CLIR at NTICR

Noriko Kando
National Institute of Informatics (NII), Japan
http://research.nii.ac.jp/ntcir/
CLEF 2002
Sept. 19, 2002

NTCIR Workshop is :

- A series of evaluation workshops designed to enhance research in information access technologies by providing large-scale re-usable testcollections and a forum of research groups.

1st workshop:Nov.1,'98- Sept.1,'99.

2nd workshop:June,2000–March,2001

3rd workshop:Sept 2001-Oct 2002

http://research.nii.ac.jp/ntcir/

Tasks NTCIR Workshop 3

(1) CLIR [23] Chinese, English, Japanese, Korean

(2) Patent IR [11]

(3) Question Answering Challenge [16]*

(4)Text Summarization Challenge [7]

(5) Web [9]

[# of active participants]

Participants of NTCIR WS 3

Chungnam Nat Univ / ETRI+
CMU CLAIR
CRL+ (3)
Fu Jen Catholic Univ.
Hitachi CRL
HongKong PolyTech
Hummingbird*
Institute of Software Chinese Academy of Sciences+
JHU/APL
JUSTSYSTEM*
Keio Univ (2)
Kent Ridge Digital Lab+
Korea Univ
MEI MSRL*
Micorsoft Res Asia*
Mie Univ
NAIST
NAIST-CRL+
Nat Taiwan Univ
NEC Kansai*
NEC MRL*
NTT Data Tech*
NTTCS*
NTTCS*+NAIST
NTTDATA*
NYU/CRL+ (2)
Oki Electric Co.*
OsakaKyoikuUniv (3)
POSTECH(2)
QCCUNY
RICOH*
Ritsumeikan Univ (2)
SICS+
Surugadai Univ
Thomson Legal & Regulatory*
Tianjin University
Titech
Tokai Univ/Beijin Japan Center
Toshiba Corp* (2)
TUT (6)
ULIS-AIST+
Univ Aizu - OASIS (2)
Univ Kanagawa
Univ Kochi RAIK
Univ of California Berkeley (2)
Univ of Tokyo
Univ of Lib and Info Sci (2)
Univ of Tokyo/RICOH*
Waterloo Univ
Yokohama Nat Univ (2)

63 groups from 10 countries

Forcus of NTCIR

- Information Access: technologies to utilize info in the huge collection of documents

IR

Asian Languages

Variety of Genre

Parallel/comparable Corpus

New Challenges

Intersection of IR + NLP

Term Extraction, Summarization, QA, etc

Realistic eval/user task

Forum of Researchers

Idea Exchange

Discussion/ Investigation on Evaluation methods/metrics

Brief History

Project start on 1997

NTCIR WS 1 (Nov.,1998～Sept.,1999)

Sept1999: IREX joined>> Summarization, QA

Nov.1999: Int'l collaboration>> Asian lang CLIR

Apr.2000: RCIR/NII: permanent host

NTCIR WS 2 (June, 2000～March,2001)

NTCIR WS 3 (Oct.,2001～Oct.,2002)

Tasks in the Previous Workshops
[# of participants]

NTCIR WS 1 28 groups from 6 countries

(1) Ad Hoc IR [18]

(2) CLIR [10]

(3) Term Extraction & Role Analysis [9]

NTCIR WS 2 36 groups from 8 countries

(1) Chinese Text Retrieval [11]

(2) Japanese-English IR [25]

(3) Text Summarization [9]

Test collections constructed

Collection	task	document			topic/sum		rel
		genre	size	lng	#	lang	grd
NTCIR-1	IR	sci.abstracts	577MB	JE	83	J	3
CIRB010	IR	news articles	200MB	C	50	CE	4
NTCIR-2	IR	sci.abstracts	800MB	JE	49	J	4
NTCIR-2SUMM	sum	news articles	180 doc	J	1260	J	
NTCIR-2TAO	sum	news articles	1000 doc	J	2000	J	
KEIB010	IR	news articles	74MB	K	30	KCJE	4
CIRB011,020 NTCIR-3CLIR	IR	news articles	870MB	CJE	50	KCJE	4
NTCIR-3PAT	IR	Patent full. Patent abst.	17GB 4GB	J JE	31	KCCJE	3
NTCIR-3QA	QA	news articles	282MB	J	200*	J	2
NTCIR-3SUMM	sum	news articles	60doc	J	1260	J	
NTCIR-3Web	IR	Web docs	100GB	J(E)	110	J	5

Chinese, English, Japanese, Korean *200+800

Analysis of Test Collections/results

(A) exhaustivity of the document pool; effect of unique contribution

(B) inter-analyst consistency and its effect for system evaluation

(C) topic-by-topic evaluation

(D) estimate difficulty of topics

(E) effect of segmentation for search effectiveness

(F) evaluation metrics: nature of MAP, correlation of trec_eval measures with multigrade judgments; stableness of metrics; significance test as paired samples, metrics for multigrade judgments, metcis for TS, QA, Web etc.

(G) effect of reuse of training data for test (under analysis)

CLIR at Asian Environment

1. **Initial Stage : English & Own language**

   internationalize = provide info in English

2. **Long (2000 years?) historical relationship**, but less interaction in 1950-early 1990s

3. **Interest getting growing rapidly-**

   Commercial/industrial exchange increased: ex. patent import

   Cultural/Social interest, Human Exchange, Travel, TV, music, magazines, etc, especially in younger generation:

   ex. Taiwanese tea service

   Korean cuisine

   Ha-Li = Japanese lovers in Taiwan

It is the time to start !!

4. **Languages structures are completely different**
   Character codes are different, Chinese characters are different each other

Quite Challenging

CLIR at Previous Workshops

NTCIR WS 1: 28 groups from 6 countries

(1) Ad Hoc IR, (2) CLIR

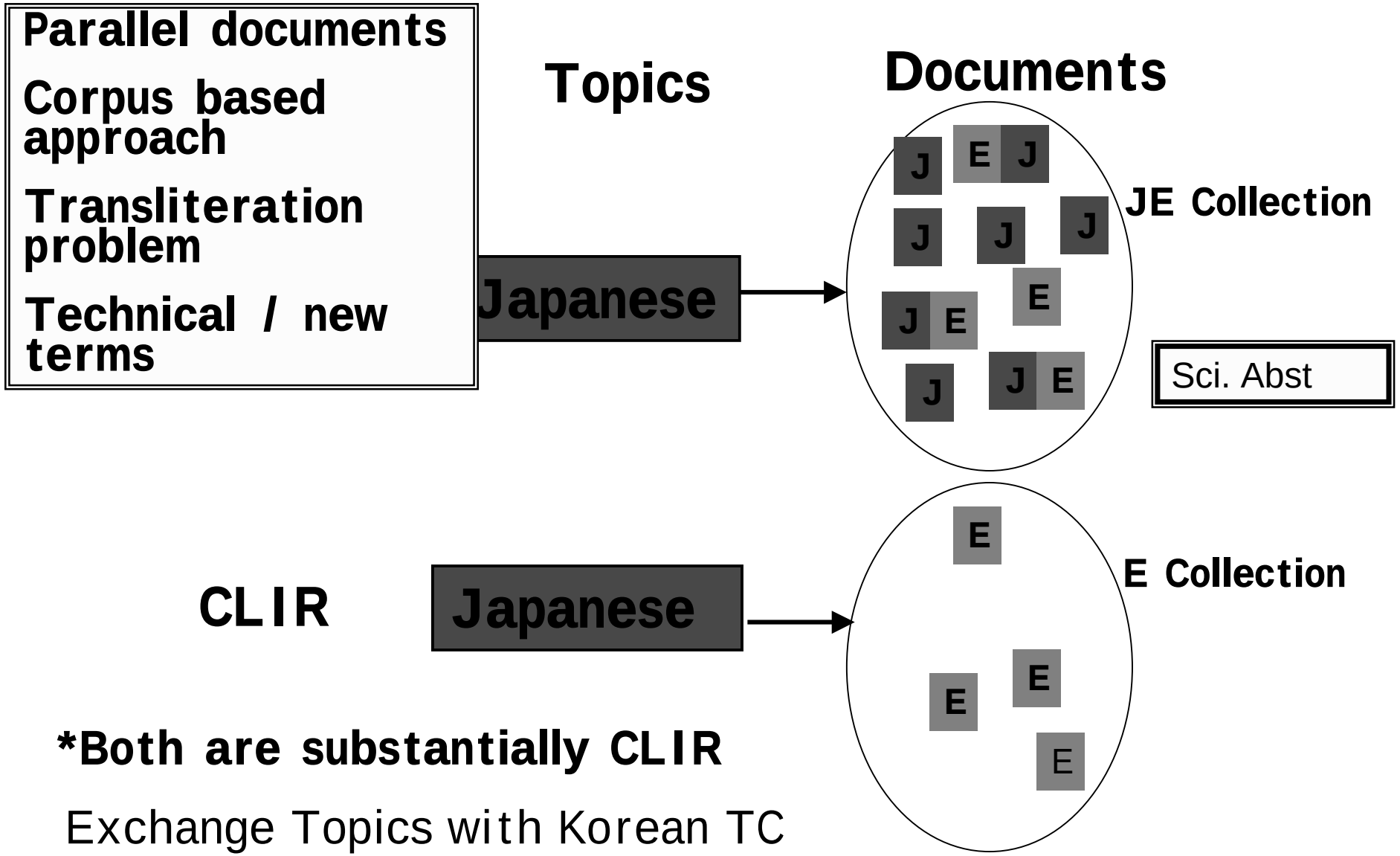
NTCIR WS 2: 36 groups from 8 countries

(1) Chinese Text Retrieval, (2) Japanese-English IR,

NTCIR WS 3: 63 groups from 10 countries

(1) CLIR, (2) Patent IR,

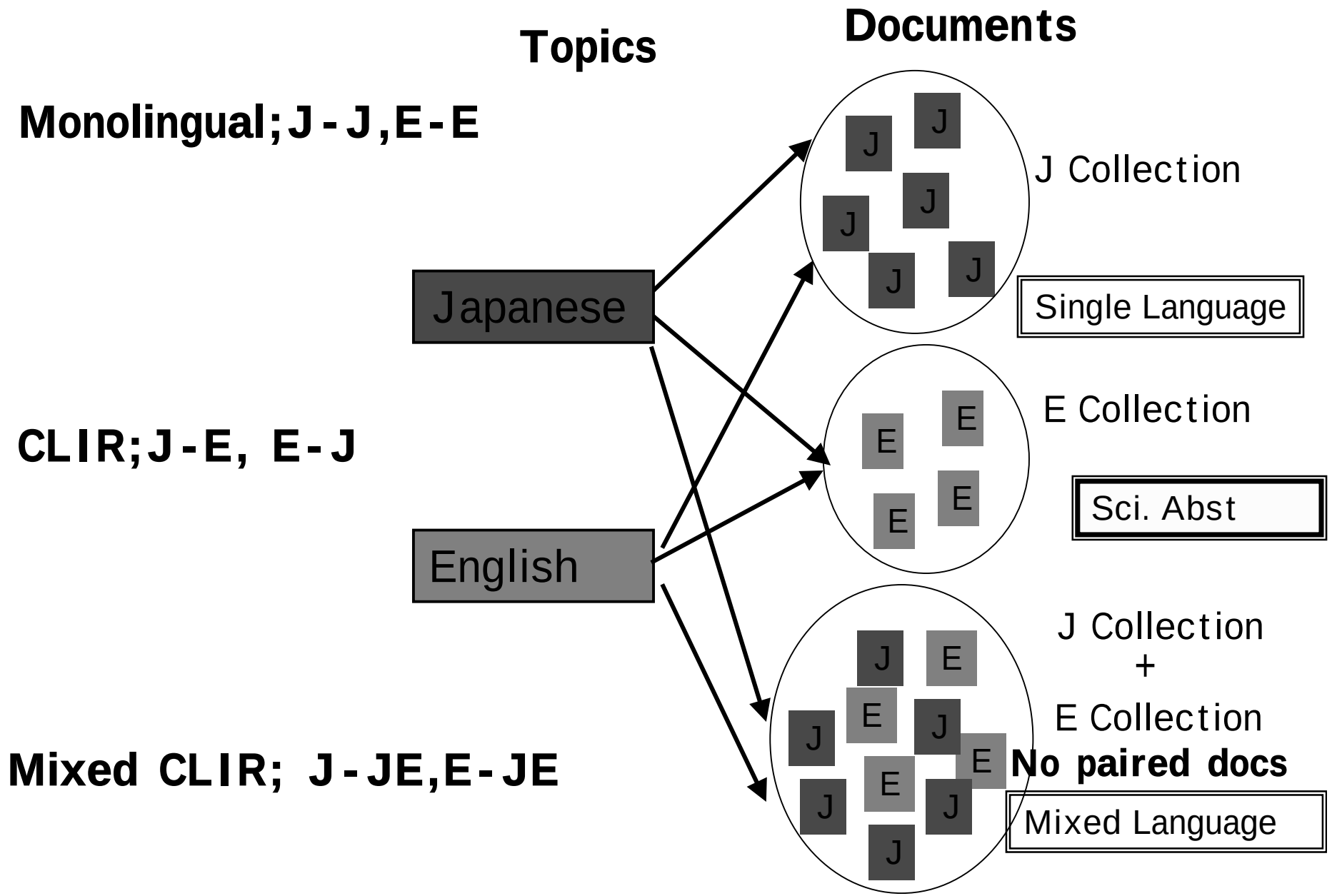
CLIR at NTCIR: **NTCIR WS 1 (1998/99)**



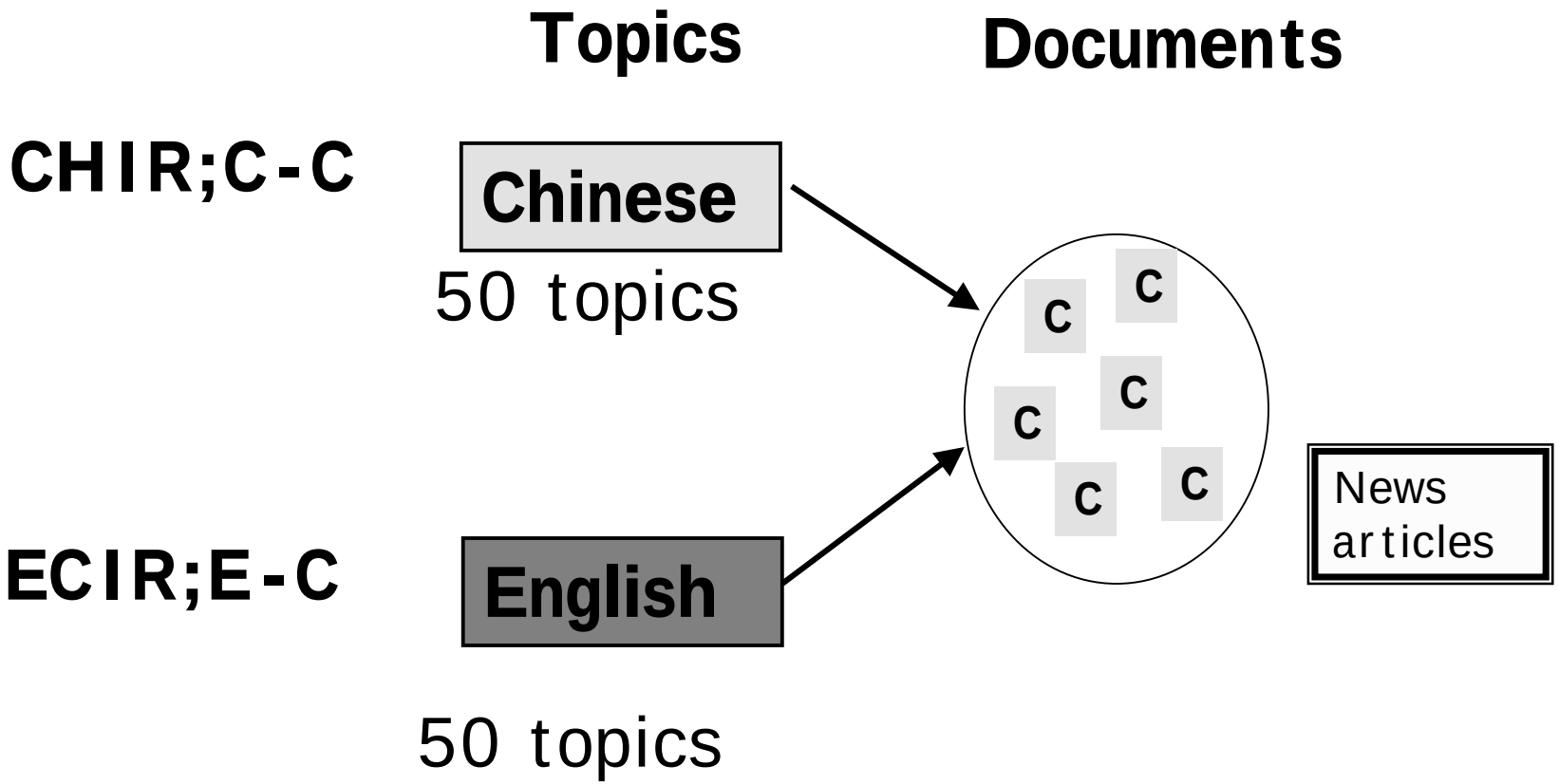
***Both are substantially CLIR**

Exchange Topics with Korean TC

NTCIR WS 2 Japanese & English

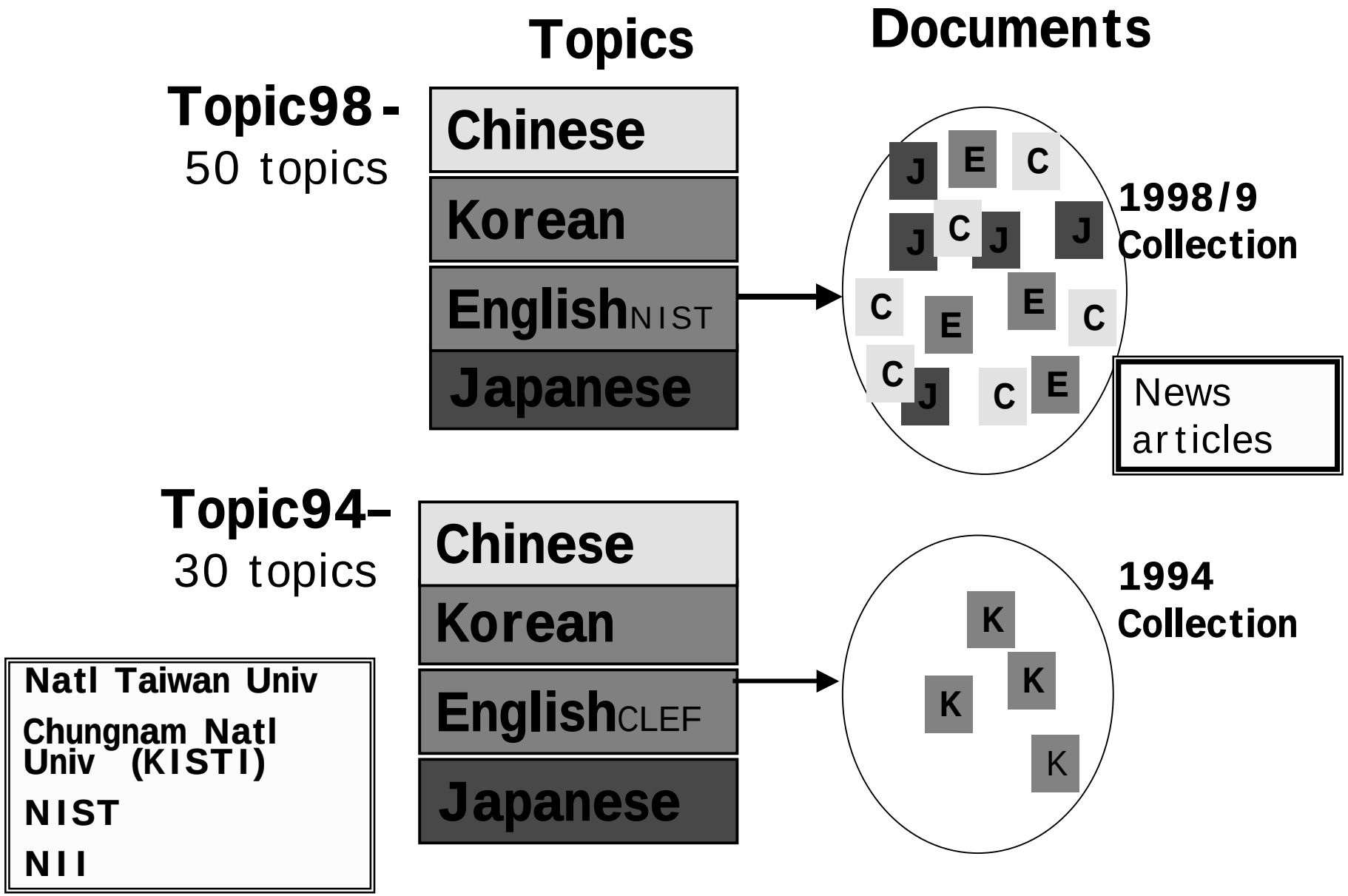


NTCIR WS 2 Chinese (CHIB010)



Organizers: Kuang-hua Chen, Hsin-Hsi Chen (National Taiwan University)

CLIR at NTCIR: **NTCIR WS 3 CLIR**



CLIR at NTCIR WS 3

TARGET: more realistic CLIR task

CHALLENGE:

Scientific -> News: ordinary life, cross-cultural/social issues, proper names

APPROACH:

try to keep balance of topics from each country, topics with local/proper names, etc.

CLIR at NTCIR WS 3

DOCUMENT:

Taiwan: 1998-1999

- CIRB011, Chinese, 132,173 docs
- CIRB020 (United Daily News), Chinese 249,508 docs
- Taiwan News/Chinatimes English News, English, 10,204 docs

Japan: 1998-1999

- Mainichi Newspaper, Japanese, 220,078 docs
- Mainichi Daily News, English, 12,723 docs

Korea: 1994

- Korean Economic Daily, Korean, 1994, 66,146 docs

TOPICS: Chinese, English, Japanese, Korean/ 2 sets

RELEVANCE JUDGMENTS: 4 grades

Sample topic

<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>

D run is mandatory

CLIR at NTCIR WS 3

SLIR (single language IR):

C-C 34 runs

K-K 17 runs

J-J 29 runs total 110 runs from 22 groups

BLIR (Bilingual CLIR: x- C, K, J):

E-C 16, E-J 11, E-K 6, C-J4, J-C 5, etc

total 50 runs from 14 groups

MLIR (Multilinaul CLIR: x- CEJ)

C-CEJ 4, E-CEJ 4, J-CEJ 3, etc

Total 29 runs from 7 groups

*Can submit “up to 3 runs per language pair”

MAP of mandatory top runs

C-C	**pircs-C-C-D-01**	**0.362**
	Brkly-C-C-D-01	**0.352**
	CRL-C-C-D-02	**0.345**
	APL-C-C-D-02	**0.340**
E-E	**TSB-E-E-D-01**	**0.465**
	Brkly-E-E-D-02	**0.437**
	APL-E-E-D-02	**0.389**
	HUM-E-E-D-01	**0.338**
J-J	**CRL-J-J-D-03**	**0.400**
	Brkly-J-J-D-01	**0.395**
	TSB-J-J-D03	**0.391**
	APL-J-J-D-02	**0.357**

C-C all, rigid

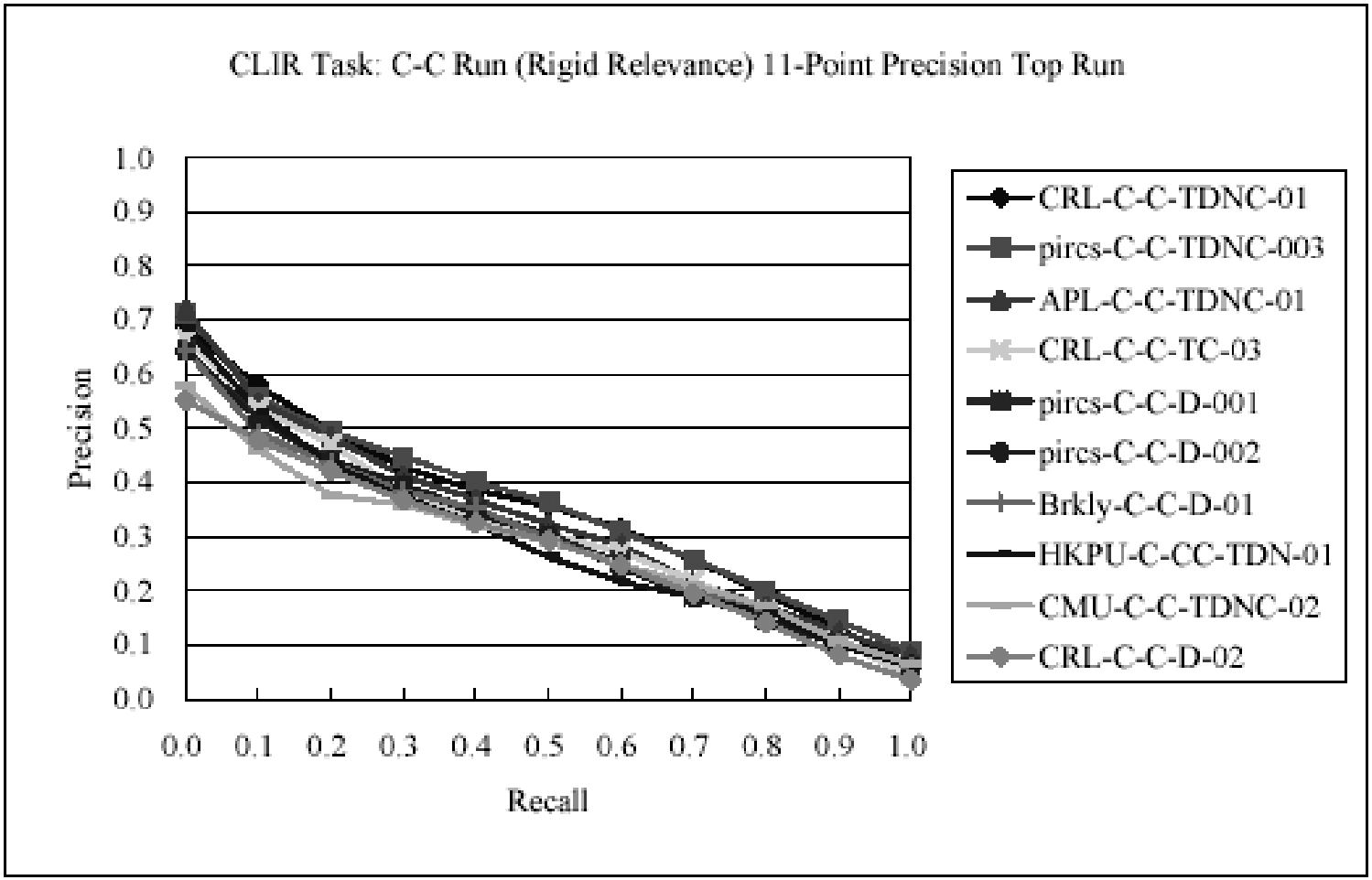


Figure 4. CLIR Task: C-C Run 11-Point Precision (Rigid Relevance)

E-E all, rigid

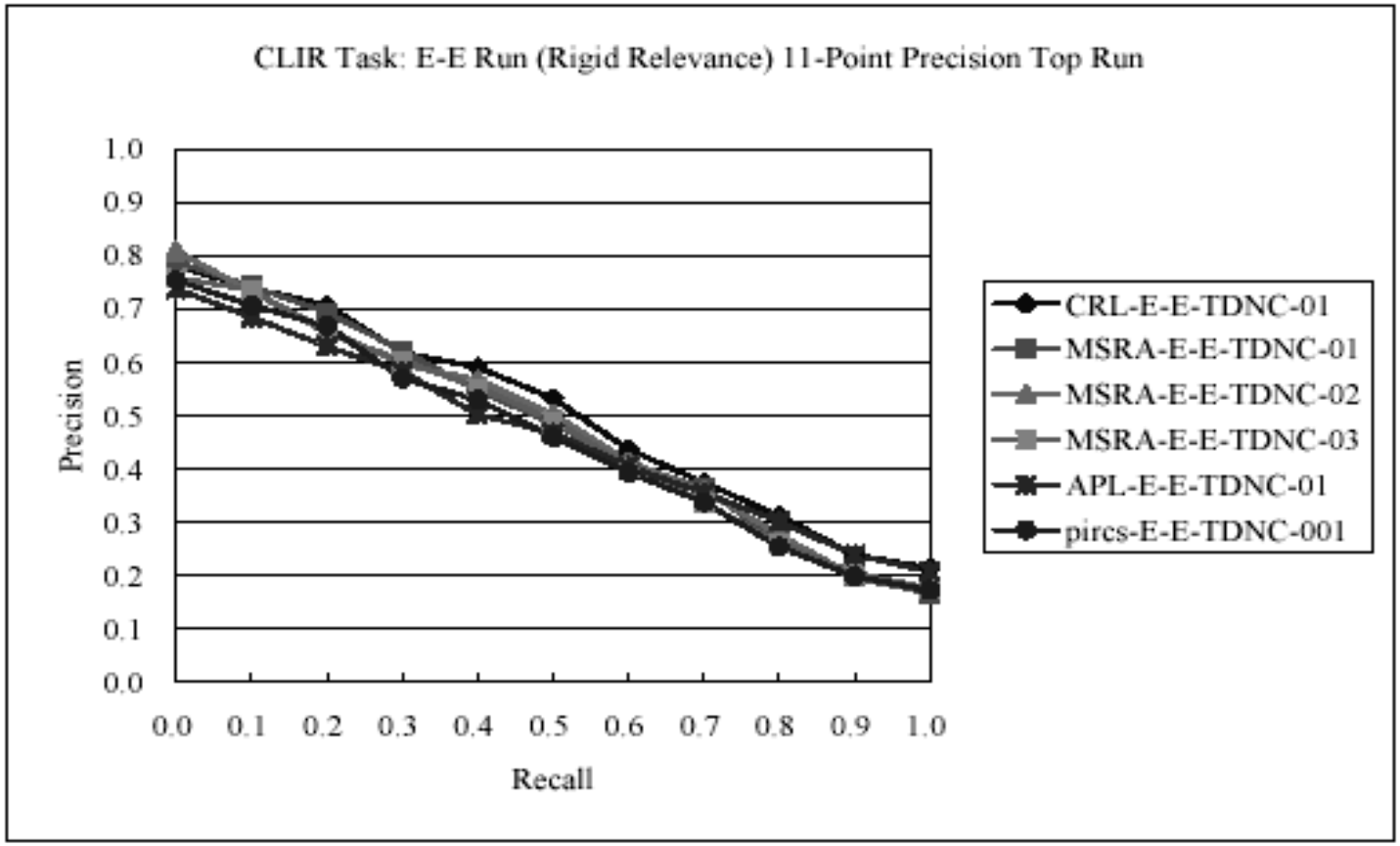


Figure 6. CLIR Task: E-E Run 11-Point Precision (Rigid Relevance)

J-J all, rigid

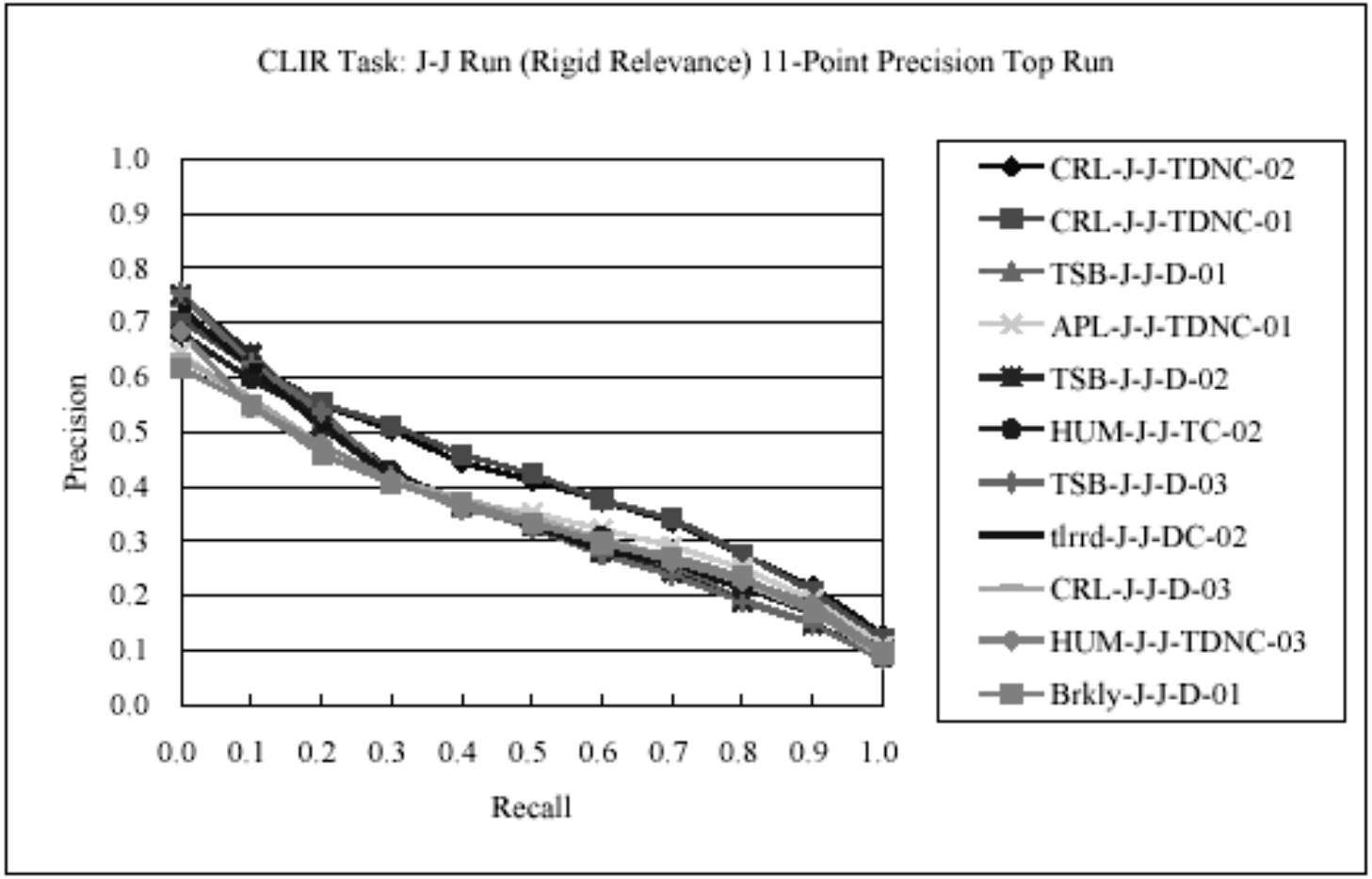


Figure 8. CLIR Task: J-J Run 11-Point Precision (Rigid Relevance)

MAP of mandatory top runs

K-K	**CRL-K-K-D-03**	**0.360**
	Brkly-K-K-D-01	**0.313**
	APL-K-K-D-02	**0.272**
	KUNLP-K-K-D-01	**0.269**

K-K all, rigid

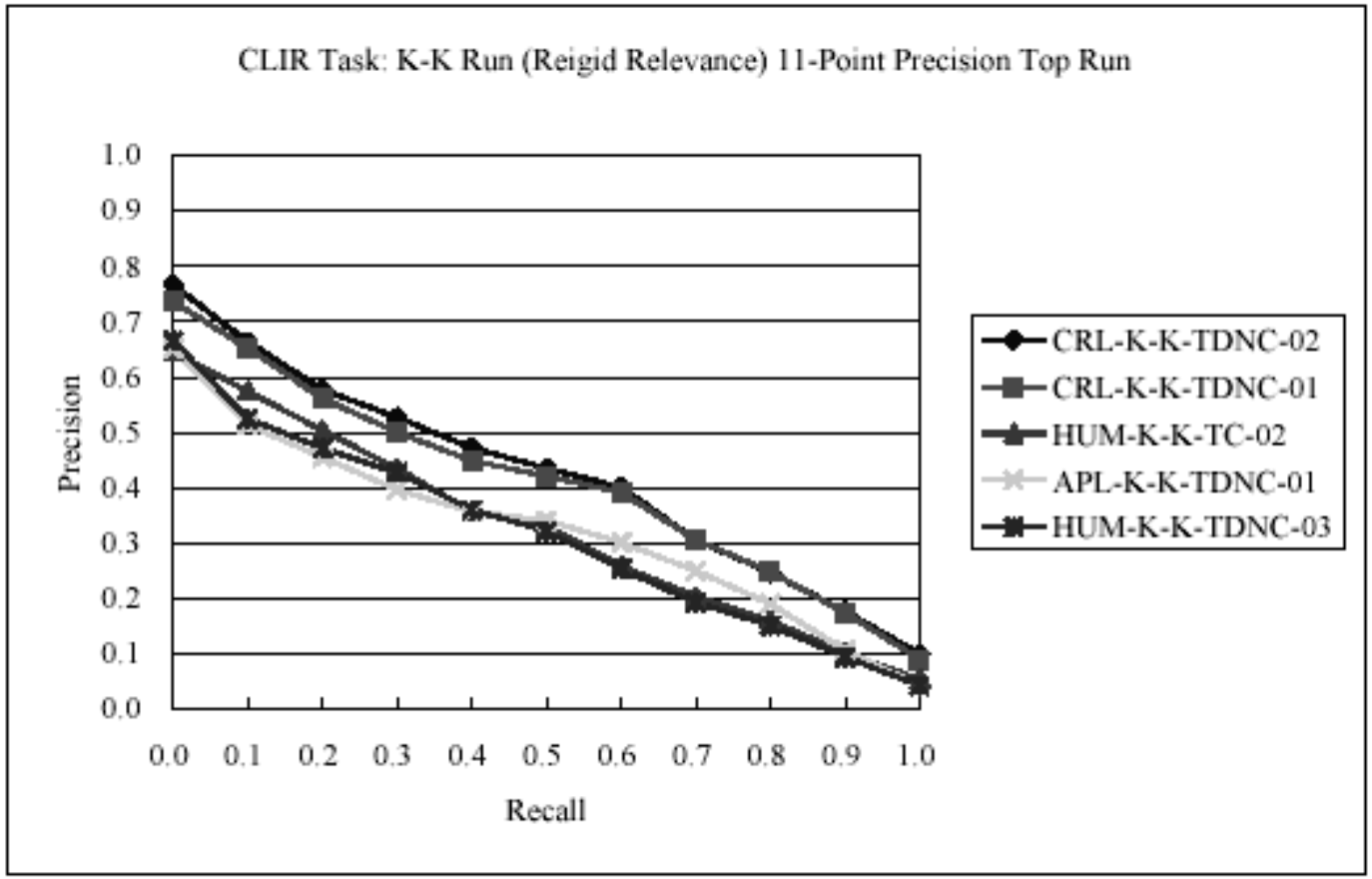


Figure 10. CLIR Task: K-K Run 11-Point Precision (Rigid Relevance)

Map of top mandatory runs

E-C	**MSRA-E-C-D-03**	**0.192**
	Brkly-E-C-D-01	**0.161**
	pircs-E-C-D-02	**0.159**
E-J	**TSB-E-J-D-02**	**0.340**
	Brkly-E-J-D-01	**0.221**
	APL-E-J-D-02	**0.111**
E-K	**KUNLP-E-K-D-03**	**0.199**
E-EC	**ISCAS-E-EC-D-03**	**0.165**
	pircs-E-EC-D-01	**0.162**

E-C all, rigid

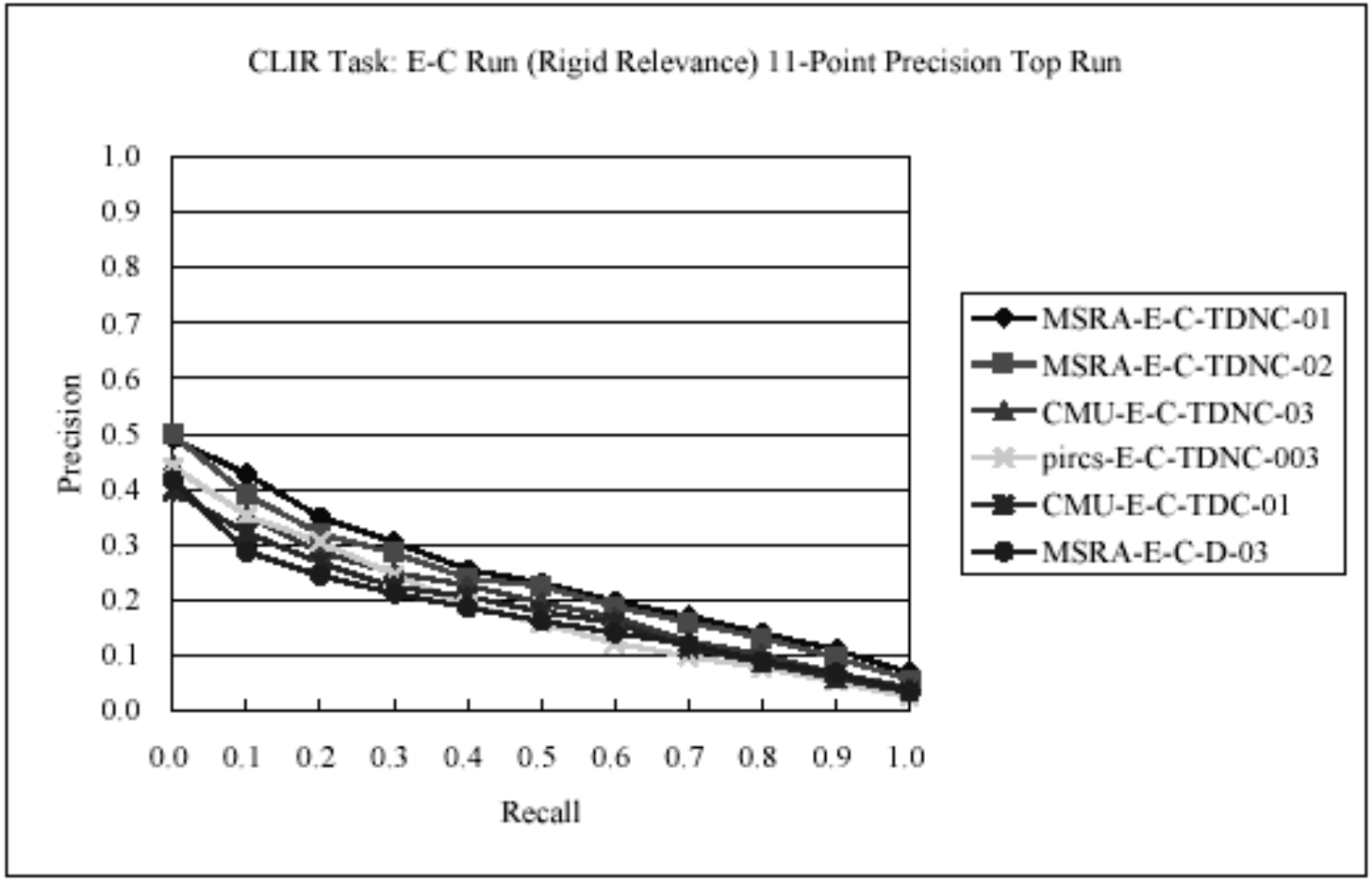


Figure 14. CLIR Task: E-C Run 11-Point Precision (Rigid Relevance)

E-J D-onlv. riaid

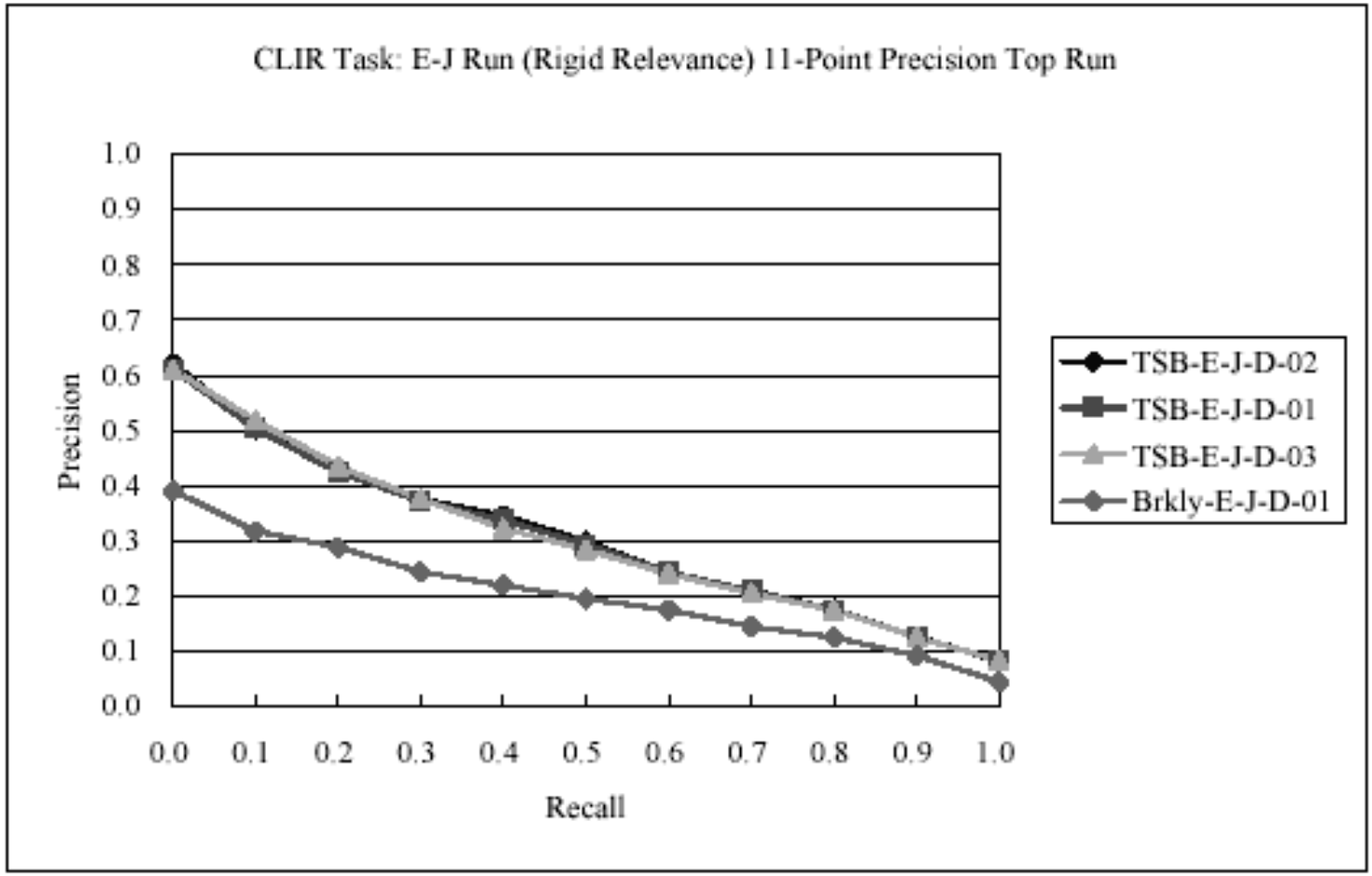


Figure 16. CLIR Task: E-J Run 11-Point Precision (Rigid Relevance)

C-J all, rigid

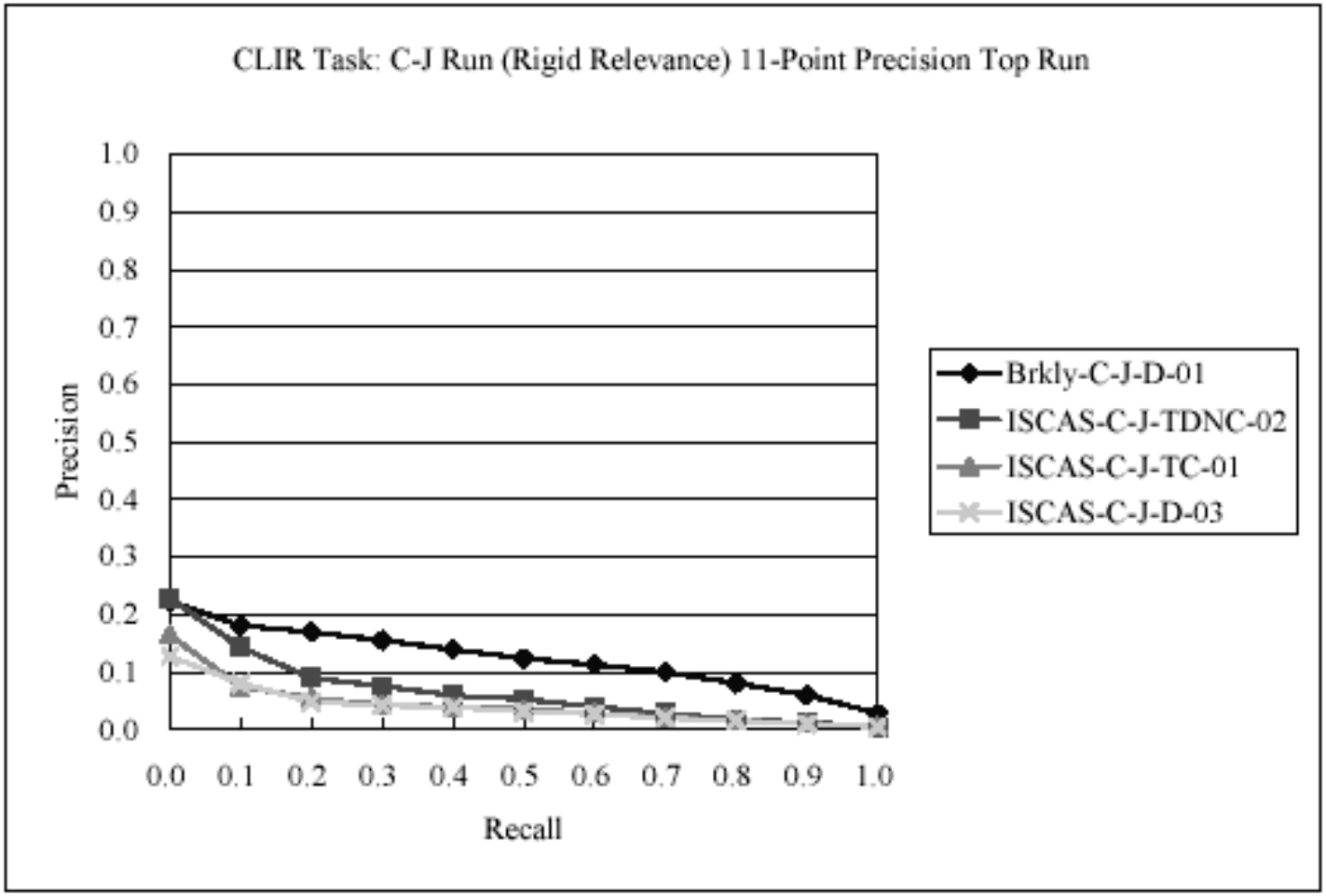


Figure 12. CLIR Task: C-J Run 11-Point Precision (Rigid Relevance)

J-C all, rigid

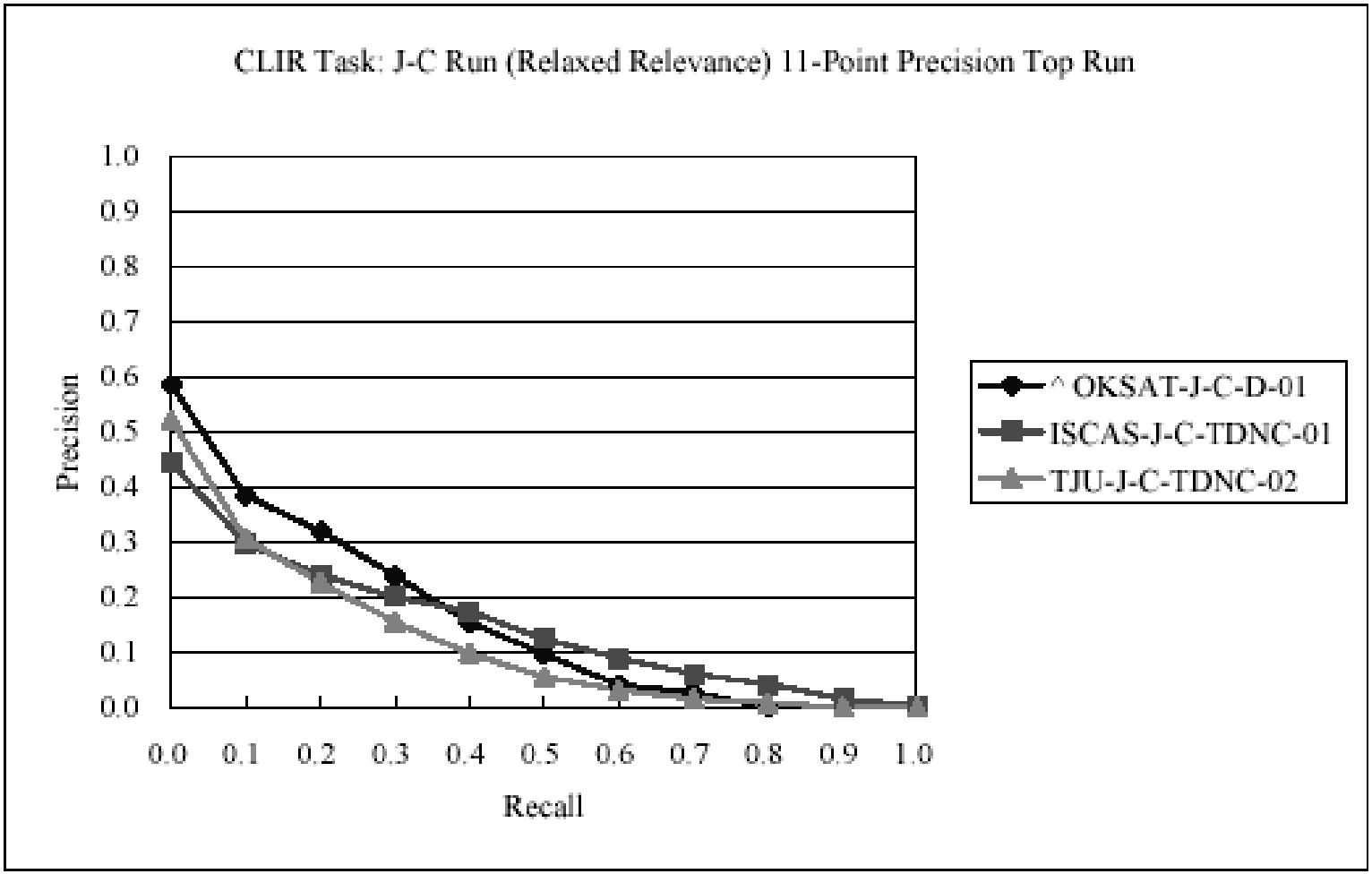


Figure 19. CLIR Task: J-C Run 11-Point Precision (Relaxed Relevance)

C-CJE all, rigid

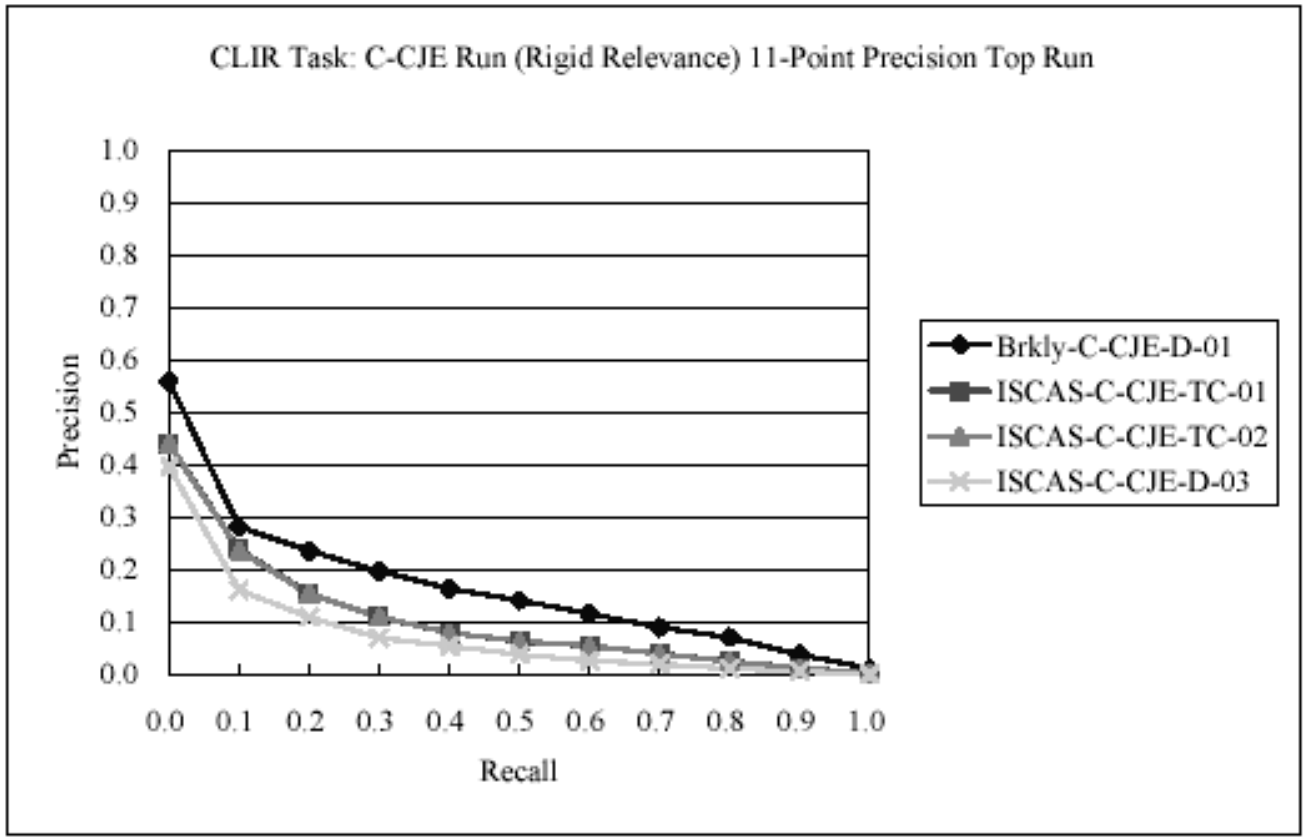


Figure 22. CLIR Task: C-CJE Run 11-Point Precision (Rigid Relevance)

Outcomes

- **Many monolingual & E-x Bilingual runs**
  - Many groups did all monolingual
- **Probablistic** models, **PRF**, Pre& post translation **query expansion**
  - HKPU : OKAPI, Pircs, logistic regression, VSM
  - HUM: doc length effect on VSM
- More MT/ Less corpus-base, but combination
  - Toshiba, CMU

Outcomes

- **Segmentation**:
  - Bi-gram for Chinese, Word-base indexing for Japanese
  - HKPU, Thompson,
- Topics in English, Japanese well investigated?
- Start of MLIR and BLIR among Asian languages
  - ISCAS, Brkly, NTU, Microsoft

Patent at NTCIR WS 3

TARGET:

Technology Survey, search patent from newspaper clip w/memo

CHALLENGE/CHARACTERISTICS:

Cross-Genre IR

long, structured documents, highly technical, new terminology, etc

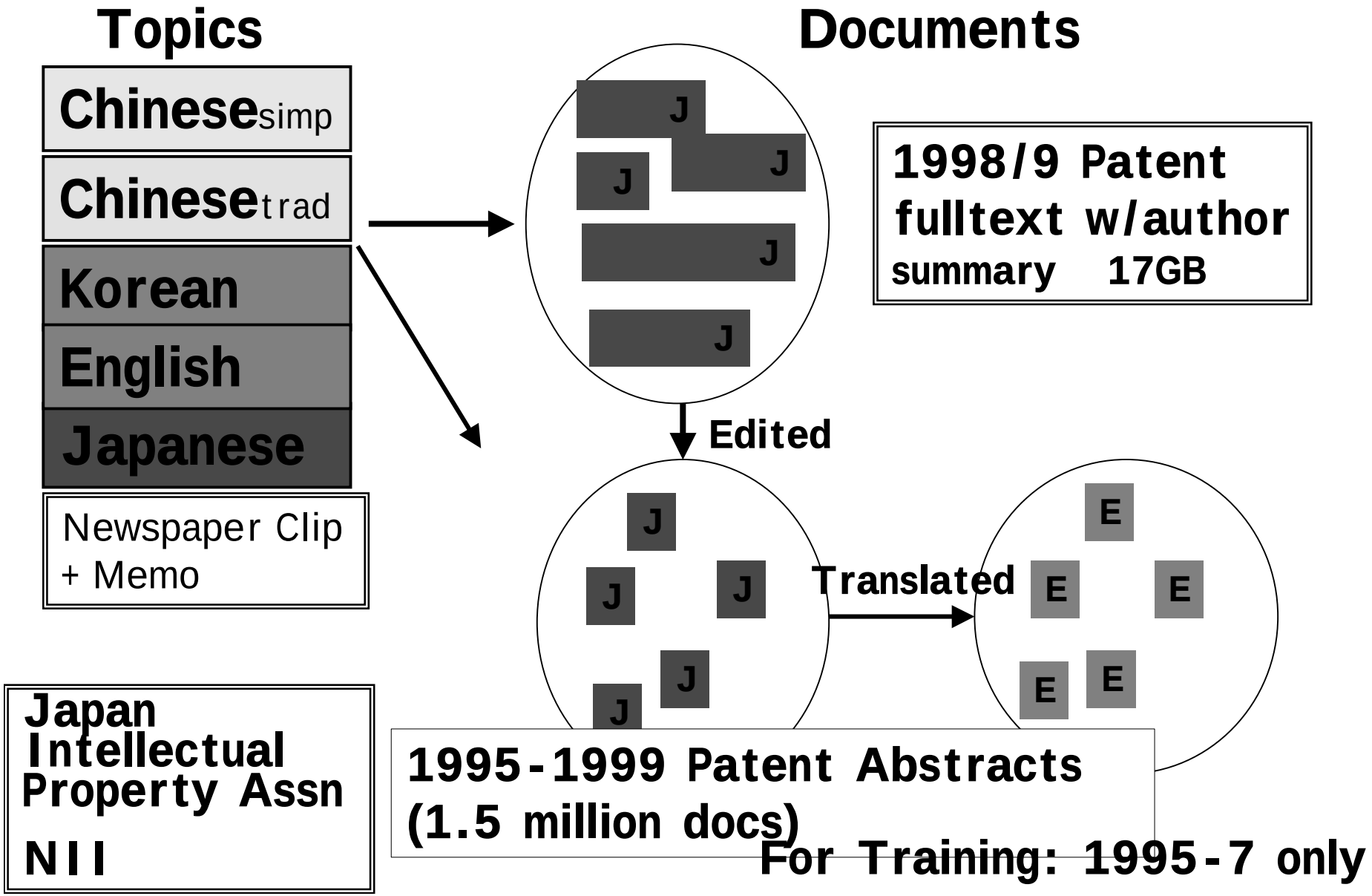
Large exactly translated English-Japanese Paird Docs

Classification Codes

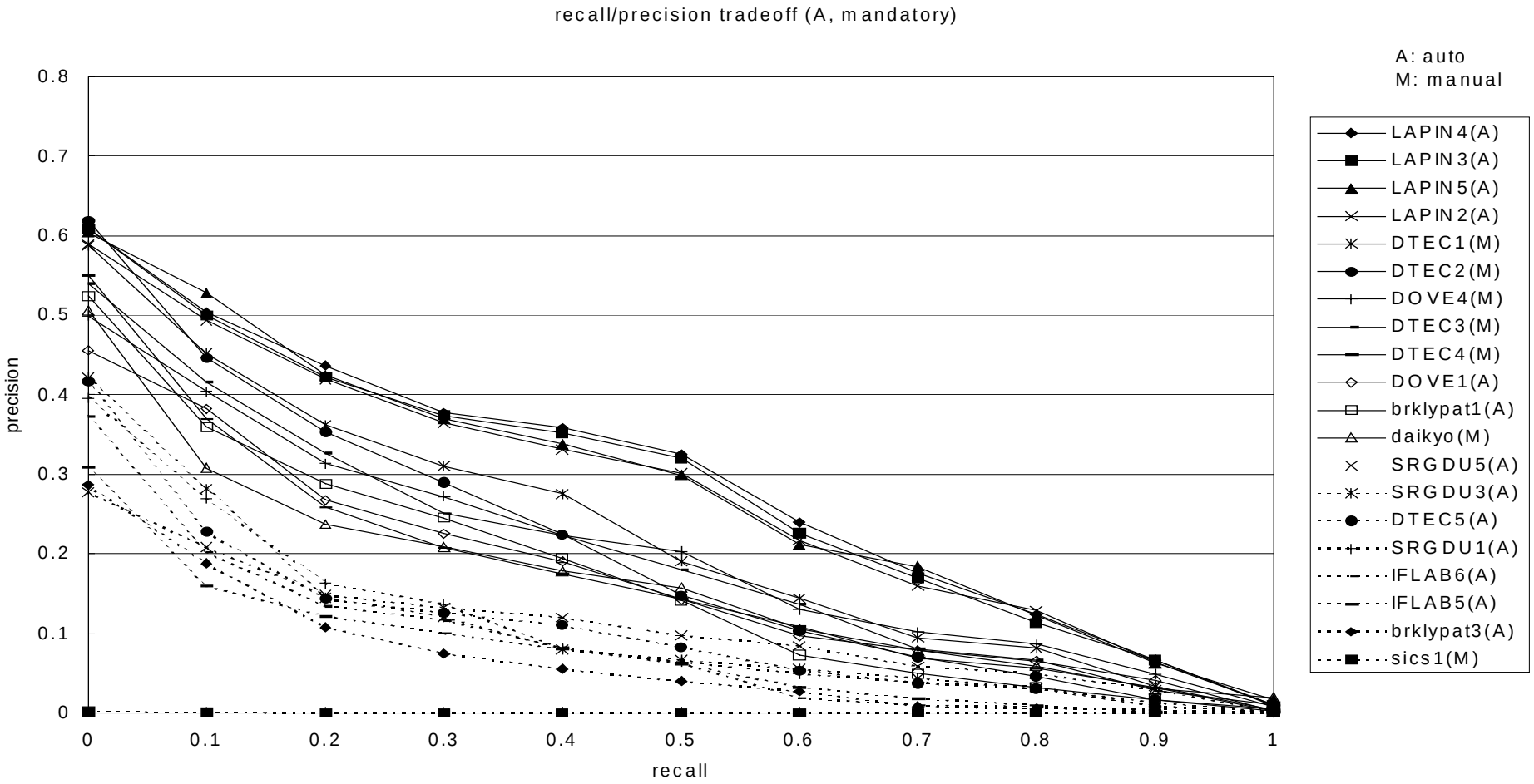
Attracts Company/"Veterans" Groups

PARALLELISM:

CLIR at NTCIR: **NTCIR WS 3 patent**

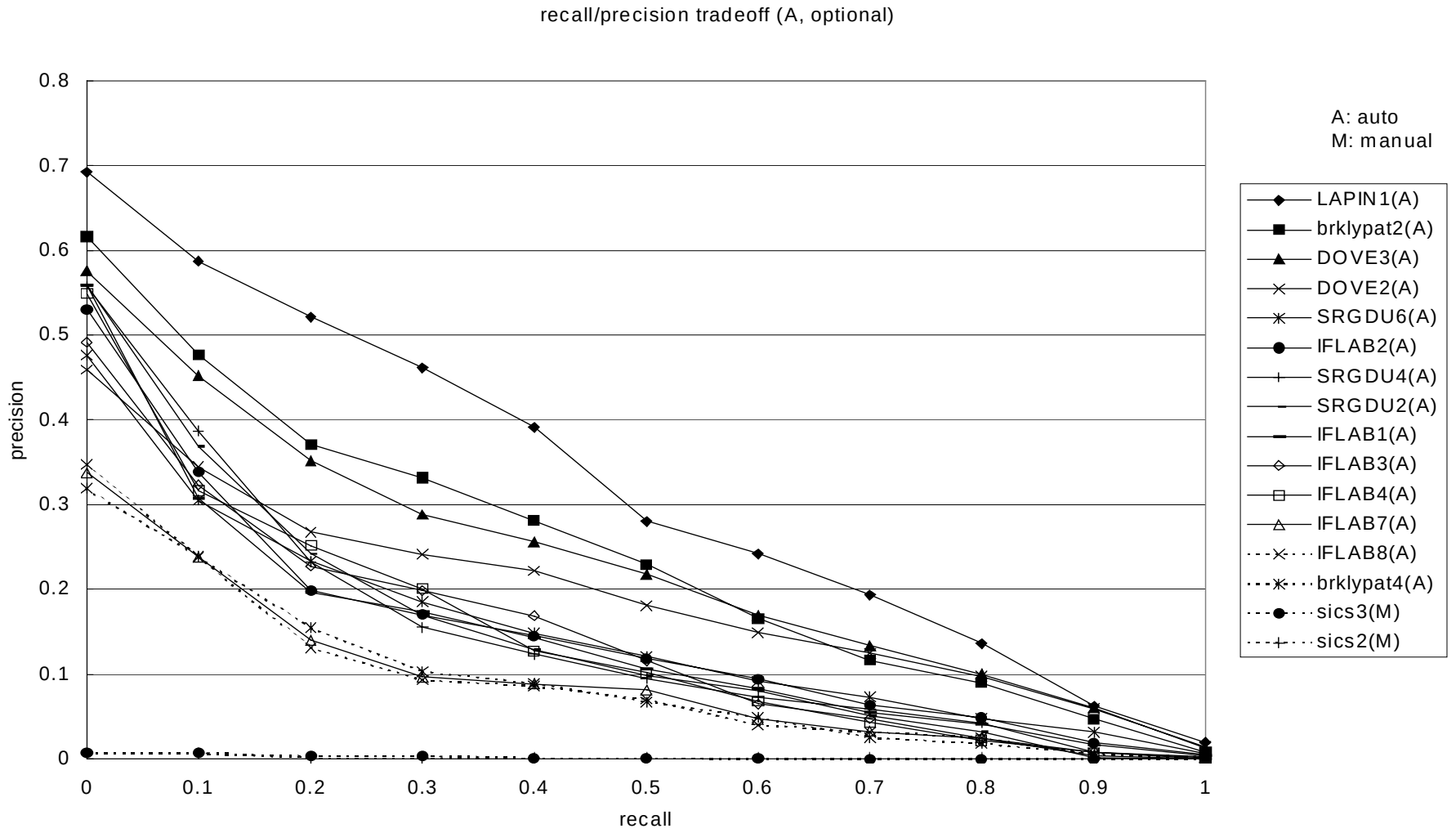


Patent mandatory

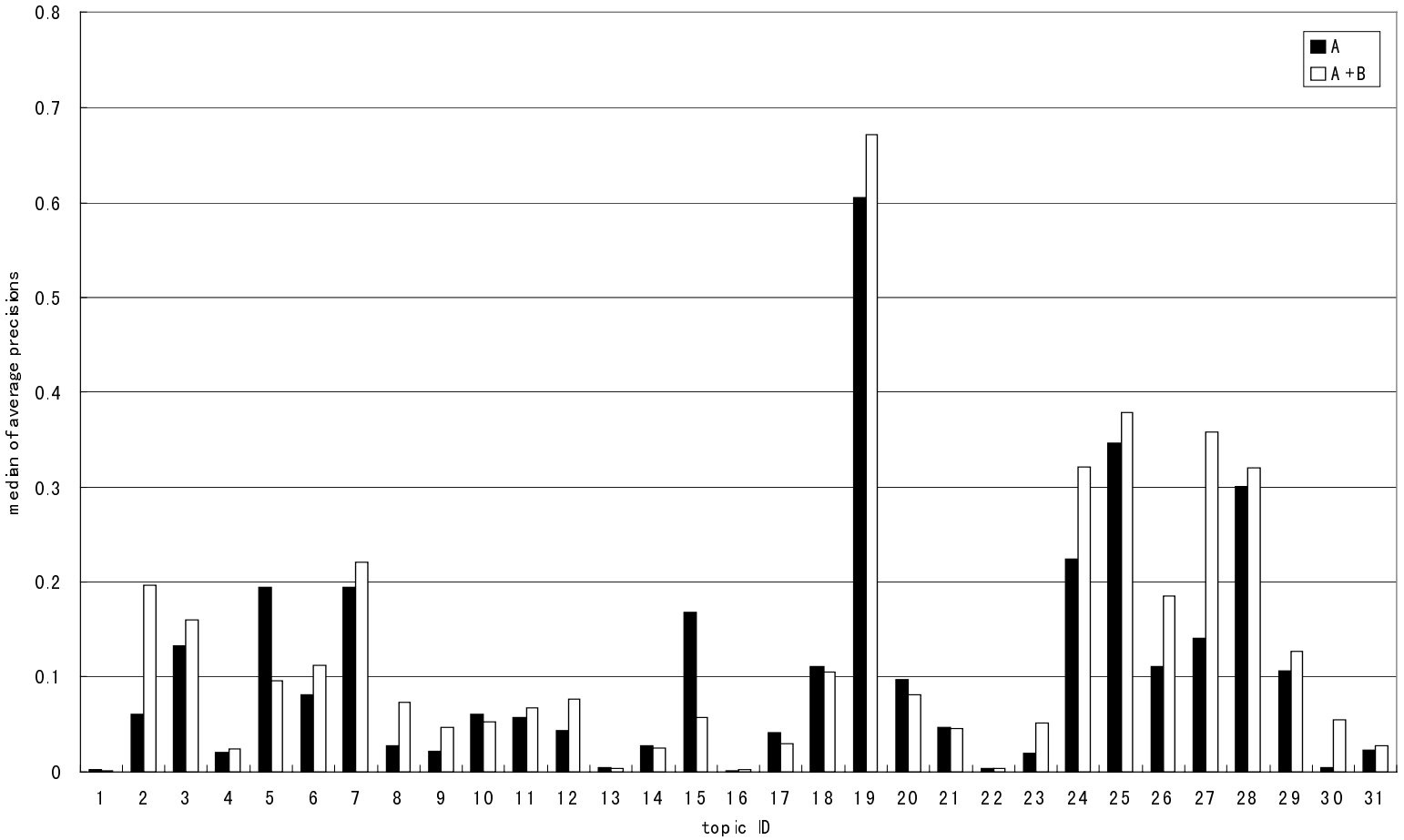


recall/precision tradeoff (A+B, mandatory)

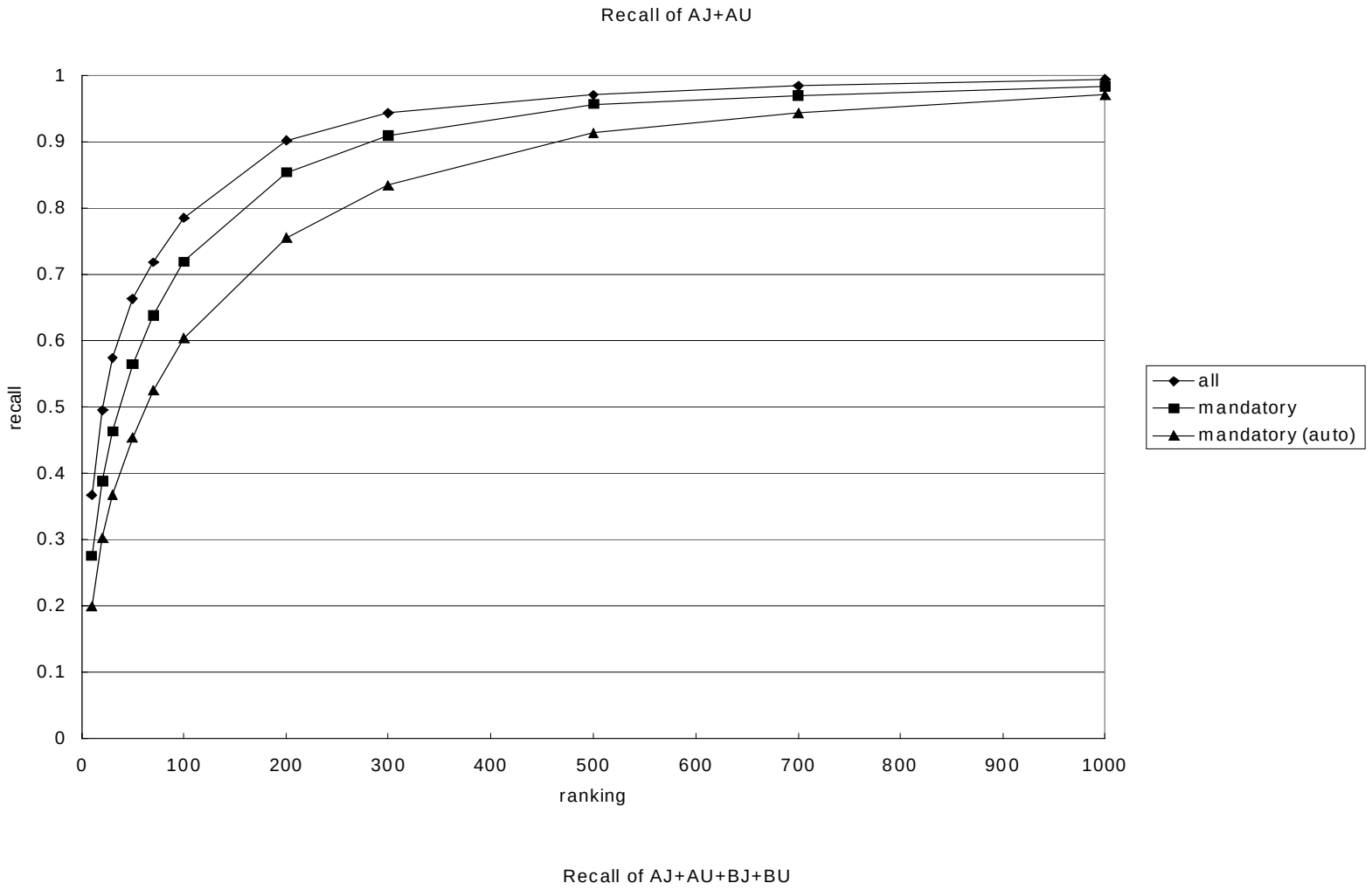
NTCIR WS 3 – Patent optional runs



NTCIR WS 3 – Patent topic by topic



Patent: human vs system



NTCIR WS 3 - QA

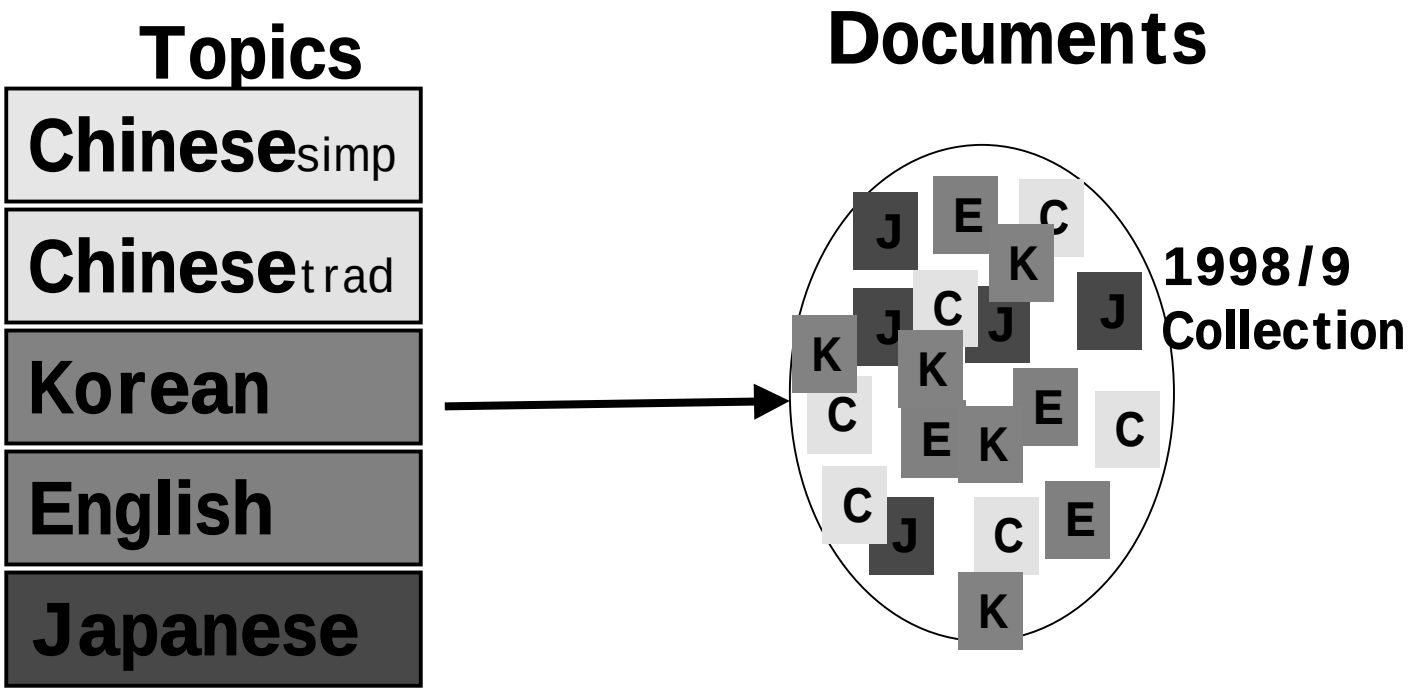
Task 1: Return 5 top answers (noun phrase) /w support information (relevant passage less than 100 characters), 100 Q

Task 2: wrong answer is penaltied. Return only correct answers. 100 Q. Support information is required.

Task 3: a series of questions. The related questions are given for the 30 of questions of Task 2.

Doc: J newspaper articles,1998-1999

NTCIR WS 4 CLIR



Future Directions -cont'd

- MORE Multilingual!
- Keep Monolingual
- Mandatory runs : <TITLE> runs
  - <DESC> is too long. Disambiguation techs did not test well on long queries.
- Pivot Language track
  - Chinese – E – x, Japanese – E – x, Korea – E - x
- CL QA ?


Thanks
Merci
Danke schön
Gracie
Gracias
Tack
Köszönöm
Kiitos
Terima Kasih
Khap Khun
Ahsante
Tak
謝謝
ありがとう

Future Directions

- What is the real needs for CLIR among East Asian languages?
  - Bilingual English – Own language is the initial interest and social needs
  - Mutual interest to other Asian countries increasing:
    - ex. FIFA World CUP, travel & human exchange increasing esp. among younger generation. >>New articles
    - Technical/industrial interest >> Patent

Axes to characterize CLIR systems

- Languages
- Type of media
- Tasks and users
- Relevance judgments or success criteria
- Document genres
- Layers of CLIR technologies
- Information access process

Tasks and purposes of the search,

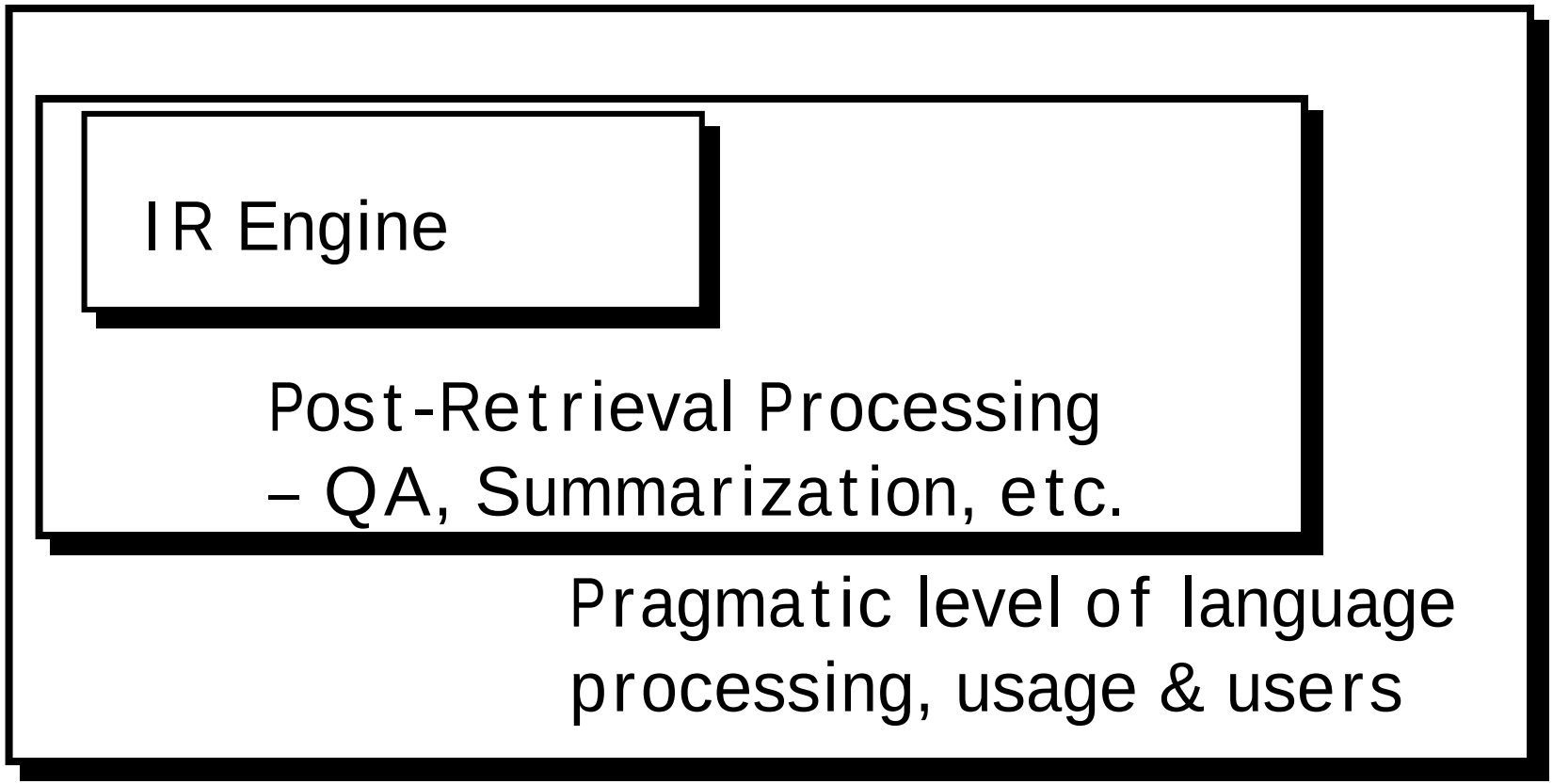
- important to set the scene of the experiments
- basis for the relevance judgment
- implication of improvements in the search effectiveness
- Document genres are often related to tasks or purpose of the search, the user communities, and success criteria.
- Traditionally IR --.> generalized systems

layers of CLIR technologies;

pragmatic layer: cultural & social aspects,
semantic layer: concept mapping
lexical layer: language identify, indexing
symbol layer: character codes
physical layer: network

- Related to tasks, relevance judgments, document genres

Evaluation of Information Access



NTCIR WS 3 – Patent-documents

- Japanese patents: 1998-1999 (about 17GB)
- abstracts (E&J): 1995-1999, ca. 1,750,000docs exactly translated-pairs.
- E-J abstracts 1995-97 are usable for translation models in mandatory runs.
- Abstracts 1998-99 are OK for additional runs both for translation models and retrieved docs.