The Chinese Text Retrieval Tasks of NTCIR Workshop 2

Kuang-hua Chen[+] and Hsin-Hsi Chen[*]

[+]Department of Library and Information Science
National Taiwan University
1, Sec. 4, Roosevelt Road
Taipei 10617, Taiwan
khchen@ccms.ntu.edu.tw

[*]Department of Computer Science and Information Engineering
National Taiwan University
1, Sec. 4, Roosevelt Road
Taipei 10617, Taiwan
hh_chen@csie.ntu.edu.tw

Abstract

This paper is a report of Chinese Text Retrieval (CHTR) tasks in NTCIR Workshop 2. CHTR tasks fall into two categories: Chinese-Chinese IR (CHIR) and English-Chinese IR (ECIR). The definitions, schedules, test collection (CIRB010), search results, evaluation, and initial analyses of search results of CHIR and ECIR are discussed in this paper.

1. Introduction

This paper describes the Chinese Text Retrieval tasks of NTCIR workshop 2 and introduces the Chinese Information Retrieval Benchmark version 1 (CIRB010) using in the tasks of NTCIR workshop 2. NTCIR-1 is the first evaluation workshop designed to enhance research in Japanese text retrieval [1]. We have negotiated this kind of joint efforts in evaluating Eastern-Asia text retrieval with Dr. Kando for a long time. NTCIR-2 is the result of the attempt and is the first evaluation workshop designed to enhance research in Japanese and Chinese text retrieval. The CHTR tasks fall into two categories: Chinese queries against Chinese documents (CHIR, a monolingual IR task) and English queries against Chinese documents (ECIR, a cross-language IR task). Both CHIR and ECIR are ad hoc IR tasks, i.e., the document set is fixed for various topics.

The goals of CHIR and ECIR tasks in NTCIR are shown as follows.

- Promote the Chinese IR researches
- Investigate effective techniques for Chinese IR
- Construct a mechanism for Chinese IR evaluation
- Provide a forum to present research results and exchange research ideas

The test collection used in CHTR tasks of NTCIR-2 is called CIRB010. It contains 132,173 documents [2]. These documents are all news stories downloaded from web sites of Chinatimes [3], Chinatimes Commercial [4], Chinatimes Express [5], Central Daily News [6], and China Daily News [7] during the period of May 1998 to May 1999. We are authorized to use these news stories for evaluation in NTCIR. The advantages of using newswire are manifold. The newswire is usually quite novel, quickly-updated and with multiple subjects in contents. In addition, the evaluation results would be more reliable through the common and popular data available from real environment.

Each participant could conduct ECIR task, CHIR task or both tasks. Sixteen groups from seven countries or areas had enrolled CHTR tasks. Among them, 14 groups enrolled CHIR task and 13 groups enrolled ECIR task. However, not all enrolled groups submit search results. Table 1 shows the distribution of groups enrolling CHIR and ECIR tasks and groups submitting search results at final. The search results of 115 runs were submitted from 11 groups. 98 runs from 10 groups are for CHIR task; 17 runs from 7 groups are for ECIR task. Table 2 shows the detailed statistics.

Table 1. Distribution of Participants

	Enrolled		Submitted	
	CHIR	ECIR	CHIR	ECIR
Canada	1	1	0	0
China	2	1	2	1
Hong Kong	2	1	1	0
Japan	3	2	3	1
Taiwan	2	2	2	2
UK	1	1	0	0
USA	3	5	2	3

Table 2. Participants of CHTR Tasks

	CHIR	ECIR	Total
# of groups enrolled	14	13	16
# of groups submitted	10	7	11
# of submitted runs	98	17	115

The rest of this report will focus on the test collection (CIRB010) and the CHIR task and ECIR task. Section 2 will introduce the task of CHTR in NTCIR workshop 2. Section 3 will describe the test collection used in the CHTR tasks. 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. CHIR Task and ECIR Task

Two kinds of IR tasks have been arranged for NTCIR-2 Chinese Text Retrieval. The first is Chinese IR Task (a monolingual IR task) and the second is English-Chinese IR Task (a cross-language IR task). Both tasks are ad-hoc-based tasks, that is to say, the document set is fixed against the different topics.

2.1 Schedule

By 2000-07-15:	Submit an application. Online application is available at: http://www.rd.nacsis.ac.jp/~ntcadm/workshop/application2/app2-en.html
2000-08-31:	CIRB-1-CH CD (132,172 documents and 50 Chinese topics) will be distributed to the participants of Chinese IR Task, and CIRB-1-EN CD (132,172 documents and 50 English topics) will be distributed to the participants of English-Chinese IR Task.
2000-09-30:	Search results and system description forms submission.
2001-01-10:	Results of Relevance Assessments will be distributed to the participants.
2001-02-12:	Papers for the working-note proceedings submission.
2001-03-07/ 2001-03-09:	Workshop meeting at NII, Tokyo, Japan.
2001-03-16:	Camera-ready copies for the proceedings.

2.2 Task Type

- Chinese IR Task ("CHIR")
  The Chinese IR Task is to assess the capability of participating systems in retrieving Chinese documents using Chinese queries. Chinese texts, which are composed of characters without explicit word boundary, make the retrieval task more challengeable than English ones. The participating systems can employ any approaches. Either word-based or character-based systems are acceptable. The organizer will not provide any segmentation tools and Chinese dictionaries.
- English-Chinese IR Task ("ECIR")
  The English-Chinese IR Task is to assess the capability of participating systems in retrieving Chinese documents using English queries. The organizer will not provide any segmentation tools and English-Chinese dictionaries.

2.3 Query Type

We distinguish each run according to the length of query. Three different types of run are defined as the follows.

- Long query ("LO"): Any query uses the narrative of the topics.
- Short query ("SO"): Any query uses no narrative of the topics.
- Very short query ("VS"): Any query uses neither narrative nor question of the topics.
- Title query ("TI"): Any query uses the title of the topics only.

The participating group could use any type of query to carry out the IR tasks.

3. The Test Collection: CIRB010

3.1 Document Set

In order to facilitate the process of identification and analysis of the contents, documents are supposed to be consistent in their format. Therefore, we edit the html documents downloaded from web and delete the noises. In addition, we add tags to mark the designated fields, which are document id-number, news reporting date, title, paragraphs, etc. Consequently, every document has the same format and tags. The documents are encoded in BIG5 with XML-style tags. We add tags to documents to mark their specifications and sections. The meaning of each tag is described below:

- <doc> </doc>: Denote the beginning and the ending of a document.
- <id> </id>: Denote the identification code of the document, which is composed of the source, the subject category, and the serial number of document. The code can also be seen as the full file path of each document in our data CD. The document is assigned a 7-digits serial number in each category, so we can identify each document and recognize its source by this unique id code.
- <date> </date>: Denote the date of the news using ISO8601 format. It is presented in the format of "year (in 4 digits)-month (in 2 digits)-date (in 2 digits)".
- <title> </title>: Denote the title of news.
- <text> </text>: Denote the text of news.
- : Denote the paragraphs of news.

Figure 1 shows a sample document. We only modify documents' format and do not change their contents.

Table 3 shows the statistics of CIRB document

set. CIRB010 contains 132,173 documents with the size of 200MB. The subjects of documents are various, such as politics, finance, social, life, sports, entertainment, international issue, and information technology and so on.

```
<doc>
<id>cts_foc_0005657</id>
<date>1999-05-07</date>
<title>解決高鐵融資 尋求第三管道</title>
<text>
<p>
【記者羅雨莎台北報導】據負責台灣高速鐵路聯合貸款的主辦銀行表示，高鐵融資問題目前仍卡在銀行團、交通部高鐵局以及台灣高鐵公司「三方合約」內容的訂定。在銀行團和交通部一直未能就相關歧見達成共識之下，三大主辦銀行原則決定，將尋求行政院經建會等第三管道與交通部協調，以儘早解決銀行團和交通部之間對融資問題的歧見。</p>
<p>
高鐵案將向國內銀行融資二千八百多億元，這項聯貸案確定由交銀、台銀和中國國際商業銀行共同主辦。不過，由於高鐵是國內首宗ＢＯＴ案，潛在風險究竟有多高，銀行無從評估。</p>
<p>
據主辦銀行主管表示，銀行當然希望債權確保不會有問題，譬如，在三方合約中訂定，由政府出面保證萬一將來台灣高鐵公司蓋不下去時，政府可以出面買下，負責把工程完成等。</p>
</text>
</doc>
```

Figure 1. Example of CIRB Document Tagging

Table 3. Document Set

News Agency	# of Document	Percentage
Chinatimes	38,163	28.8%
Chinatimes Commercial	25,812	19.5%
Chinatimes Express	5,747	4.4%
Central Daily News	27,770	21.0%
China Daily News	34,728	26.3%
Total	132,173	(200MB)

3.2 Topic

Three main procedures constructing topics of CIRB010 are shown as follows:

(1) Collecting information request
In order to increase the similarity between our benchmark and real environment, we build the topics using real users' information requests. We collected 405 requests through questionnaire on web. There are both closed and open-ended questions about the types and subject of requests, narratives of requests, and other related information. The basic assumption of the method is that users may state their specific information request distinctly and exhaustively.

(2) Selecting information request
The responses of questionnaires gained from Internet was not entirely so qualitative, complete and exhaustive, and the type and subject of the request provided by user is not necessarily suitable for evaluation purpose in our IR benchmark. Therefore, we pick out 50 best requests from 405 collected requests according to some important criteria in three phases explained as follows. In the first phase we examine the statements and narratives, and then delete too simple, short, ambiguous, and subjective ones. Besides, we also exclude the requests with the following criteria: (a) the coverage of subject is too broad; (b) request shows great gap in document set; (c) the answer to request change rapidly from time to time. In the second phase, a full-text information retrieval system is used to test the number of possible relevant documents. The motive of this method is to determine if the request is too comprehensive or too limited in their subject coverage. In the final phase, we select 50 requests best fitting in the requirements. The main consideration lies in the similarity between requests and degree of the explicitness and distinctness of the request. The results of request selection are shown as Table 4.

Table 4. The Selection of Information Request

	1st selection	2nd selection	3rd selection
Selection method	Judged by researcher	Supported by full-text IR software	Judged by researcher
Number of topics deleted	163	173	19
Number of topics reserved	242	69	50

(3) Constructing Topics
The main task of this phase is to establish the topics in accordance with the 50 final requests. We use four fields: title, question, narrative, and concepts to represent topics in accordance with the TREC's convention. The meaning, content, syntax and resources of each field are shown in Table 5. The "title" field has the widest coverage in its content with comparison to the other three fields. The coverage of "question" field is the second to "title" field. The "narrative" field is the most specific because of its detailed description. The keywords in "concepts" field touch on the contents of above three fields. The relationship between the four fields can be illustrated in Figure 2. The 50 topics can be roughly classified into 9 categories. The average number of words in a topic is 169. No significant differences show in the corresponding fields among different topics with comparison to other test collections. The statistics is shown in Table 6.

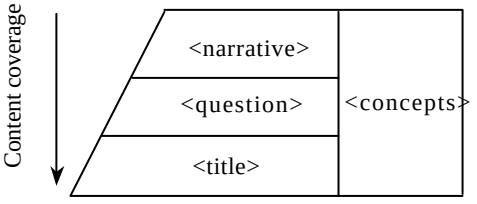

Figure 2. The Relation between Topic Fields

Table 5. Fields of CIRB Topics

Fields	Content	Syntax	Resources
<title>	Concise representation of information request subject.	Noun or Noun Phrase	The subject of information request
<question>	Brief descriptions of content of information request.	One or two sentences	The whole information request
<narrative>	Narratives of information request such as the further interpretation to request and proper nouns, the list of relevant or irrelevant information, and the specific requirements or limitations of relevant documents.	Several sentences	The whole information request
<concepts>	Keywords relevant to the whole topic.	One or more keywords	Relevant or irrelevant keywords about information request

<topic>
<number> CIRB010TopicZH011</number>
<title>**金融機構合併**。</title>
<question>
查詢我國政府單位鼓勵金融機構合併之各項措施。
</question>
<narrative>
財政部等相關單位為健全金融市場、改善金融體質，推動了一連串鼓勵銀行、證券商及保險公司等金融機構合併的措施。相關文件內容包括各項具體的獎勵優惠辦法、施行細節、法令中明定之規範條文、以及各界對相關政策的討論與評估。若文件中只陳述金融機構合併之個案，視為不相關。
</narrative>
<concepts>
金融機構、合併、銀行合併、租稅優惠、租稅減免、稅前盈餘、低利融資、促進產業升級條例、財政部、經濟部、央行、中央銀行、增值稅、印花稅、證交稅。
</concepts>
</topic>

Figure 3. Example of CIRB Topic

Table 6. Document Length of CIRB and TREC

	Fields	Minimum	Maximum	Average	Standard Deviation	Standard Deviation/Average
CIRB	<title>	3	13	6.52	2.23	0.34
	<question>	12	37	23.64	5.92	0.25
	<narrative>	57	141	93.90	20.43	0.22
	<concepts>	26	74	44.68	11.58	0.26
	Total	103	244	168.74	27.77	0.16
TREC Chinese Topic 1-54	<title>	4	29	12.30	5.58	0.45
	<desc>	6	35	17.48	7.40	0.43
	<narr>	31	174	81.54	30.28	0.37
	Total	53	204	111.32	31.36	0.28
TREC-6 Topic 301-350 (Chinese translation)	<title>	3	13	6.80	2.28	0.34
	<desc>	7	87	30.14	16.88	0.56
	<narr>	26	217	94.56	42.15	0.45
	Total	64	237	131.5	42.03	0.32

4. Evaluation

This is our first attempt to organize Chinese IR evaluation workshop. We follow the method used in TREC and NTCIR-1. The TREC's evaluation program is used to score the research results. It provides the interpolated recall and precision at 11 points, average precision (non-interpolated) over al 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 CIRB010 document "docid" (a string extracted from the <id> </id> field, e.g. <id> cts_cec_1999111514 </id>) retrieved by query "qid" (an integer extracted from the last 3 digits in <number> </number> field of topic, e.g., <number> CIRB010TopicEN002 </number>, 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.

A sample file is shown as follows.

001	0	CTS_POL_0003207	0	0.992917 RUN-ID
001	0	CTS_POL_0003217	1	0.992338 RUN-ID
001	0	CDN_FOC_00057312		0.988612 RUN-ID
001	0	CDN_FOC_00041853		0.987523 RUN-ID
001	0	CHD_POL_00058924		0.985803 RUN-ID
001	0	CDN_FOC_00056435		0.984673 RUN-ID
001	0	CTS_POL_0002947	6	0.978978 RUN-ID
001	0	CTS_POL_0003249	7	0.973372 RUN-ID
001	0	CDN_FOC_00041968		0.966231 RUN-ID
001	0	CTS_POL_0003216	9	0.931390 RUN-ID
001	0	CHD_POL_000251210		0.867546 RUN-ID
001	0	CDN_FOC_000695111		0.857641 RUN-ID

A list of relevance between each topic and documents in a benchmark is needed to facilitate the comparison and evaluation of IR system effectiveness. This is the so-called "relevance judgment." The relevance judgment is undertaken by pooling method. Due to the fact that each run submits about 1,000 documents, the following criteria are used to form the pool.

- For each participating group, only one run for each query type run is selected, i.e., at most 4 runs are selected for each participating group.
- A document qualifying as a member of pool has to be retrieved by more than one run.

On average, the size of pool for each topic is about 900 documents. Table 7 shows the total number of documents in pool.

Table 7. The Size of Pool for each Topic

topic 001	1035	topic 018	716	topic 035	1008
topic 002	922	topic 019	896	topic 036	898
topic 003	1016	topic 020	880	topic 037	918
topic 004	956	topic 021	987	topic 038	540
topic 005	1015	topic 022	1271	topic 039	720
topic 006	703	topic 023	845	topic 040	1134
topic 007	1175	topic 024	905	topic 041	935
topic 008	1177	topic 025	895	topic 042	912
topic 009	663	topic 026	1158	topic 043	765
topic 010	1145	topic 027	812	topic 044	897
topic 011	995	topic 028	807	topic 045	881
topic 012	1086	topic 029	830	topic 046	764
topic 013	748	topic 030	923	topic 047	720
topic 014	928	topic 031	704	topic 048	1017
topic 015	882	topic 032	736	topic 049	989
topic 016	926	topic 033	710	topic 050	754
topic 017	889	topic 034	776	average	898.48

While performing relevance judgments, every judge should read and understand the meaning of the topic carefully and assign each of them to the most appropriate category (mentioned below) from their viewpoint mainly according to "question" field of topic. In order to keep judges' criterion consistent, the judges must complete the judgments for a topic in a period of time. Each topic is judged by 3 judges. In total, 23 judges spend 799 hours in relevance judgment.

The "subject relevance" concept is adopted in relevance judgment. That is to say, we pay more attention to the concrete meaning, which can be perceived from the text. Based on this concept, the judges should make an objective link between document and topic. This will increase the consistency and reliability of judgments performed by different judges. 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: "Very Relevant", "Relevant", "Partially relevant", and "Irrelevant." Each kind of relevance is assigned a relevance score. "Very relevant" is 3, "Relevant" is 2, "Partially relevant" is 1, and "Irrelevant" is 0.

Since one unified relevance score have to be produced for final relevance judgment using TREC's scoring program, we combine judgment results of three judges, and then decide how to interpret the meaning of the score and how can it be applied to IR evaluation. Based on the following philosophy, we devise a method to integrate 3 relevance scores to form one relevance score.

- Each judge has equal contribution to final relevance score.
- Each judgment is independent.

The following formula is used to combine 3 judges' relevance score,

$$R = \frac{(X_A + X_B + X_C)/3}{3}$$

where X means the relevance category assigned by each judge, and A, B, C represent the three different judges. The value of R will be between 0 and 1. The closer the score to 1, the more relevant the document to the topic.

As mentioned above, TREC scoring program is used to calculate recall and precision. Since it uses binary relevance judgment, we have to decide the threshold. Two thresholds are decided: one is 0.6667; the other is 0.3333. The so-called rigid relevance means the final relevance score should between 0.6667 and 1. That is to say, it is equivalent that each person assigns "relevant (2)" to the document.

[(2+2+2)/3/3=0.6667]

The so-called relaxed relevance means the final relevance score should between 0.3333 and 1. That is to say, it is equivalent that each person assigns "partially relevant (1)" to the document.

[(1+1+1)/3/3=0.3333]

5. Search Results

We will report the search results in a broad view and analyze some of runs using different query types in this section. The different techniques which each participating group took could be referred to each paper in workshop proceedings.

5.1 CHIR Task

The search results of CHIR task are submitted from 10 participating groups listed in Table 8. Some groups are "Full Participation" and some are "Anonymous Participation". The number of submitted runs by the query types is shown in Table 9. The query types have been mentioned in Section 2.

(1) All Runs

The recall/precision graphs of top runs of CHIR task are showed in Figure 4 (relaxed relevance) and Figure 5 (rigid relevance). The techniques used in these top runs are showed in Table 10 and Table 11. From Table 10 and Table 11, it is easy to find out that all runs use query expansion techniques except Brkly-CHIR-LO-01. CRL group performs well in all query types, since it uses automatic feedback to carry out query expansion. We also find that stop-word list seems a good resource for Chinese information retrieval. Brkly group applies logistic regression technique to tune the various parameters. The unique technique shows good performance in CHIR task.

Basically, most of the participating groups use tf/idf approach in a little different form. This shows that the long-history tf/idf approach still play an important role in information retrieval.

Both CRL group and PIRCS group adopt probabilistic model and they are the leading groups in CHIR task. This implies that the probabilistic model shows good performance in CHIR task of NTCIR workshop 2.

In general, the performance of "Very Short Query" is the best; the performance of "Short Query" is better than that of "Long Query". The performance of "Title Query" is worst. It seems that the long query conveys many noises. On the contrary, the title query conveys little information. Observing the search results, we conclude that the short query is appropriate for CHIR task. It seems that the name of "Very Short Query" will mislead us to draw a direct conclusion that query should be short. In fact, the "Very Short Query" means that the participants could use the concepts identified in the topic. In the case of CIRB010, the <concepts> field contains many significant keywords. As a result, the runs applying "Very Short Query" perform well.

Since the leading groups show good performance in all runs, we will not expla in the details in each query type and the technical details of each participating group should be referred to the corresponding papers in conference proceedings.

Table 8. List of CHIR Participating Groups

1	Communications Research Laboratory
2	University of California at Berkeley
3	Queens College, CUNY
4	Chinese research group of Lab Furugori, the University of Electro-Communications
5	Department of Computing Hong Kong Polytechnic University
6	Umemura Lab. Department of Information and Computer Sciences, Toyohashi University of Technology
7	NTHU NLP Lab and Knowledge Express Technology Inc
8	Institute of Software, Chinese Academy of Sciences
9	Trans-EZ Information Technology Inc.
10	Fujitsu R&D Center

Table 9. Number of Submitted Runs of CHIR

# of Groups	10
# of Total RUN	98
# of "LO" RUN	30
# of "SO" RUN	12
# of "VS" RUN	27
# of "TI" RUN	29

(2) "Long Query" Runs

The recall/precision graphs of top runs of CHIR "Long Query" are showed in Figure 6 (relaxed relevance) and Figure 7 (rigid relevance). Since the top runs measured in relaxed relevance and in rigid relevance are the same, we only use one table (Table 12) to describe the techniques used in these top runs.

(3) "Short Query" Runs

The recall/precision graphs of top runs of CHIR "Short Query" are showed in Figure 8 (relaxed relevance) and Figure 9 (rigid relevance). Since the top runs measured in relaxed relevance and in rigid relevance are the same, we also use one table (Table 13) to describe the techniques used in these top runs.

(4) "Title Query" Runs

The recall/precision graphs of top runs of CHIR "Title Query" are showed in Figure 10 (relaxed relevance) and Figure 11 (rigid relevance). The techniques used in these top runs are showed in Table 14 and Table 15.

(5) "Very Short Query" Runs

The recall/precision graphs of top runs of CHIR "Very Short Query" are showed in Figure 12 (relaxed relevance) and Figure 13 (rigid relevance). The techniques used in these top runs are showed in Table 16 and Table 17.

5.2 ECIR Task

The search results of ECIR task are submitted from 7 participating groups listed in Table 18. All groups are "Full Participation". The number of submitted runs by the query types is shown in Table 19.

Table 18. List of ECIR Participating Groups

1	University of California at Berkeley
2	University of Maryland
3	Queens College, CUNY
4	Umemura Lab. Department of Information and Computer Sciences, Toyohashi University of Technology
5	NTHU NLP Lab and Knowledge Express Technology Inc
6	Institute of Software, Chinese Academy of Sciences
7	Trans-EZ Information Technology Inc.

Table 19. Number of Submitted Runs of ECIR

# of Groups	7
# of Total RUN	17
# of "LO" RUN	8
# of "SO" RUN	2
# of "VS" RUN	6
# of "TI" RUN	1

(1) All Runs

The recall/precision graphs of top runs of ECIR task are showed in Figure 14 (relaxed relevance) and Figure 15 (rigid relevance). Since there are only 17 ECIR runs, we use one table (Table 20) to list the techniques which participating groups used in ECIR task. Observing Figure 14, we find that PIRCS group outperforms other groups using relaxed relevance metric. Further investigating Table 20, we have an idea that PIRCS uses MT software to carry out translation commission. On the contrary, most groups use dictionaries with select-all, select-top-1, select-top-n, or select-all approaches. Among select-X approach, select-all is better than select-top-3; select-top-3 is better than select-top-2; select-top-2 is better than select-top-1. We could not conclude directly that select-all is the best among all select-X approaches, since some groups also apply corpus-based approach at the same time. However, no enough information shows how participating groups utilize corpus. Did they calculate mutual information? Did they calculate the bilingual mutual information? The detailed information should be referred to the corresponding papers.

Observing the index unit, we find that word-based approaches are much better than other approaches in ECIR task. In addition, PIRCS group combines word-based and character-based approaches to construct index file.

Since only PIRCS group submits runs of all query types in ECIR task, we will compare the performances of each query type based on the search results of PIRCS. The "Title query" is the worst among all query types. The difference among "Long Query", "Short Query", and "Very Short Query" is little. However, the "Short Query" is better than others. As mentioned before, the "Very Short Query" in CIRB010 conveys many important keywords, so the performance is good. This phenomenon is different from the observation pointed out in NTCIR-1 Japanese IR task [1].

Since the leading groups show good performance in all runs, we will not explain the details in each query type. Again, the interested readers have to refer to the corresponding papers in the conference proceedings for technical details.

(2) "Long Query" Runs

The recall/precision graphs of top runs of CHIR "Long Query" are showed in Figure 16 (relaxed relevance) and Figure 17 (rigid relevance).

(3) "Short Query" Runs

The recall/precision graphs of top runs of CHIR "Short Query" are showed in Figure 18 (relaxed relevance) and Figure 19 (rigid relevance).

(4) "Title Query" Runs

Only one group submits ECIR "Title Query" run. The recall/precision graphs of top runs of CHIR "Title Query" are showed in Figure 20 (relaxed relevance) and Figure 21 (rigid relevance).

(5) "Very Short Query" Runs

The recall/precision graphs of top runs of CHIR "Very Short Query" are showed in Figure 22 (relaxed relevance) and Figure 23 (rigid relevance).

6. Conclusions and Future Works

The NTCIR Workshop 2 is the first international joint effort in providing an evaluation mechanism for Japanese and Chinese Text Retrieval. We hope this mechanism could encourage the IR 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 IR researches, develop the effective method in construction of test collection, and enhance the effectiveness of IR systems.

Through the initial analyses on the submitted runs, some findings are shown as follows. The further and detailed analyses on particular systems could be referred to each system report in this workshop.

- Most participating groups apply inverted file approach for index structure.
- Many participating groups adopt tf/idf-based approaches.
- "Short Query" and "Very Short Query" show much better performance.
- Query expansion is a good method to increase system performance.
- In general, the probabilistic model shows better performance.
- For CHIR task, stop-word list is a good resource for enhancing system performance.
- For ECIR task, select-all approach seems to be better than other select-X approaches, if no further techniques are adopted.
- For ECIR task, MT approach is much better than dictionary-based approach.
- For ECIR task, word-based indexing approach is better.

We would like to say again that these findings are drawn from the submitted runs using the "CIRB010" test collection. They cannot directly apply to other test collection, since each test collection has its own characteristics and each language also has its own characteristics. We have to carry out more detailed analyses using other test collections.

As mentioned previously, the keywords in concepts filed of topic provide the crucial information and make the performance higher than other IR evaluation forum. We would like to explain the procedure for keywords preparation. We have executed a pre-test for CIRB010 test collection. As a result, the positive documents and negative documents for each topic have been constructed. We then analyze these documents and extract the good keywords for each topic. According our analysis and Brkly's experiment, using concepts field as the query will produce the best performance with comparison to the other fields. We could say that the concepts field makes the Chinese IR tasks much easier. Therefore, we are considering the role of concepts field.

Since the limitation in staff and the lack of experience, we make some mistakes in CHTR tasks. We all work in university and the teaching load is always an issue. The time is crucial for us. As a result, we will plan to cooperate with NII in a more tight way next year. We would like to apologize for making any inconvenience this year.

Some important issues are worthy of presenting in this report.

- We have planed to execute a CLIR task next year, which is cross-language in document collection level. We have discussed this issue with Dr. Noriko Kando in Taipei.
- The copyright issue of continuing uses CIRB010 in researches is under negotiation. The original agreement is that we could use it in the tasks of NTCIR workshop. We are about to solve this problem.
- According to experience of TREC, the fields of topic are changeable year by year. Do we have to proceed the same experiment?
- Since the high quality of keywords in concepts field of topic, we will reserve them for relevance judgment only next year and then make Chinese IR tasks more difficult.

Acknowledgments

We would like to thank Chinatimes, Chinatimes Commercial, Chinatimes Express, Central Daily News, China Daily News, and United Daily News for their kindly providing test materials. We are grateful to all pioneers in the area of IR evaluation for their efforts in paving a smooth way for followers. We would like to thank the participants for their contributions and the relevance judges for their hard working. Special thanks are due to Dr. Noriko Kando for her substantial help.

References

[1] Kando, N. et al. Overview of IR Tasks at the First NTCIR Workshop. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 1-46, Tokyo, 1999.

[2] Chiang, Y.-T. *A Study on Design and Implementation for Chinese Information Retrieval Benchmark*, Master Thesis, Department of Library and Information Science, National Taiwan University, 1999.

[3] China Times. http://news.chinatimes.com.tw/

[4] Chinatimes Commercial. http://news.chinatimes.com.tw/

[5] Chinatimes Express. http://news.chinatimes.com.tw/

[6] Central Daily News. http://www.cdn.com.tw/

[7] China Daily News. http://www.cdns.com.tw/

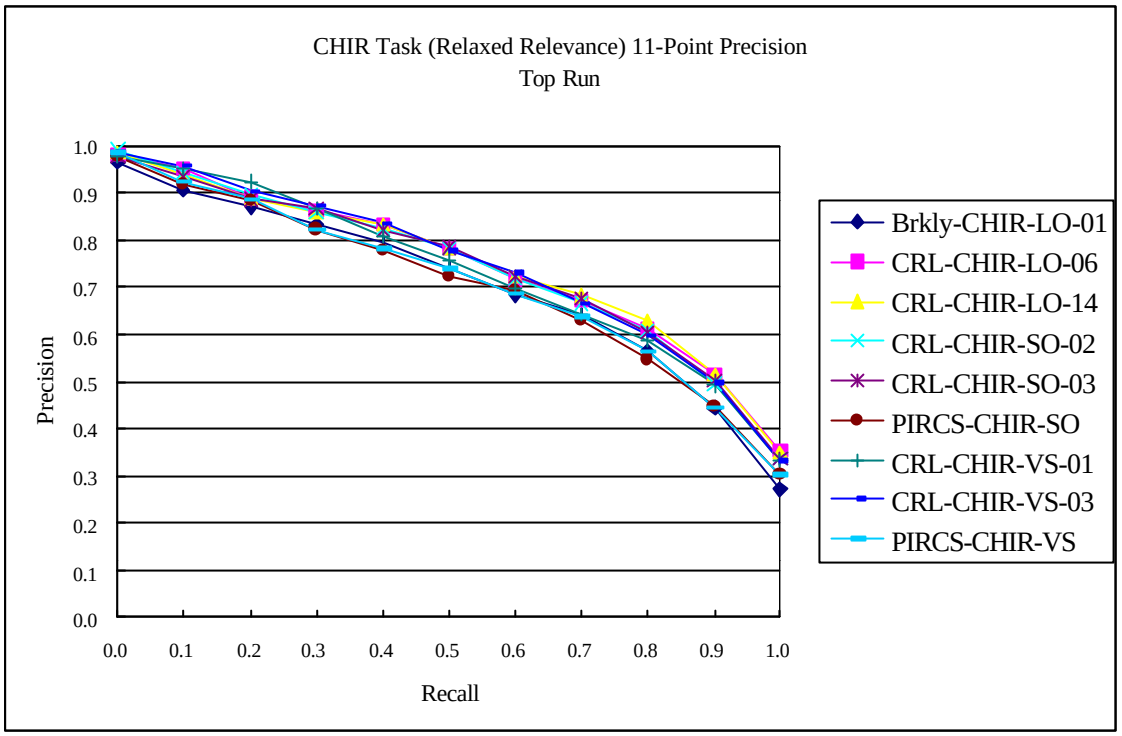

Figure 4. CHIR Task 11-Point Precision (Relaxed Relevance)

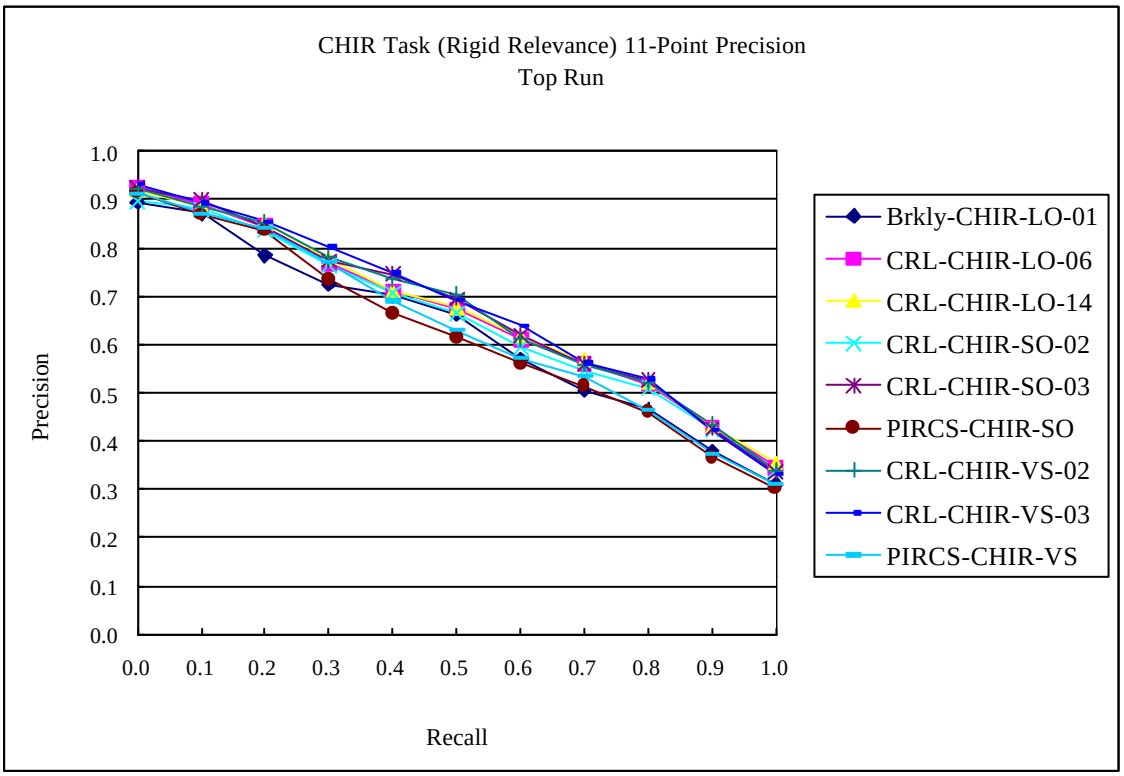

Figure 5. CHIR Task 11-Point Precision (Rigid Relevance)

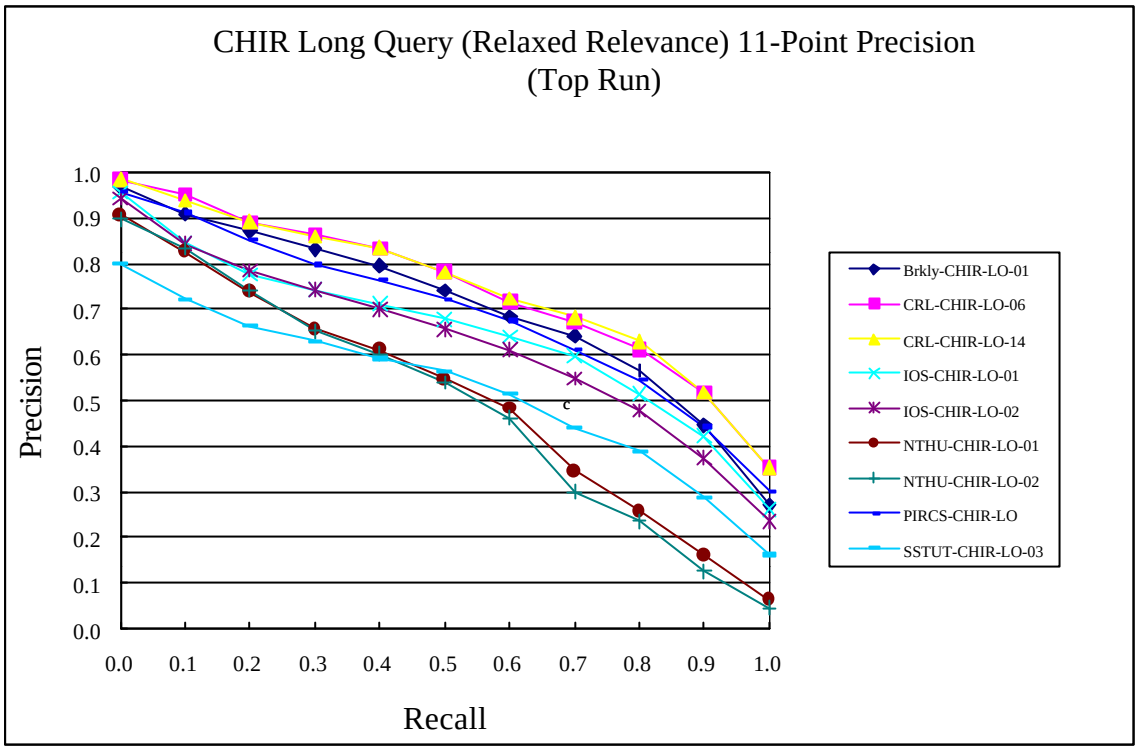

Figure 6. CHIR Long Query 11-Point Precision (Relaxed Relevance)

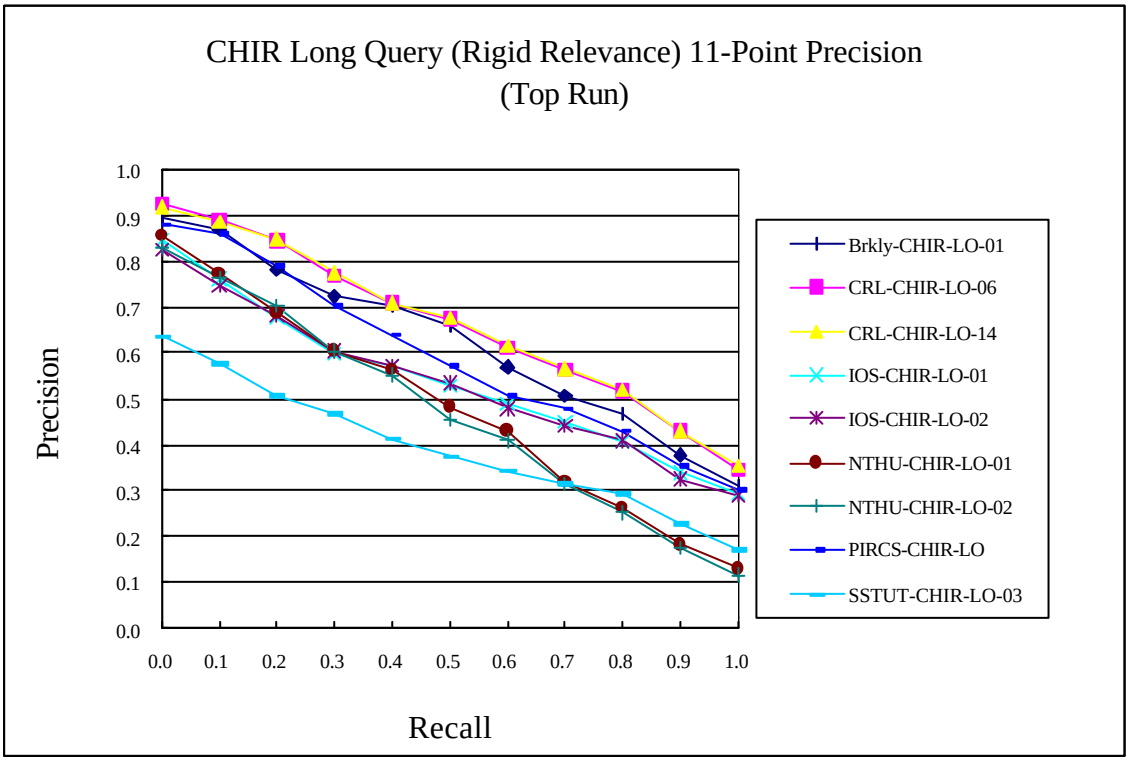

Figure 7. CHIR Long Query 11-Point Precision (Rigid Relevance)

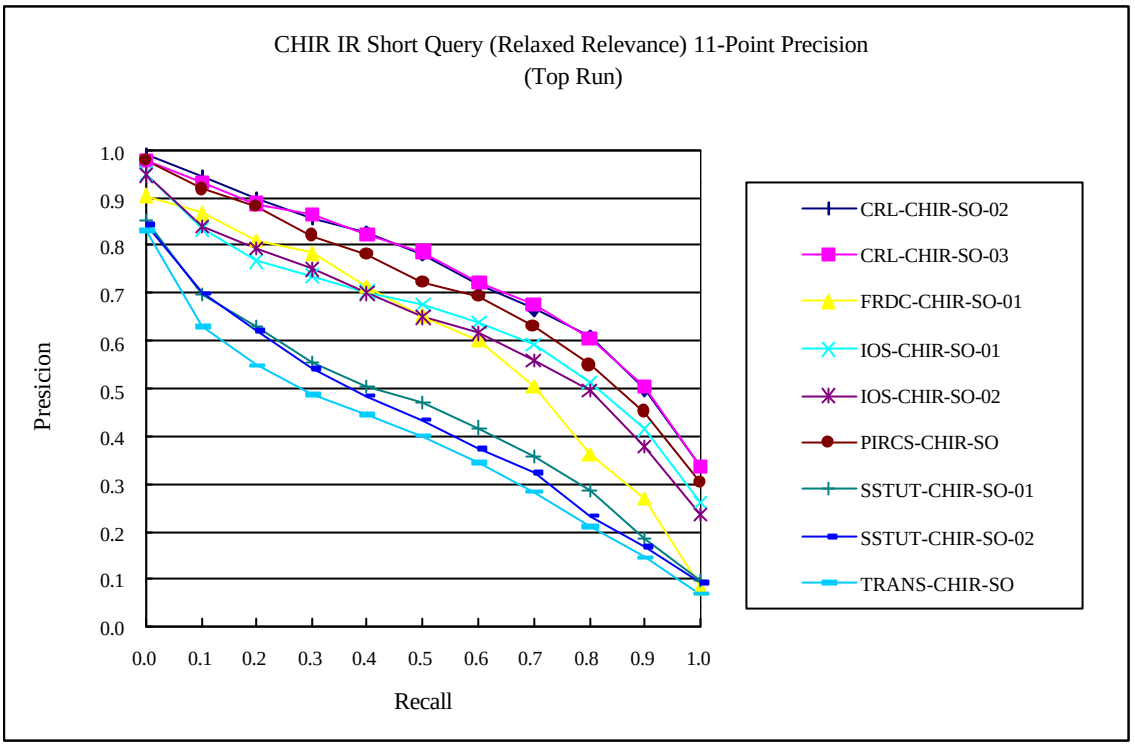

Figure 8. CHIR Short Query 11-Point Precision (Relaxed Relevance)

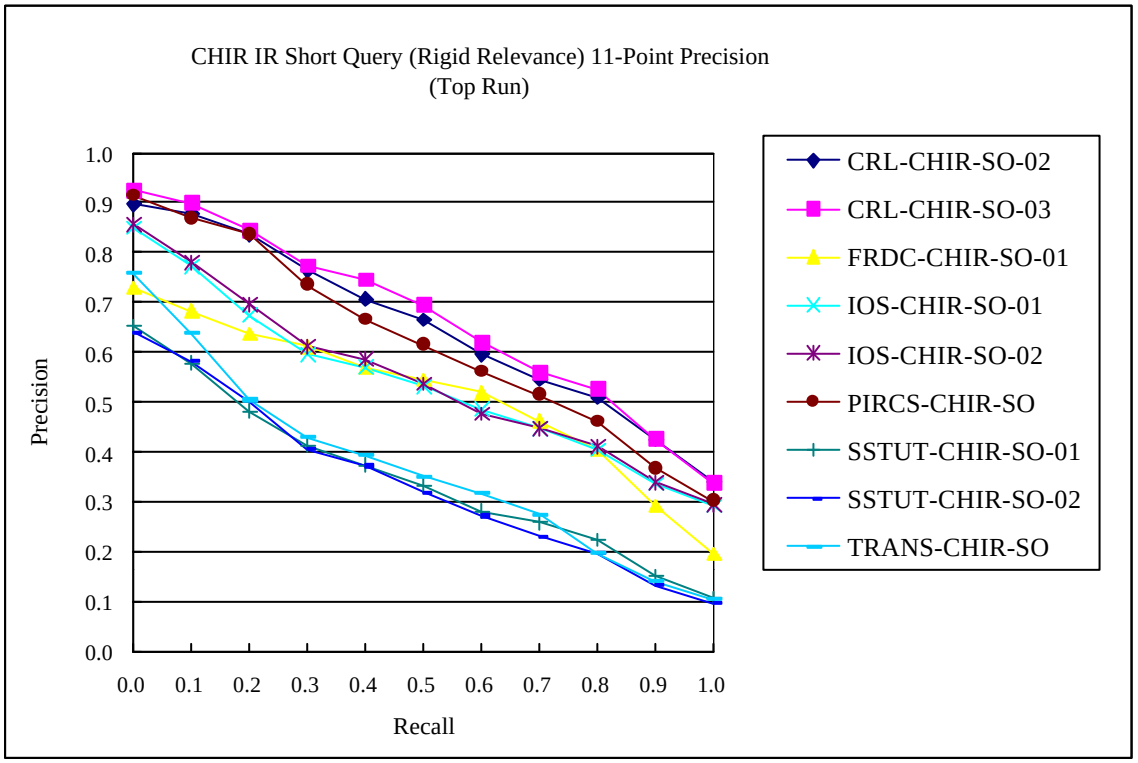

Figure 9. CHIR Short Query 11-Point Precision (Rigid Relevance)

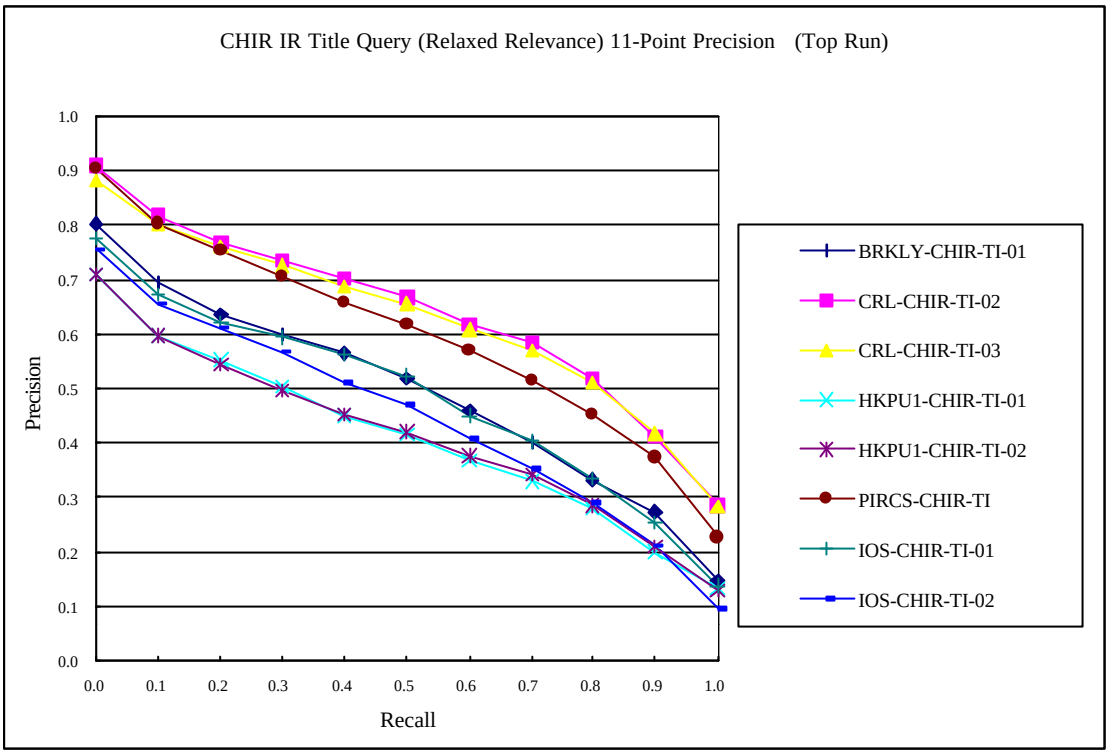

Figure 10. CHIR Title Query 11-Point Precision (Relaxed Relevance)

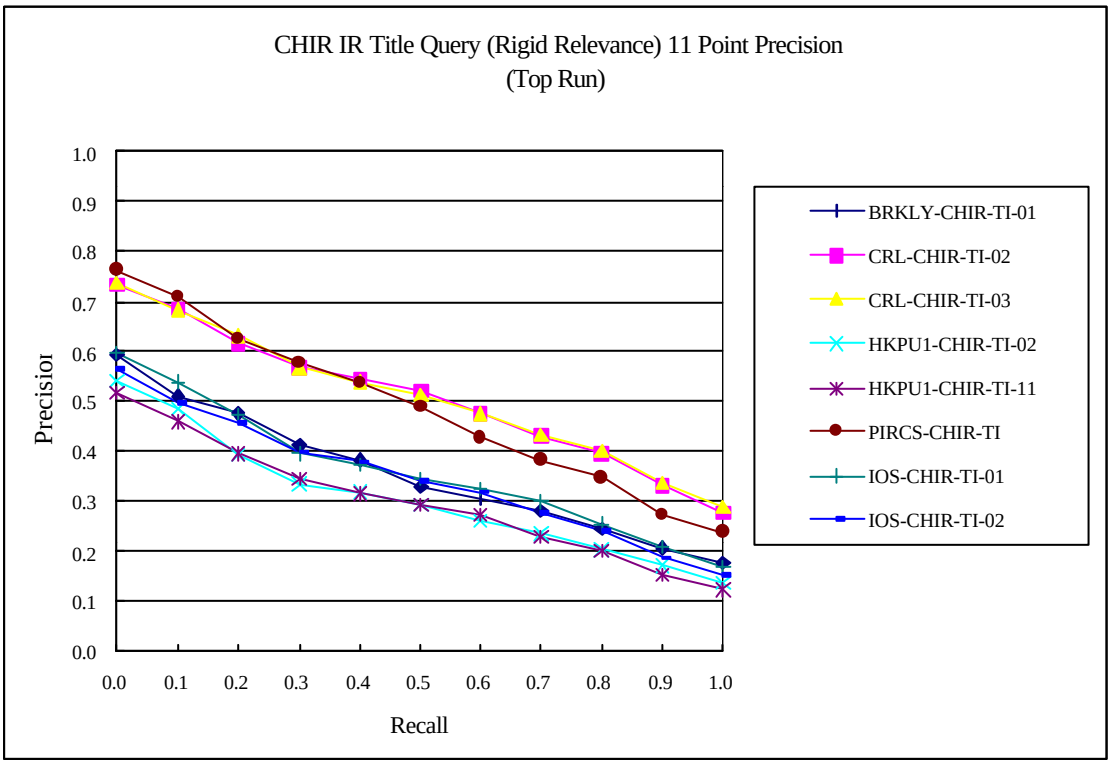

Figure 11. CHIR Title Query 11-Point Precision (Rigid Relevance)

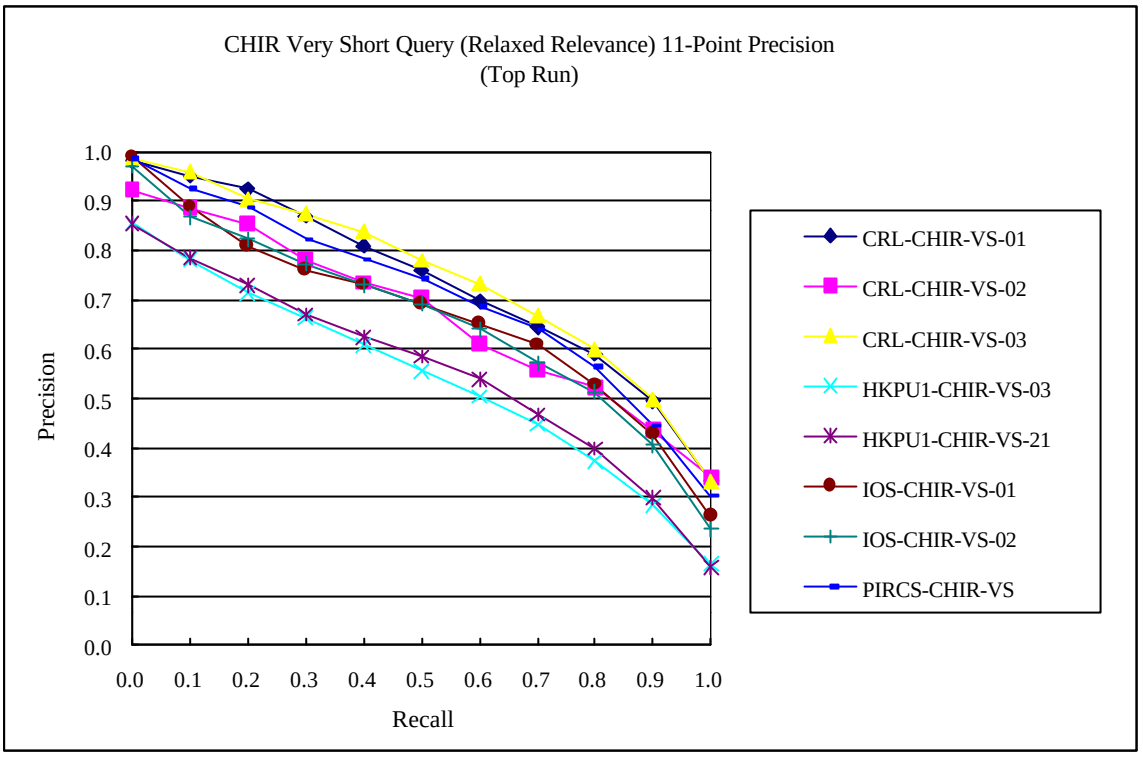

Figure 12. CHIR Very Short Query 11-Point Precision (Relaxed Relevance)

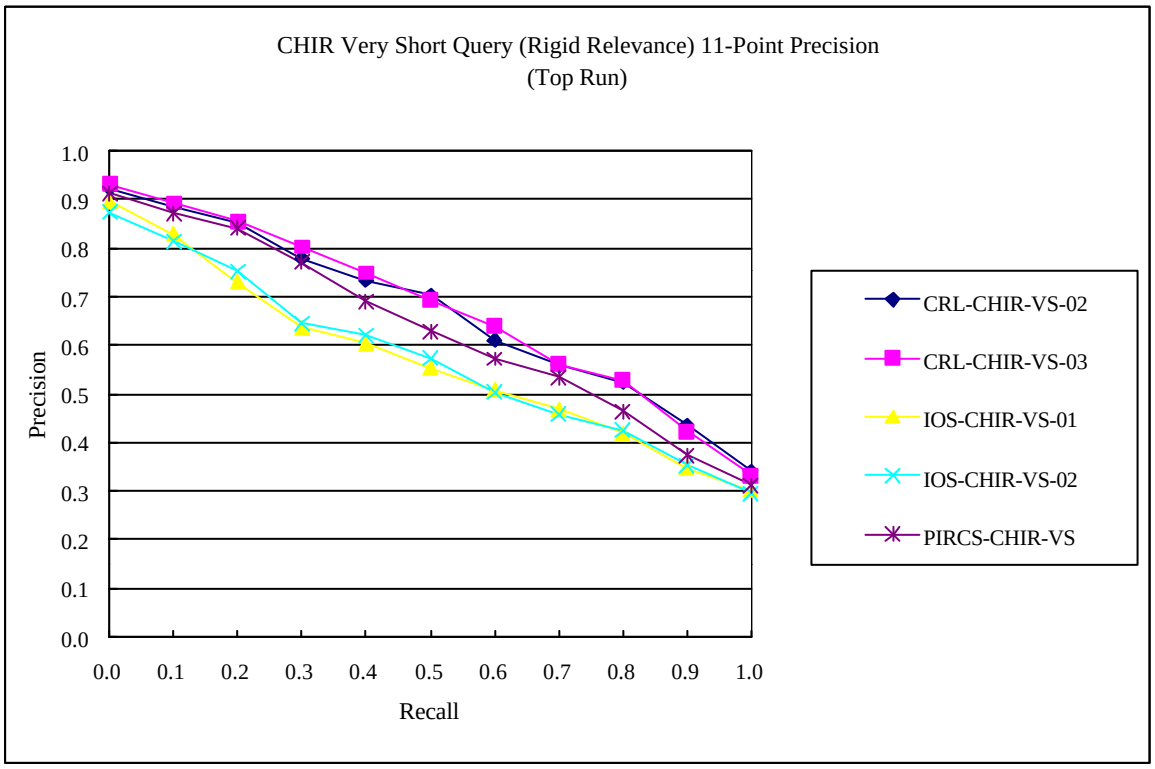

Figure 13. CHIR Very Short Query 11-Point Precision (Rigid Relevance)

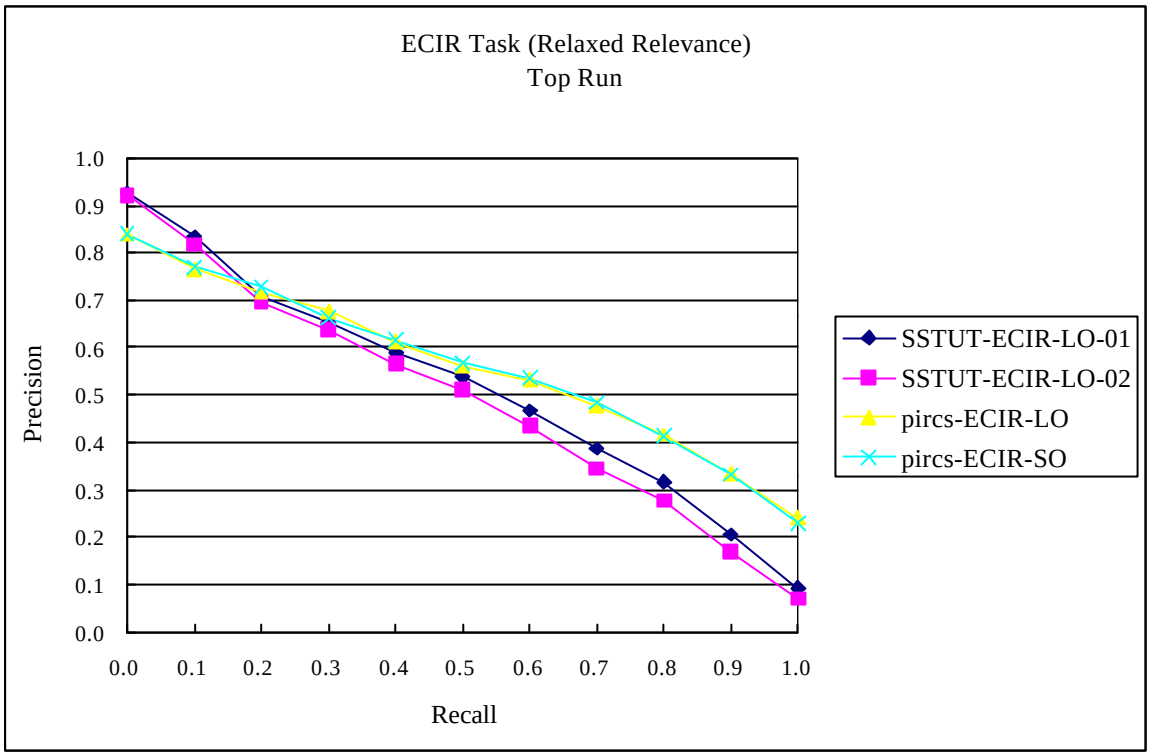

Figure 14. ECIR All Task 11-Point Precision (Relaxed Relevance)

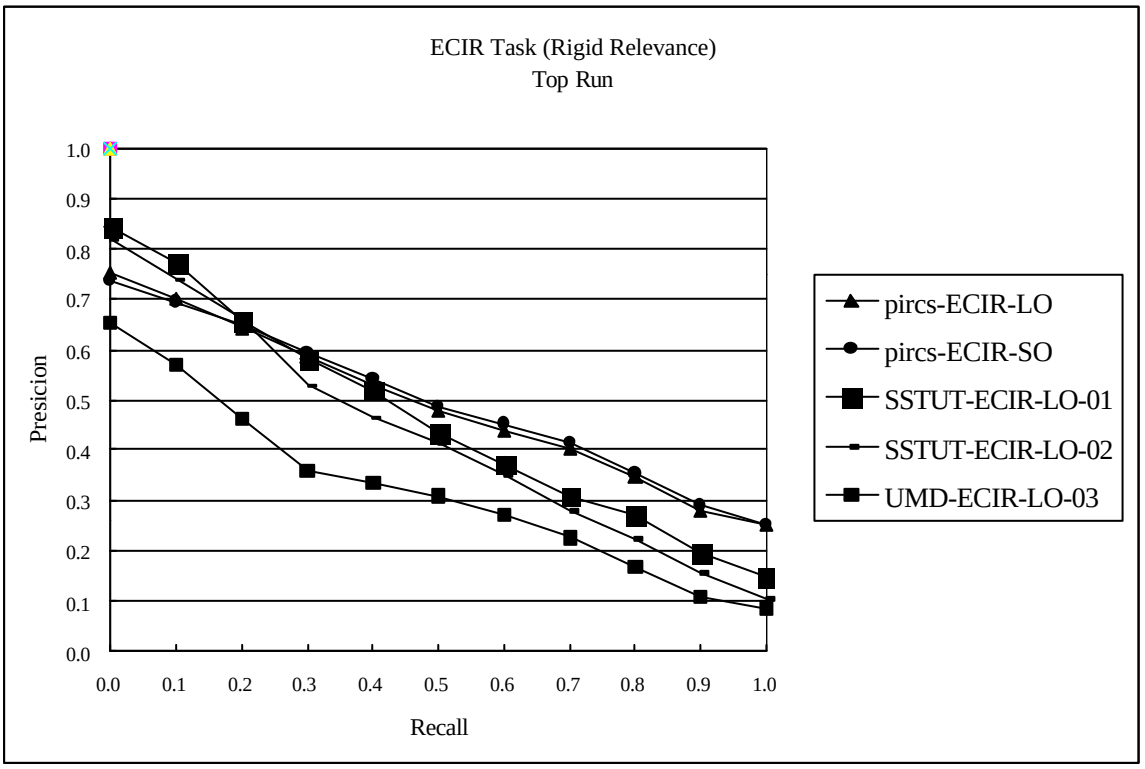

Figure 15. ECIR Task 11-Point Precision (Rigid Relevance)

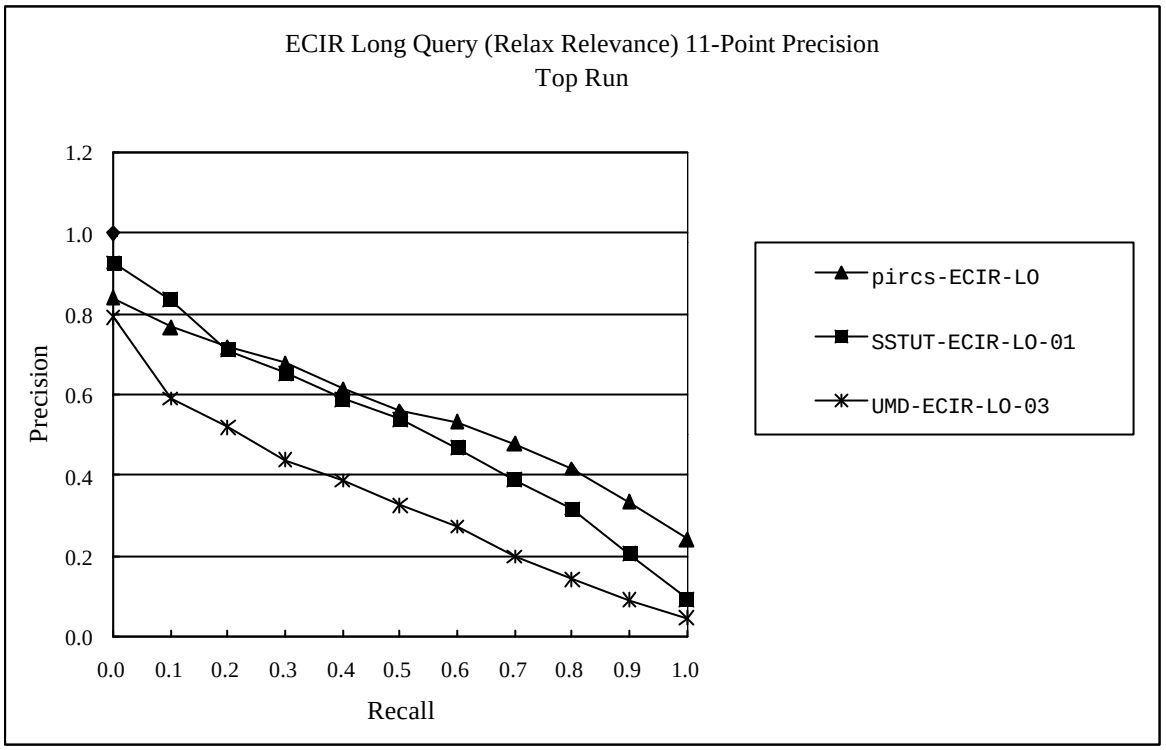

Figure 16. ECIR Long Query 11-Point Precision (Relaxed Relevance)

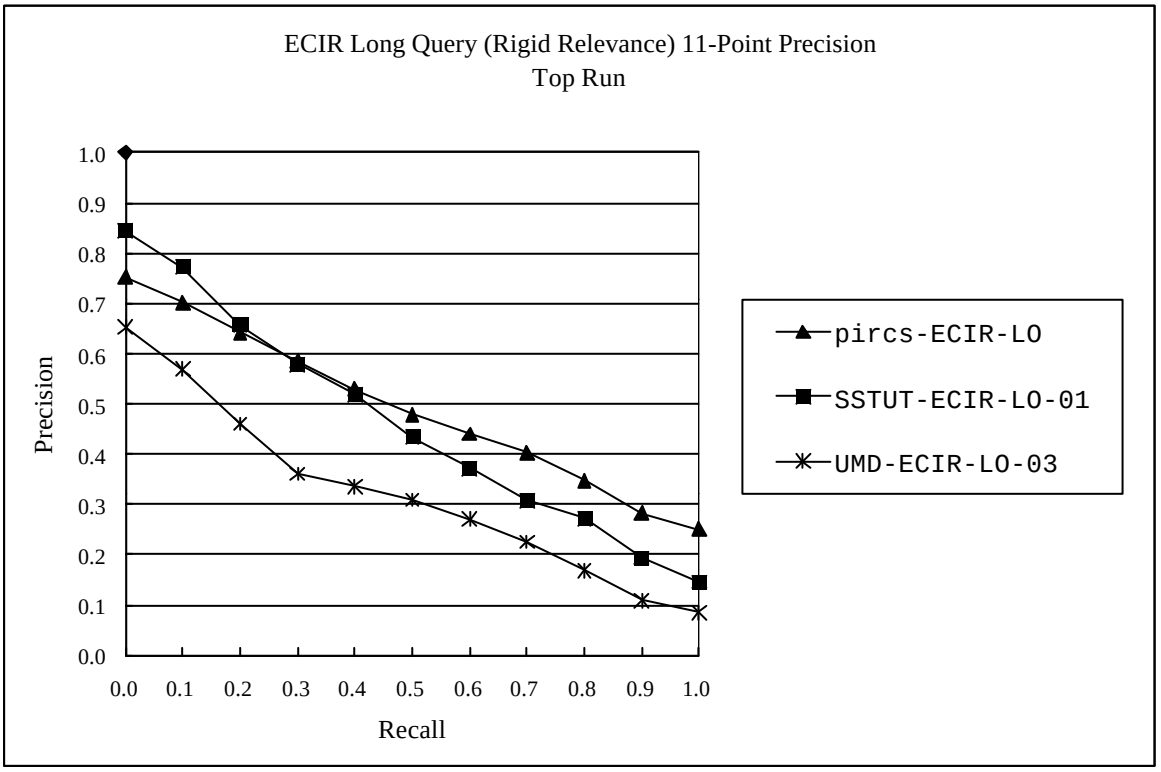

Figure 17. ECIR Long Query 11-Point Precision (Rigid Relevance)

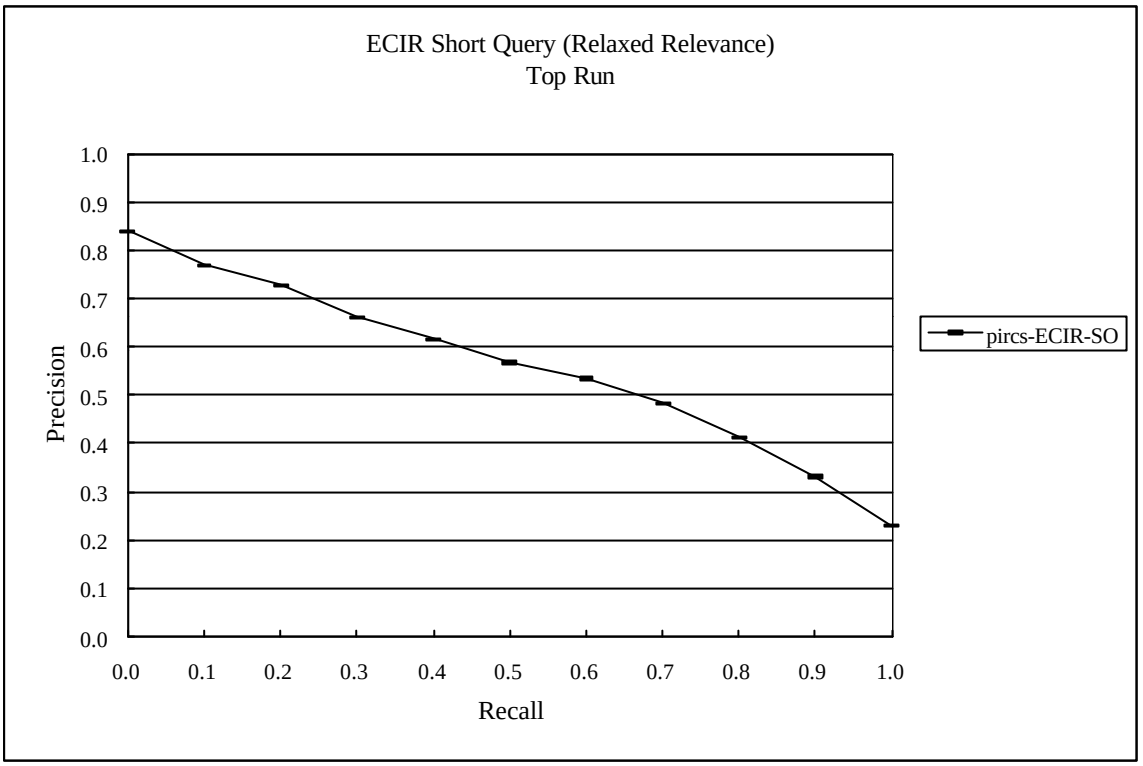

Figure 18. ECIR Short Query 11-Point Precision (Relaxed Relevance)

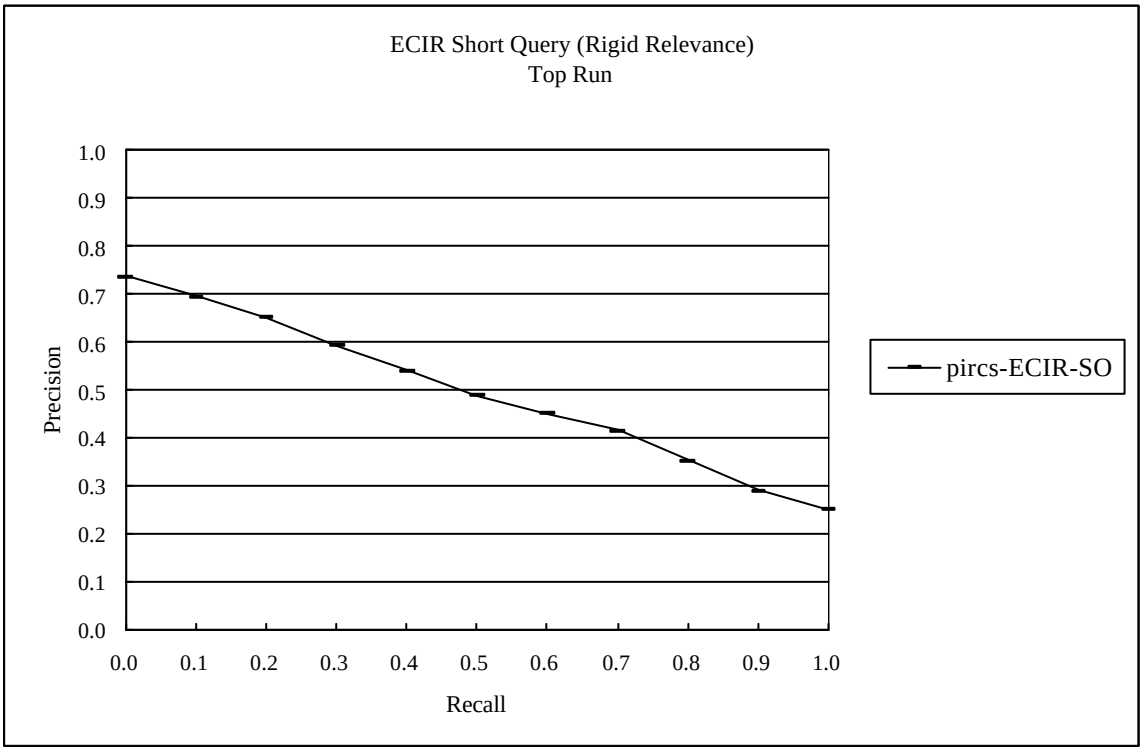

Figure 19. ECIR Short Query 11-Point Precision (Rigid Relevance)

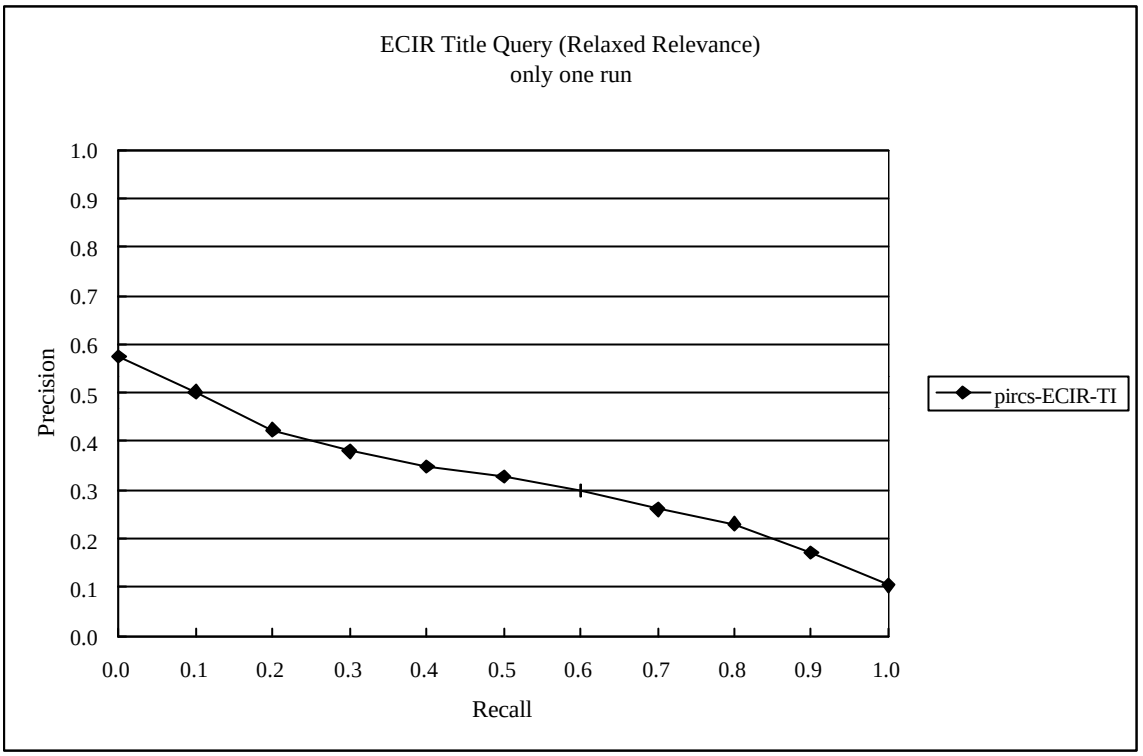

Figure 20. ECIR Title Query 11-Point Precision (Relaxed Relevance)

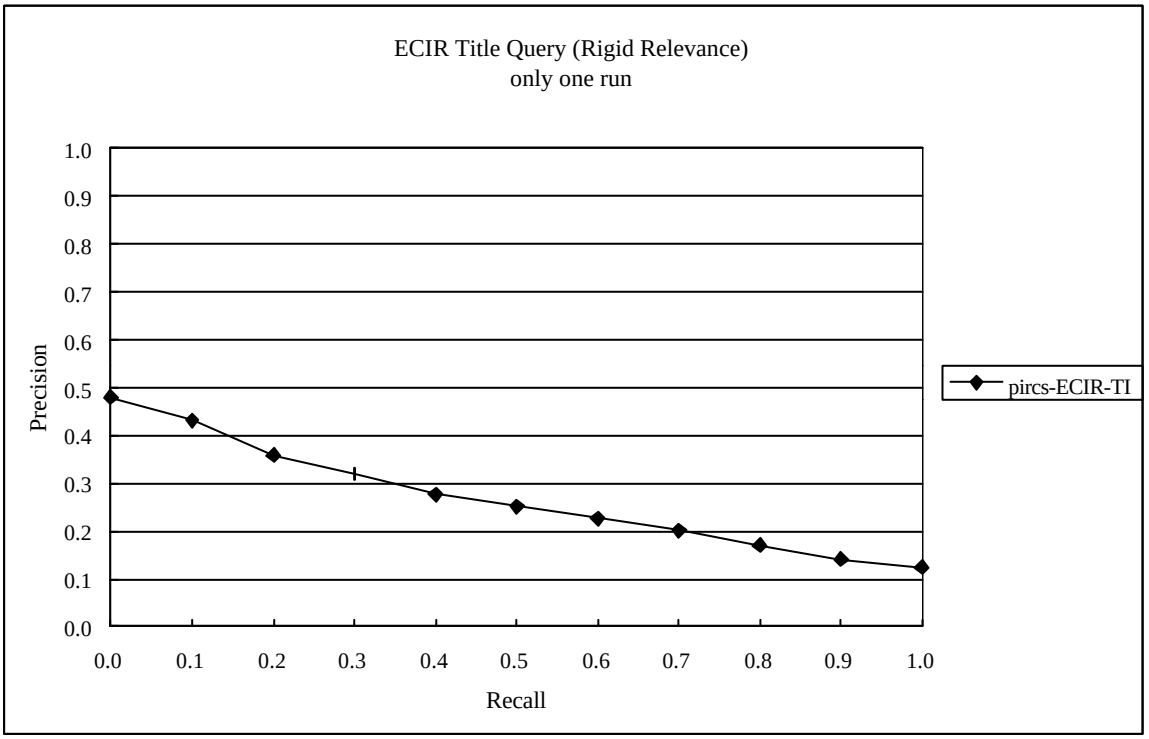

Figure 21. ECIR Title Query 11-Point Precision (Rigid Relevance)

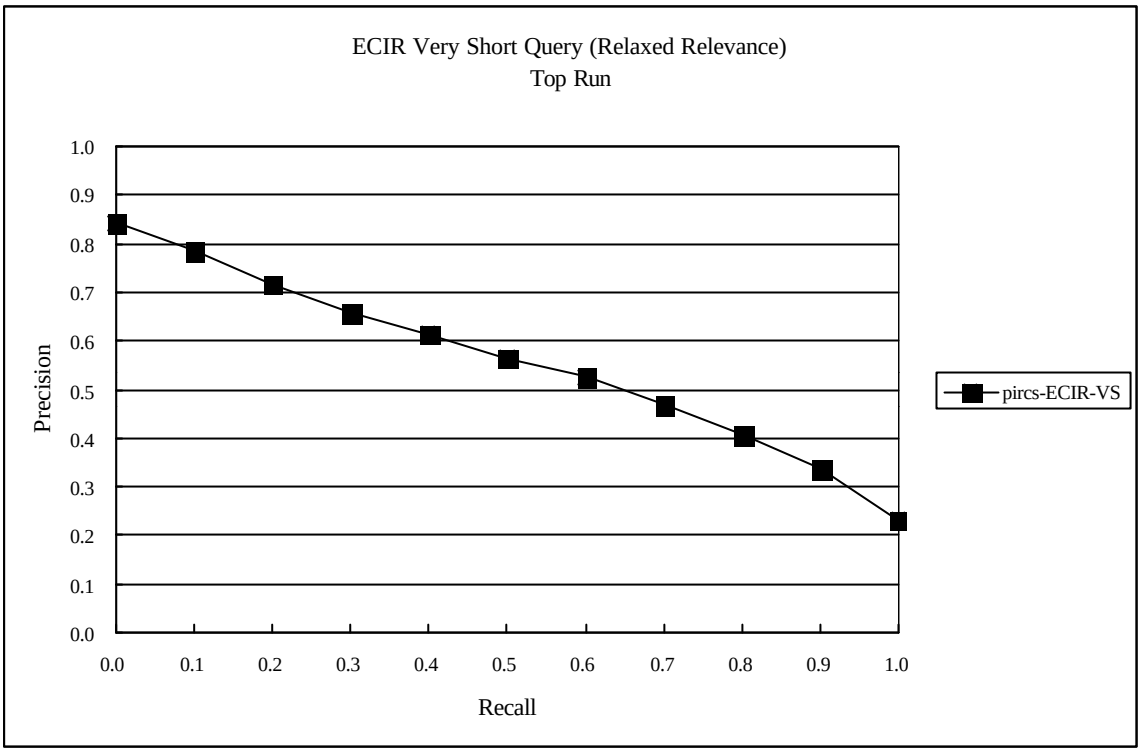

Figure 22. ECIR Very Short Query 11-Point Precision (Relaxed Relevance)

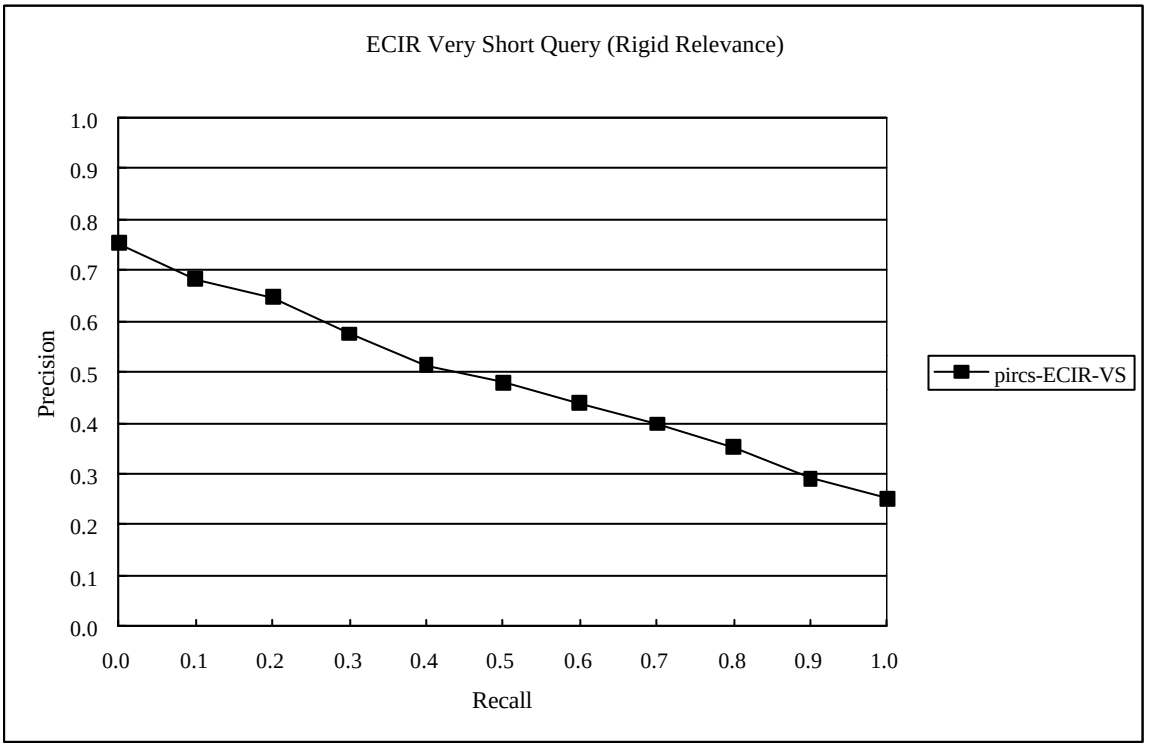

Figure 23. ECIR Very Short Query 11-Point Precision (Rigid Relevance)

Table 10. CHIR Task (Relaxed Relevance) Top Runs

RunID	IndexUnit	IndexTech	IndexStru	QueryUnit	IRModel	Ranking	QueryExpan
Brkly-CHIR-LO-01	bi-character	stopword	Inverted fille	bi-character	logistic regression	tf/idf/dl/ql/cl/cf	NO
CRL-CHIR-LO-06	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-LO-14	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-SO-02	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-SO-03	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
PIRCS-CHIR-SO	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ	tf/ ictf	top40doc+ 100term
PIRCS-CHIR-VS	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ	tf/ ictf	top40doc+ 100term
CRL-CHIR-VS-01	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-VS-03	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback

Table 11. CHIR Task (Rigid Relevance) Top Runs

RunID	IndexUnit	IndexTech	IndexStru	QueryUnit	IRModel	Ranking	QueryExpan
Brkly-CHIR-LO-01	bi-character	stopword	Inverted file	bi-character	logistic regression	tf/idf/dl/ql/cl/cf	NO
CRL-CHIR-LO-06	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-LO-14	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-SO-02	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-SO-03	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
PIRCS-CHIR-SO	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ	tf/ ictf	top40doc+ 100term
PIRCS-CHIR-VS	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ	tf/ ictf	top40doc+ 100term
CRL-CHIR-VS-02	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-VS-03	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback

Table 12. CHIR Long Query (Relaxed Relevance and Rigid Relevance) Top Runs

RunID	IndexUnit	IndexTech	IndexStru	QueryUnit	IRModel	Ranking	QueryExpan
Brkly-CHIR-LO-01	bi-character	stopword	Inverted file	bi-character	logistic regression	tf/idf/dl/ql/cl/cf	NO
CRL-CHIR-LO-06	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-LO-14	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
IOS-CHIR-LO-01	2-gram	2-gram	Inverted file	2-gram	Vector space	tf/idf	No
IOS-CHIR-LO-02	word	Stopword+dictionary	Inverted file	word	Vector space	tf/idf	No
NTHU_CHIR_LO_01	bi-word	morphology	Inverted file	word	vector space module	tf/idf	No
NTHU_CHIR_LO_02	bi-word	morphology	Inverted file	word	vector space module	tf/idf	No
PIRCS-CHIR-LO	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ	tf/ ictf	top40doc+ 100term
SSTUT-CHIR-LO-03	bigram	as is	suffix array	bigram	probabilistic model	tf, idf, burstiness,expansion frequency	Yes

Table 13. CHIR Short Query (Relaxed Relevance and Rigid Relevance) Top Runs

RunID	IndexUnit	IndexTech	IndexStru	QueryUnit	IRModel	Ranking	QueryExpan
CRL-CHIR-SO-02	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-SO-03	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
FRDC-CHIR-SO-01	character+bi-gram characters	dictionary	Inverted file	word	Vector space Mode	tf/idf, word association, document length	QE based on term contribute
IOS-CHIR-SO-01	2-gram	2-gram	Inverted file	2-gram	Vector space	tf/idf	No
IOS-CHIR-SO-02	word	Stopword+dictionary	Inverted file	word	Vector space	tf/idf	No
PIRCS-CHIR-SO	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ	tf/ ictf	top40doc+ 100term
SSTUT-CHIR-SO-01	bigram	as is	suffix array	bigram	probabilistic model	tf, idf, burstiness	No
SSTUT-CHIR-SO-02	bigram	as is	suffix array	bigram	probabilistic model with dynamic programming	tf, idf, burstiness	No
Trans-CHIR-SO	bi-word	No	inverted file	phrase	vector space model	tf	no

Table 14. CHIR Title Query (Relaxed Relevance) Top Runs

RunID	IndexUnit	IndexTech	IndexStru	QueryUnit	IRModel	Ranking	QueryExpan
Brkly-CHIR-TI-01	bi-character	stopword	Inverted file	bi-character	Logistic regression	tf/idf/dl/ql/cl/cf	No
CRL-CHIR-TI-02	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-TI-03	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
HKPU1-CHIR-TI-01	Hybrid (Word + Bigram)	Stop word + dictionary	Inverted File	Hybrid (Word + Bigram)	Vector Space	Tf/idf	No
HKPU1-CHIR-TI-02	Hybrid (Word + Bigram)	Stop word + dictionary	Inverted File	Hybrid (Word + Bigram)	Vector Space	Tf/idf + Term Length	No
PIRCS-CHIR-TI	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ	tf/ ictf	top40doc+ 100term
IOS-CHIR-TI-01	2-gram	2-gram	Inverted file	2-gram	Vector space	tf/idf	No
IOS-CHIR-TI-02	word	Stopword + dictionary	Inverted file	word	Vector space	tf/idf	No

Table 15. CHIR Title Query (Rigid Relevance) Top Runs

RunID	IndexUnit	IndexTech	IndexStru	QueryUnit	IRModel	Ranking	QueryExpan
Brkly-CHIR-TI-01	bi-character	stopword	Inverted file	bi-character	Logistic regression	tf/idf/dl/ql/cl/cf	No
CRL-CHIR-TI-02	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-TI-03	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
HKPU1-CHIR-TI-02	Hybrid (Word + Bigram)	Stop word + dictionary	Inverted File	Hybrid (Word + Bigram)	Vector Space	Tf/idf + Term Length	No
HKPU1-CHIR-TI-11	Word	Stop word + dictionary	Inverted File	Word	Vector Space	Tf/idf	No
PIRCS-CHIR-TI	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ	tf/ ictf	top40doc+ 100term
IOS-CHIR-TI-01	2-gram	2-gram	Inverted file	2-gram	Vector space	tf/idf	No
IOS-CHIR-TI-02	word	Stopword+dictionary	Inverted file	word	Vector space	tf/idf	No

Table 16. CHIR Very Short Query (Relaxed Relevance) Top Runs

RunID	IndexUnit	IndexTech	IndexStru	QueryUnit	IRModel	Ranking	QueryExpan
CRL-CHIR-VS-01	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-VS-02	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-VS-03	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
HKPU1-CHIR-VS-03	Hybrid (Word + Bigram)	Stop word + dictionary	Inverted File	Hybrid (Word + Bigram)	Vector Space	Tf/idf + Term Length	No
HKPU1-CHIR-VS-21	Bigram	Stop word + dictionary	Inverted File	Bigram	Vector Space	Tf/idf	No
IOS-CHIR-VS-01	2-gram	2-gram	Inverted file	2-gram	Vector space	tf/idf	No
IOS-CHIR-VS-02	word	Stopword+dictionary	Inverted file	word	Vector space	tf/idf	No
PIRCS-CHIR-VS	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ	tf/ ictf	top40doc+ 100term

Table 17. CHIR Very Short Query (Rigid Relevance) Top Runs

RunID	IndexUnit	IndexTech	IndexStru	QueryUnit	IRModel	Ranking	QueryExpan
CRL-CHIR-VS-02	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-VS-03	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
IOS-CHIR-VS-01	2-gram	2-gram	Inverted file	2-gram	Vector space	tf/idf	No
IOS-CHIR-VS-02	word	Stopword+dictionary	Inverted file	word	Vector space	tf/idf	No
PIRCS-CHIR-VS	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ	tf/ ictf	top40doc+ 100term

Table 20. ECIR All Runs

RunID	IndexUnit	IndexTech	IndexStru	QueryUnit	IRModel	Ranking	QueryExpan	TransTech
Brkly-ECIR-LO-01	word	stopword+ dictionary	Inverted file	word	logistic regression	tf/idf/dl/ ql/cl/cf	NO	Dictionary-based, select two
IOS-ECIR-*	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
NTHU-ECIR-LO-01	bi-word	morphology	invertedfile	word	vector space module	tf/idf	No	dictionary-based, corpus-based
PIRCS-ECIR-*	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ + retrieval combination	tf/ ictf	top40doc+ 100term	bi-word list + MT software
SSTUT-ECIR-LO-01	all n-grams	as is	suffix array	word	probabilistic model	tf, idf, burstiness	No	dictionary-based handmade
SSTUT-ECIR-LO-02	all n-grams	as is	suffix array	word	probabilistic model with dynamic programming	tf, idf, burstiness	No	dictionary-based handmade
Trans-ECIR-SO	bi-word	No	inverted file	word	vector space model	tf	no	Dictionary-based and corpus-based,select-top-1
UMD-ECIR-LO-01	overlapping character bigram	hexicificati on of Chinese characters	inverted file	within word overlapping character bigram	probabilistic model	tf/idf	no	dictionary-based, select-all
UMD-ECIR-LO-02	overlapping character bigram	Chinese character hexifying	inverted file	within word overlapping character bigram	probabilistic model	tf/idf	no	dictionary-based, select-top-3
UMD-ECIR-LO-03	word	Chinese character hexifying	inverted file	word	probabilistic model	tf/idf	no	dictionary-based, select-top-3