# CLIR Evaluation at NTCIR: Japanese-English CLIR and its challenges

*Noriko Kando*

*National Institute of Informatics (NII), Japan*

*( formerly NACSIS)*

http://www.rd.nacsis.ac.jp/~ntcadm/

CLEF 2000

September 21, 2000

# NTCIR Workshop is :

- A series of evaluation workshops designed to enhance research in Japanese IR, cross-lingual IR and related area by providing large-scale Japanese test collections and a forum of research groups.
- The first workshop was held on Nov.1,1998 - Sept.1,1999.
- The 2nd: June, 2000 – Feb.,2001

# CLIR is critical ..

- Information access across languages is important in the Internet environment.
- …especially between languages with different origins, such as English and Japanese.
- CL is needed in IR with Japanese or other Asian language docs since a concept in the docs of the language may be represented as an English term.

# Ex. in Japanese documents, a concept can be represented …

- **Original English spelling**;
  Ex. cross language information retrieval,
  cross-lingual information retrieval,
  translingual information retrieval
- **Acronym**; Ex. CLIR
- **Transliterated form**; Ex. ••••••••
  •
- **Japanese**; Ex. ••••••

# Tasks of NTCIR Workshop 1

## IR

Ad Hoc IR: **J-JE**

Cross-lingual IR: **J-E**

scientific docs, more than half are J-E paired, 3 grade judgment

## Term Recognition & Role Analysis

(a) to extract terms ,

(b) to identify "object", "method" and "main operation"

# Test Collection1 (NTCIR-1)

## (1) Documents

Japanese & English, 330,000 documents

Abstracts of Conference papers, 65 societies

## (2) Topics

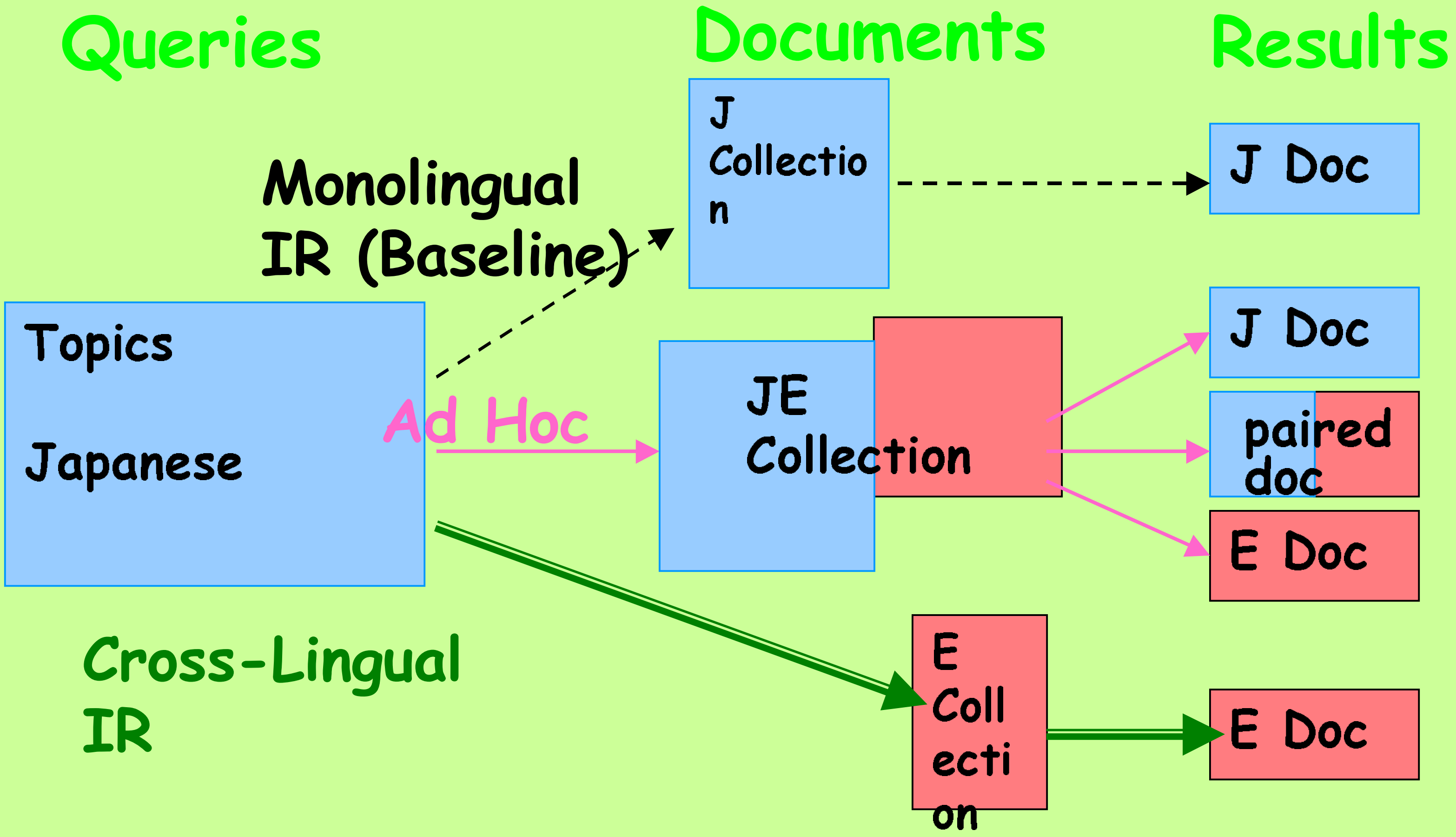
Japanese, 83 topics. SGML-like tagged

## (3) Relevance Judgments (Right Answers)

Relevant, Partial relevant, Non-relevant

## (4) Tagged Corpus (Term Recognition Tasks only)

# Collection & Tasks (NTCIR-1)



*Ad Hoc IR is also cross-lingual at NTCIR

# NTCIR Workshop 2

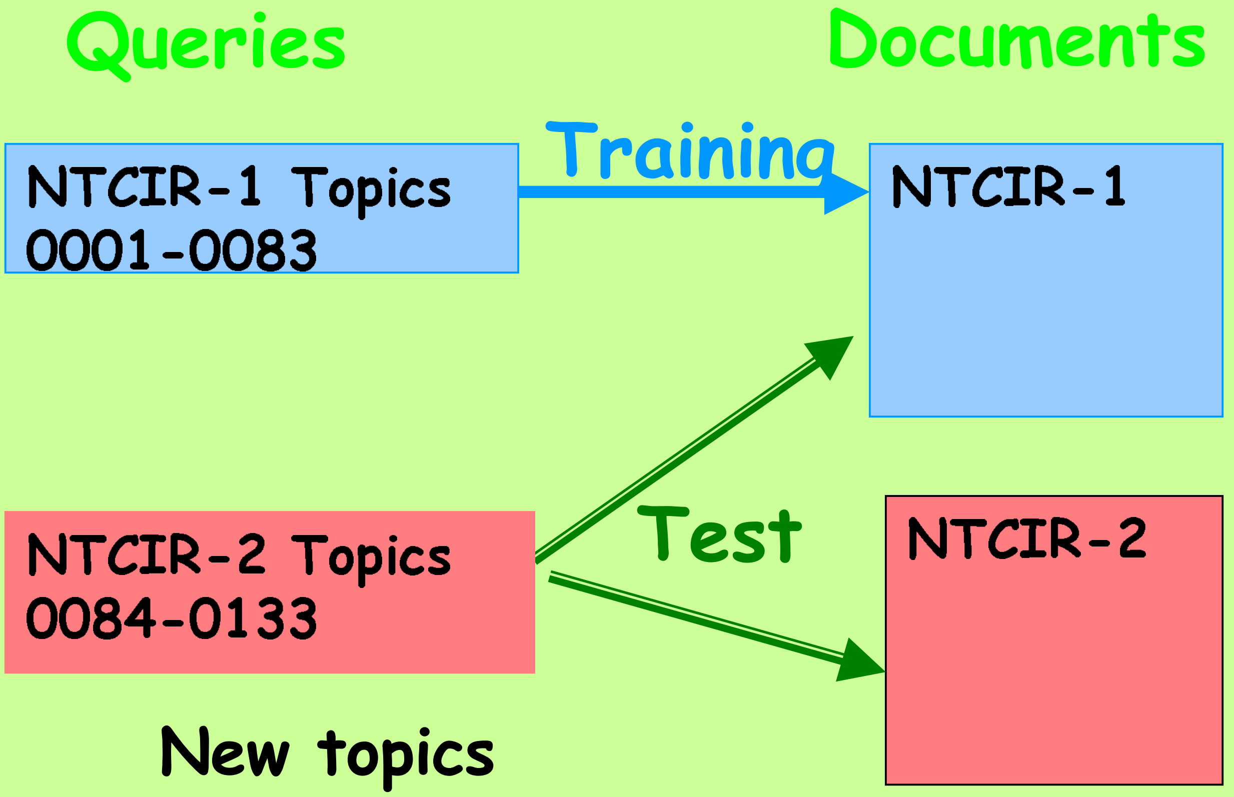
Asian language IR

- Chinese IR : Taiwan Nat. Univ.
  - C-C, E-C
- Japanese IR : NII
  - J-J, J-E, J-JE, E-E, E-J, E-JE
  - Segmented Data
  - 4grade judgement
- (Contact with Korean Eval.)

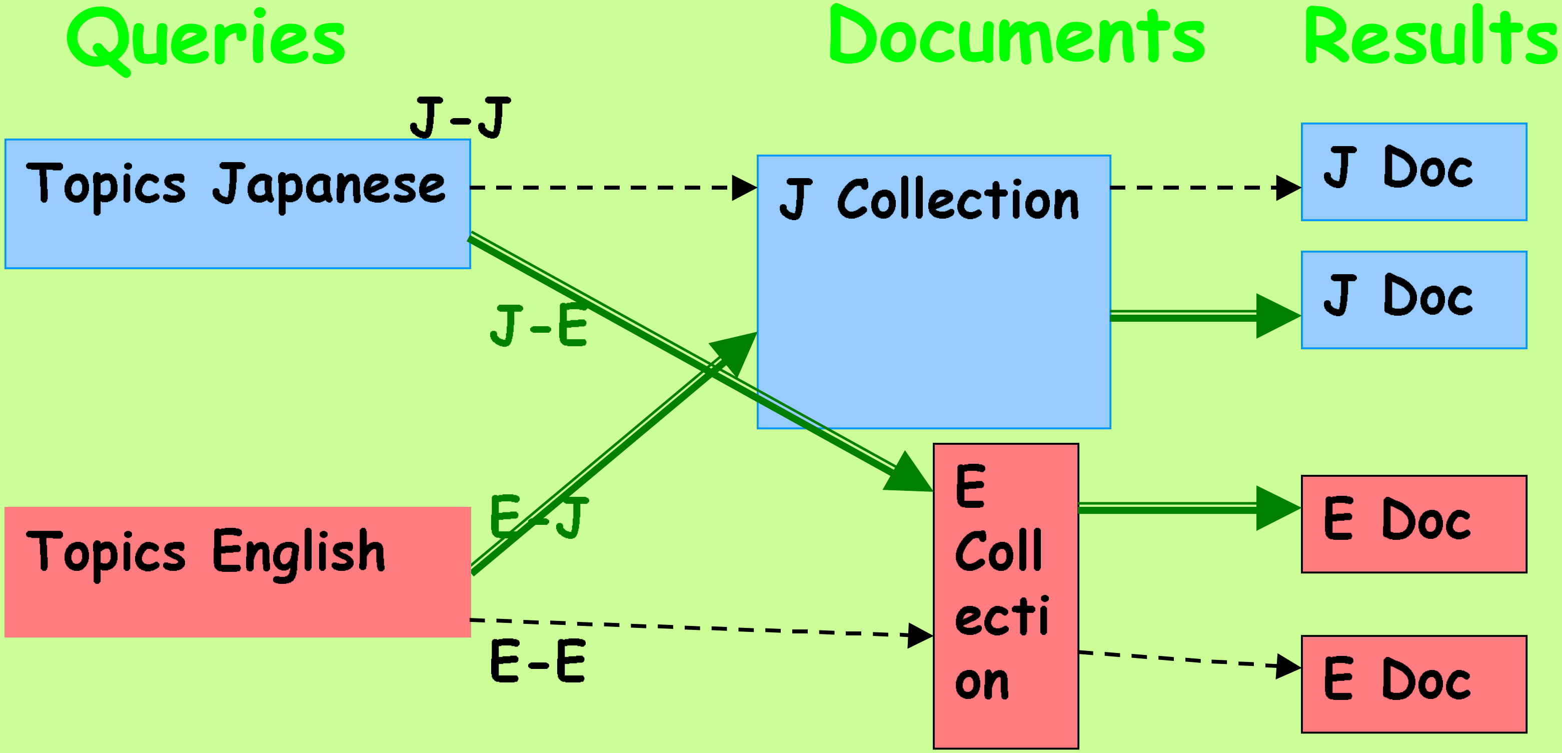
Text summarization of Japanese texts

Insentric / Exsentric (IR task based)

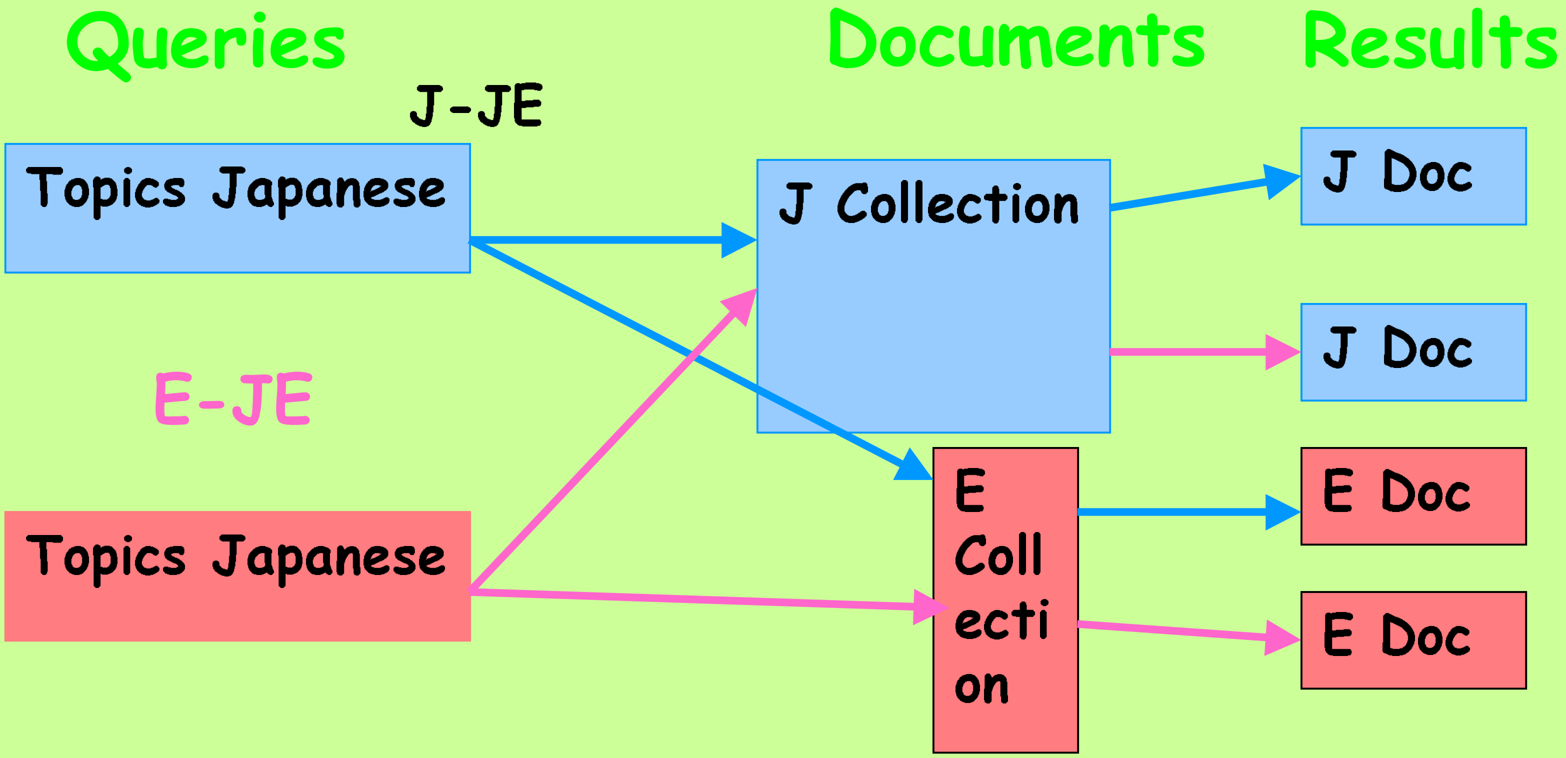
# Collection & IR Tasks in NTCIR Workshop 2



# Collection & Tasks



# Collection & Tasks



# Original Japanese doc's

<TITL>•••••••••••••••       ••••••
</TITL>

<ABST>••••••••••••••••••••••••••
•••••••••••••••••••••••.…

No explicit boundaries (spaces) between words in a sentence.

# NTCIR-1, manually tagged (ca. 2,000docs)

- A part of the collection contains detailed part-of-speech tags
- Because of absence of explicit boundary between words in Japanese sentences, we set the two levels of lexical boundaries

**w •••• m •• m ••• w • w ••• w •• m •• m •••• w ••••**

vector space model (of) based information retrieval system (-) ...

(whereas: w : word boundary m : morpheme boundary)

# NTCIR-2, automatically segmented (whole corpus: 340,000+400,000 docs)

<PJNM>•• • •• ••• •• • ••_•• • •• • ••<PJNM>

<ABST>••_•• • •• • ••• ••_••_•• • ••_•• •• • ••_• • •• • ••• • •• • •• • •• •• •• • …

Hard Segmentation: Space, longer units (~compound terms)

Soft Segmentation• _ Under score, components in a term

# Participants of NTCIR Workshop 1

Communications Research Laboratory (MPT)
Fuji Xerox
Fujitsu Laboratories
Hitachi
JUSTSYSTEM
Kanagawa University (2 groups)
KAIST/KORTERM
Manchester Metropolitan Univ.
Matsushita Electric Industrial
NACSIS
National Taiwan Univ.
NEC (2 groups)
NTT
RMIT & CSIRO

Tokyo Univ. of Technology
Toshiba
Toyohashi Univ. of Technology
Univ. of California Berkeley
Univ. of Lib. and Inf. Science
Univ. of Maryland
Univ. of Tokushima
Univ. of Tokyo
Univ. of Tsukuba
Yokohama National Univ.
Waseda Univ.

*28 groups from 6 countries

* 9 groups from companies

# Participants of NTCIR Workshop 2

ATT Labs & Duke Univ.(US)

Chinese Univ HK (HK)

Communications Research Laboratory (MPT)

Fuji Xerox

Fujitsu Laboratories (2)

Gifu Univ.

Hitachi co.

HK Polytechnic(HK)

IoS (China)

Johns Hopkins Univ.(US)

JR Res. Labs.

JUSTSYSTEM

Kanagawa University

KAIST/KORTERM(Korea)

Matsushita Electric Industrial

Nat. TsinHua Univ (TW)

NII (3)

NEC

New Mexico Univ.(US)

NTT & NAIST

OASIS

Queen College (US)

Ricoh Co.(2)

Surugadai Univ.

Toshiba/Cambrige/MicroSoft(UK)

Trans EZ (TW)

Toyohashi Univ. of Technology (2)

Univ. of California Berkeley (US)

Univ. of Electro-Communication(2)

Univ. of Exeter (UK)

Univ. of Lib. and Inf. Science

Univ. of Maryland (US)

Univ. of Montreal (Canada)

Univ. of Osaka

Univ. of Tokyo (2)

Yokohama National Univ.(2)

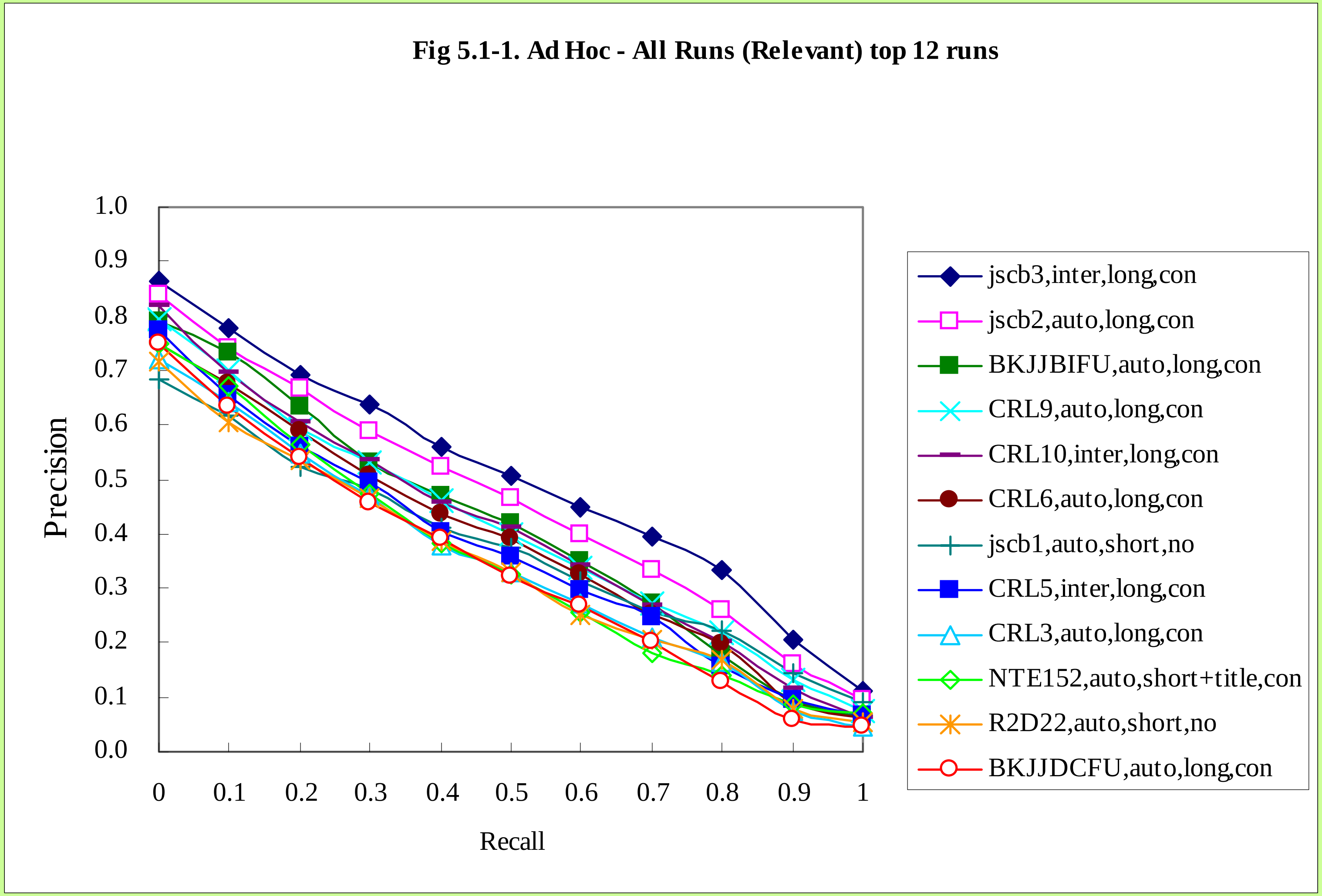
Waseda Univ.

*46 groups from 8 countries

# Results (NTCIR Workshop 1)

- *Various approaches* Pls visit and see the Online Proceedings
- Long Query > Short Query, but some runs are vice versa (short is mandatory)
- Interactive > Automatic
- Indexing: Bi-gram vs Morphological Analysis vs Suffix Array, Extended N-gram (adaptive segmentation,etc)
- Query expansion worked well in both interactive and automatic searches.
- Query Tansl. Only. Dictionary or MT based. No Corpus based

# Evaluation results: Ad Hoc All Runs



# Technical terms

- One of the difficult part of NTCIR-1
- few dictionaries.
- Needs to investigate the automatic construction of bilingual lexicon from texts
- Phonetics & transliteration of Katakana words was proposed and worked well

# NTCIR 2000/2001 schedule

Aug. 10, 2000: (JE)• Test documents• & topics

Aug. 30, 2000: (C) Test documents & topics

Sept. 8, 2000: (Summ) DryRun

Sept.18, 2000: (JE)•Results submission

Sept. 30, 2000: (c) Results submission

Oct, 2000 : (Summ) Test

Feb.21-23,2001: Workshop meeting, Tokyo.

# Results, NTCIR Workshop 2 (JE IR) as of 19th September 2000

| Task: | Run | Group | AutoRun(Group) | ManRun(Group) |
|---|---|---|---|---|
| J-J: | 85 | 15 | 73 (14) | 12 (4) |
| E-E: | 15 | 6 | 15 (6) | 0 |
| J-E: | 37 | 11 | 26 (11) | 0 |
| E-J: | 25 | 8 | 25 (8) | 0 |
| J-JE: | 15 | 7 | 15 (7) | 0 |
| E-JE: | 11 | 4 | 11 (4) | 0 |
| | | | | |
| Total | 188 | 23 | 176 (22) | 12 (4) |

# Future Plan- 2-directions plan

- Traditional Approach
  - Language; Japanese, Asian Language
  - CLIR → International Collaboration
  - Resource creation & Share
- Challenging Issue

  More realistic!
  - We have to think about; Document types & Characteristics of the users who use the type of documents, ex. Web, Patent, Images, etc.
  - How user can be involved to improve CLIR? How symtem can help user to create appropriate queries?
  - Post retrieval processing: needs IR+NLP

# Future Plan (2)

- Challenging issues:

  *post-retrieval processing & support human information work by making retrieved documents usable (ex.pinpointing the answer, text summarization, comparing multiple docs., text mining, etc)

  *Various documents; patents, Web documents, multimedia, etc., etc.

- More investigation on evaluation scheme

# Future Plan (3)

- These two direction for the future plan must be supported by the Forum and Discussion on Evaluation
  - We have done; multigrade judgment, topic oriented judgment vs situation oriented judgment, analysis of stableness of the eval. using TC., trying to estimate the difficultiness of the topics
  - We would like to think about; Relevant or more? Reliability? Freshness? Usefulness? Also eval. of component (a module in a whole IR system)? for example, transliteration will improve search effectiveness of CLIR? Named entities? Segmentation (N-gram vs Word/phrase based?),
  - What shall we evaluate?

# Things to be thought

- International Collaboration & Real MultiLingual TC
- Resource Creation and Sharing
- Language Specific Aspects
- Component & Whole
- Retrieval and More (post-retrieval processing)
- Social environment in the Asia
  - Diversity
- Too many evaluation?

*Proceedings is available online :

http://www.rd.nacsis.ac.jp/~ntcadm/workshop/OnlineProceedings/

*To subscribe the ML: send a message

subscribe ntcir your_email_adress

as a body of message to

majordomo@rd.nacsis.ac.jp

*The test collection NTCIR-1 is available for outside the Workshop:

http://www.rd.nacsis.ac.jp/~ntcadm/

Thanks

Obrigada