translation. We select Japanese topics as original queries. The original Japanese topics were translated into English. Then the translated English topics were translated into Chinese.
*12: Every query was processed twice: First search generated 2 docs. Words consisted of 3 characters were considered as candidates for expansion. Words, which occurred four times, were selected from this set. Their number did not exceed a half of the word number in the original query. In the case of the necessity, the random selection were utilized
*13: Katakana sequences were considered as a word; hiragana characters were discarded; from the rest of the text overlapping bi-gram were taken into account