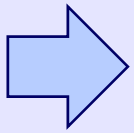
Introduction to NTCIR-7

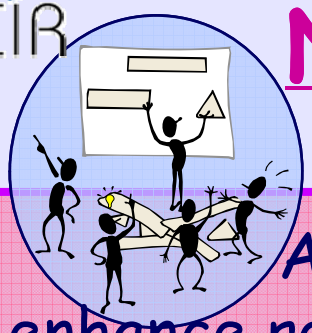
Noriko Kando
National Institute of Informatics, Japan
http://research.nii.ac.jp/ntcir/
kando (at) nii. ac. Jp

NTCIR

Road map



- What is NTCIR
- Leason learned from past NTCIRs
- Brief Introction to NTCIR-7
- Conclusion



NTCIR: NII Test Collection for Information Retrieval

Research Infrastructure for Evaluating IA

A series of evaluation workshops designed to enhance research in **information-access** technologies by providing an **infrastructure** for large-scale evaluations.

- Data sets, evaluation methodologies, and forum

Project started in late 1997

- Once every 18 months

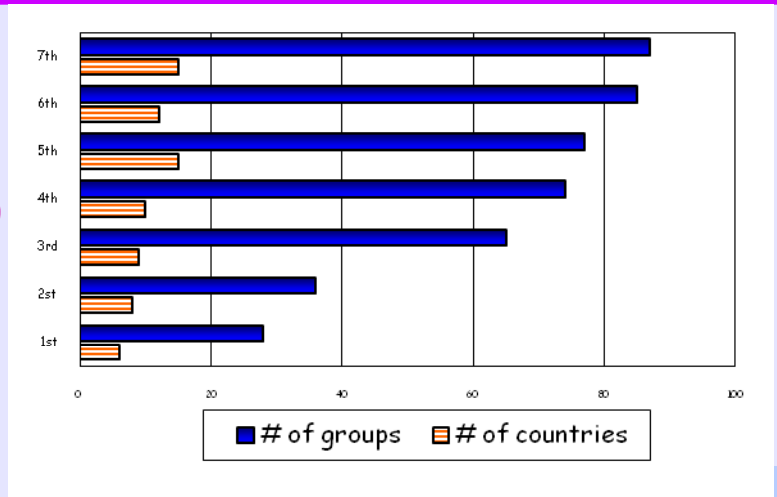
Data sets (Test collections or TCs)

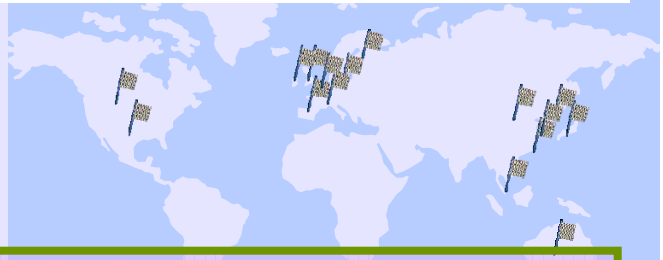
- Scientific, news, **patents**, and web
- Chinese, Korean, Japanese, and English

Tasks

- IR: Cross-lingual tasks, patents, web,
- QA: Monolingual tasks, cross-lingual tasks
- Summarization, trend info., patent maps
- Opinion analysis, text mining

Community-based Research Activities





NTCIR-7 participants
82 groups from 15 countries

Noriko Kando

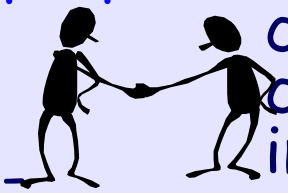
NTCIR

NTCIR provides;

- A scientific basis for understanding the effectiveness of automated search systems
- Large-scale "**Test Collections**" or TC
  - Organizers provide a data set
  - Participants use the same data set to compare the effectiveness of their systems
  - TCs are available for research purpose
- Forum of researcher groups
- Show-case of the State-of-the-art technologies
- Investigations into evaluation methodologies and metrics

Document set, a set of topics, and a list of relevant documents for each topic

NTCIR enables:



* Cross-system comparison on a common infrastructure

* Speeds up R&D and technology transfers

NTCIR

Information retrieval

- Retrieve RELEVANT information from vast collection to meet users' information needs

Using computers since the 1950s

First CS uses human assessments as success criteria
  - Judgments vary
  - Comparative evaluations on the same infrastructure

NTCIR

Information access (IA)

- Whole process ofpreparing information from the vast collection of documents usable by users.
- For example, IR, text summarization, QA, text mining, and clustering
- Use human assessments as success criteria

Focus of NTCIR

Lab-type IR Test

Asian Languages/cross-language

Variety of Genre

Parallel/comparable Corpus

New Challenges

Intersection of IR + NLP

To make information in the documents more usable for users!

Realistic eval/user task

Forum for Researchers

Idea Exchange

Discussion/Investigation on Evaluation methods/metrics

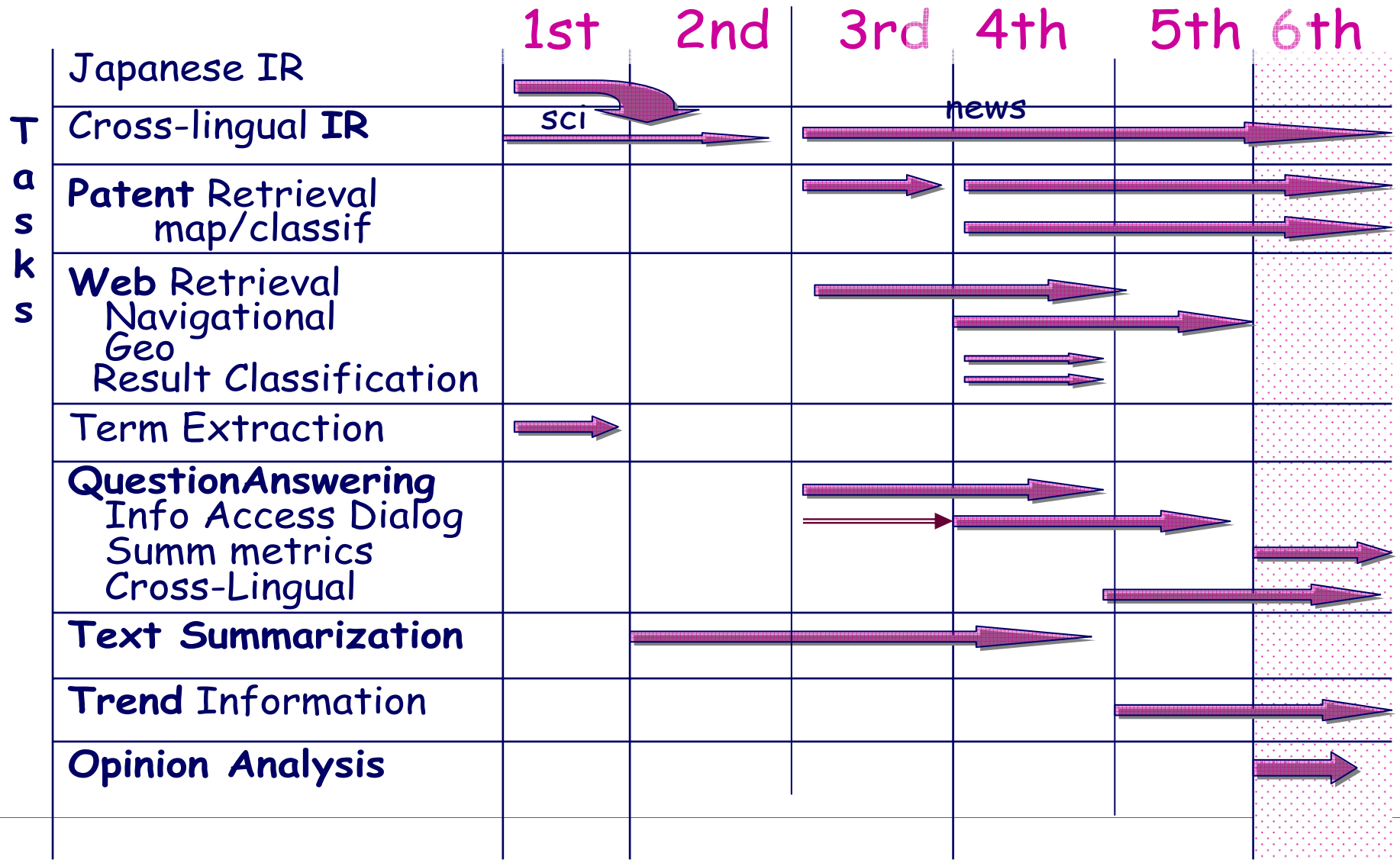
History…

Project starts late 1997

Nov '98 – Sep '99	NTCIR-1
Jun '00 – Mar '01	NTCIR-2
Sep '01 – Oct '02	NTCIR-3
Apr '03 – Jun '04	NTCIR-4
Oct '04 – Dec '05	NTCIR-5
Apr '06 – May '07	NTCIR-6
Oct '07 – Dec '08	NTCIR-7

NTCIR-7 Workshop Meeting Dec 16-19

Tasks at past NTCIRs



NTCIR

NTCIR-7 Clusters

Cluster 1. Advanced CLIA
- Complex CLQA (Chinese, Japanese, English)
- IR for QA (Chinese, Japanese, English)

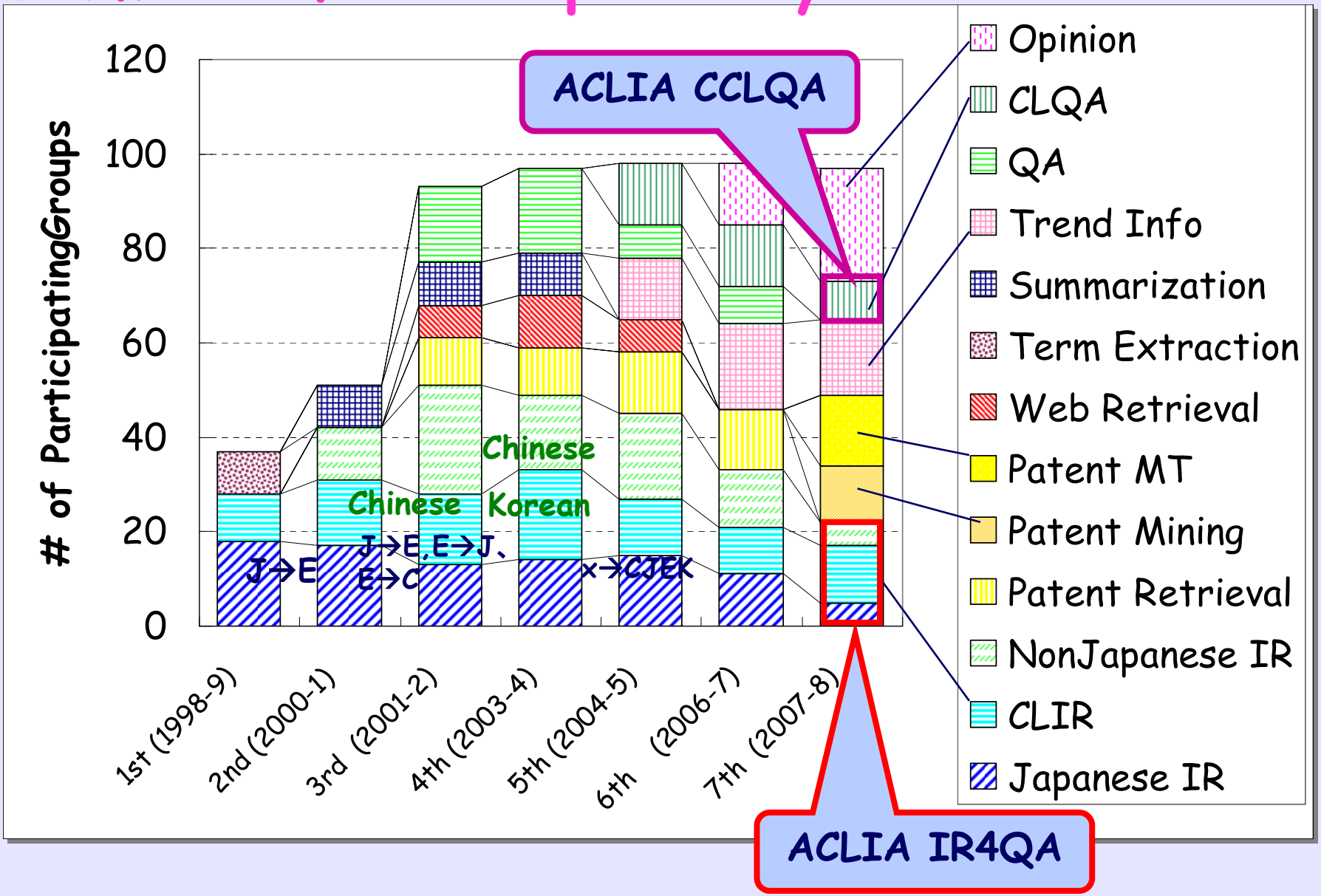
Cluster 2. User-Generated :
- Multilingual Opinion Analysis

Cluster 3. Focused Domain : Patent
- Patent Translation ; English -> Japanese,
- Patent Mining paper -> IPC

Cluster 4. MuST :
- Multi-modal Summarization of Trends

MuST: Visualization Challenge

Number of Participants by Tasks



For details, please click on each column in the table.

	Ad Hoc/ CLIR [Scientific Abstracts] (Japanese/ English IR))	Chinese IR	CLIR [News] (Closs-Lingual QA)	CLQA (Closs-Lingual IR)	MuST (Multimodal Summarization for Trend Information)	OPINION (Opinion Analysis)	PATENT	QAC (Question Answering)	TMREC (Term Recoginition)	TSC (Summa-rization)	WEB
NTCIR-1	NTCIR-1 **Ad Hoc/ CLIR**	-	-	-	-	-	-	-	NTCIR-1 **TMREC**	-	-
NTCIR-2	NTCIR-2 **Ad Hoc/ CLIR**	**CIRB 010**	-	-	-	-	-	-	-	NTCIR-2 **SUMM**	-
NTCIR-3	-	-	NTCIR-3 **CLIR**	-	-	-	NTCIR-3 **PATENT**	NTCIR-3 **QA**	-	NTCIR-3 **SUMM**	NTCIR-3 **WEB**
NTCIR-4	-	-	NTCIR-4 **CLIR**	-	-	-	NTCIR-4 **PATENT**	NTCIR-4 **QA**	-	NTCIR-4 **SUMM**	NTCIR-4 **WEB**
NTCIR-5	-	-	NTCIR-5 **CLIR**	NTCIR-5 **CLQA**	-	-	NTCIR-5 **PATENT**	NTCIR-5 **QA**	-	-	NTCIR-5 **WEB**
NTCIR-6	-	-	NTCIR-6 **CLIR**	NTCIR-6 **CLQA**	NTCIR-6 **MuST**	NTCIR-6 **OPINION**	NTCIR-6 **PATENT**	NTCIR-6 **QA**	-	-	-

NTCIR-7 PC Meeting@NTCIR-6

Mark Sanderson, Doug Oard, Atsushi Fujii, Tatsunori Mori, Fred Gey, Noriko Kando (and others)

NTCIR-7: Advanced CLIA

Teruko Mitamura (CMU)
Eric Nyberg (CMU)

Ruihua Chen (MSRA)
Fred Gey (UCB),
Donghong Ji (Wuhan Univ)
Noriko Kando (NII)
Chin-Yew Lin (MSRA)
Chuan-Jie Lin (Nat Taiwan Ocean Univ)
Tsuneaki Kato (Tokyo Univ)
Tatsunori Mori (Yokohama N Univ)
Tetsuya Sakai (NewsWatch)

Advisor: K.L.Kwok (Queen College)

NTCIR-7: UGC (Blog)

David K Evans (NII -> Amazon Japan)
Yohei Seki (Toyohashi U Tech -> Columbia U)

LunWei Ku (National Taiwan Univ)
Le Sun (Chinese Academy of Science)
Hsin-Hsi Chen (National Taiwan Univ)
Noriko Kando (NII)

NTCIR-7: Focused Domain (Patent)

Atsuhi Fujii (Univ Tsukuba)

Taiich Hashimoto (Tokyo Insti Tech)
Makoto Iwayama (Tokyo Insti Tech/ Hitach)
Hidetsugu Nanba (Hiroshima City Univ)
Masao Utiyama (NICT),
Mikio Yamamoto, U Tsukuba)
Takehito Utsuro (U Tsukuba)

MuST: Multimodal Summarization for Trend Information

Tsuneaki Kato (Tokyo Univ)
Mitsunori Matsushita (NTT Comm Sci Lab → Kansei Univ)

[CCLQA]
•Academia Sinica
•Beijing Univ of Posts & Telecoms, China
•Carnegie Mellon Univ
•NICT
•NTT Corporation
•Shenyang Institute of Aeronautical Engineering
•Wuhan Univ
•Yokohama National Univ

[IR4QA]
•Carnegie Mellon Univ
•Chaoyang Univ of Technology
•Chinese Academy of Sciences(ICT)
•Harbin Institute of Technology + Heilongjiang Institute of Technology
•National Taiwan Univ
•Open Text Corporation
•Shenyang Institute of Aeronautical Engineering
•Toyohashi Unive of Technology
•Univ of California, Berkeley
•Univ of Montreal
•Wuhan Univ
•Wuhan Univ of Science and Technology

[MOAT]
•Beijing univ
•Chinese Academy of Sciences(NLPR-IACAS)
•Chinese Univ of Hong Kong + Hong Kong Polythechnic Univ+ Tsinghua Univ
•DAEDALUS, S.A.
•Hiroshima City Univ
•Information and Communications Univ
•Chinese Academy of Sciences(ISCAS)
•Keio Univ
•City Univ of Hong Kong
•National Taiwan Univ
•NEC
•Northeastern Univ
•Peking Univ
•Pohang Univ of Science and Technology
•Swedish Institute of Computer Science
•Technical Univ of Darmstadt
•Graduate Univ for Advanced
•Tornado Technologies Co., Ltd.,
•Toyohashi Univ of Technology
•Univ of Neuchatel
•Univ of Sussex

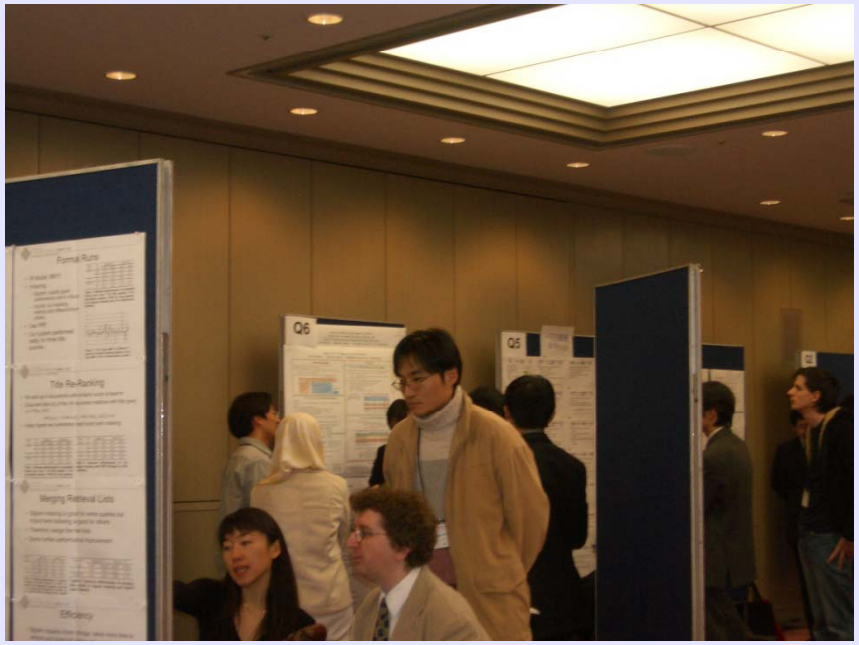
[Must]
•Hiroshima City Univ
•Keio Univ
•Mie Univ
•NICT
•NEC
•Ochanomizu Univ (2 Groups)
•Okayama Univ
•Osaka Prefecture Univ
•Otaru Univ of Commerce
•Tokyo Metropolitan Univ
•Tokyo Denki Univ
•Univ of Sheffield
•Yokohama National Univ

[PAT MIN]
•Hiroshima City Univ
•Hitachi, Ltd.,
•Huafan Univ
•Nagaoka Univ of Technology
•Northeastern Univ
•NTT Corporation
•Peking Univ
•Shenyang Institute of Aeronautical Engineering
•Toyohashi Univ of Technology
•Univ of California, Berkeley
•Univ of Montreal
•Xerox

[PAT MT]
•Fudan Univ
•Harbin Institute of Technology + Heilongjiang Institute of Technology
•Hitachi,Ltd.,
•Japan Patent Information Organization
•Kyoto Univ
•Massachusetts Institue of Technology
•Nara Institute of Science and Technology + NTT
•NICT
•National Taiwan Normal Univ
•NTT Corporation
•Pohang Univ of Science and Technology
•TOSHIBA
•Tottori Univ
•Toyohashi Univ of Technology + Hosei University
•Univ of Tsukuba

Oral presentation session

Poster session of past NTCIR Meeting



Break out session

Breakout session

NTCIR-6 (2007) banquet

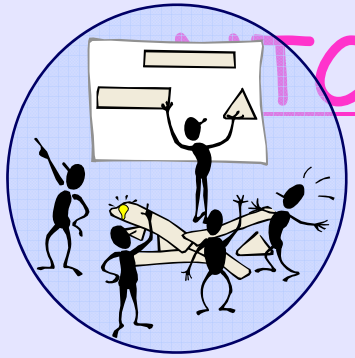
NTCIR Office members + friends

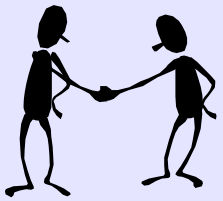
Evaluation Workshops

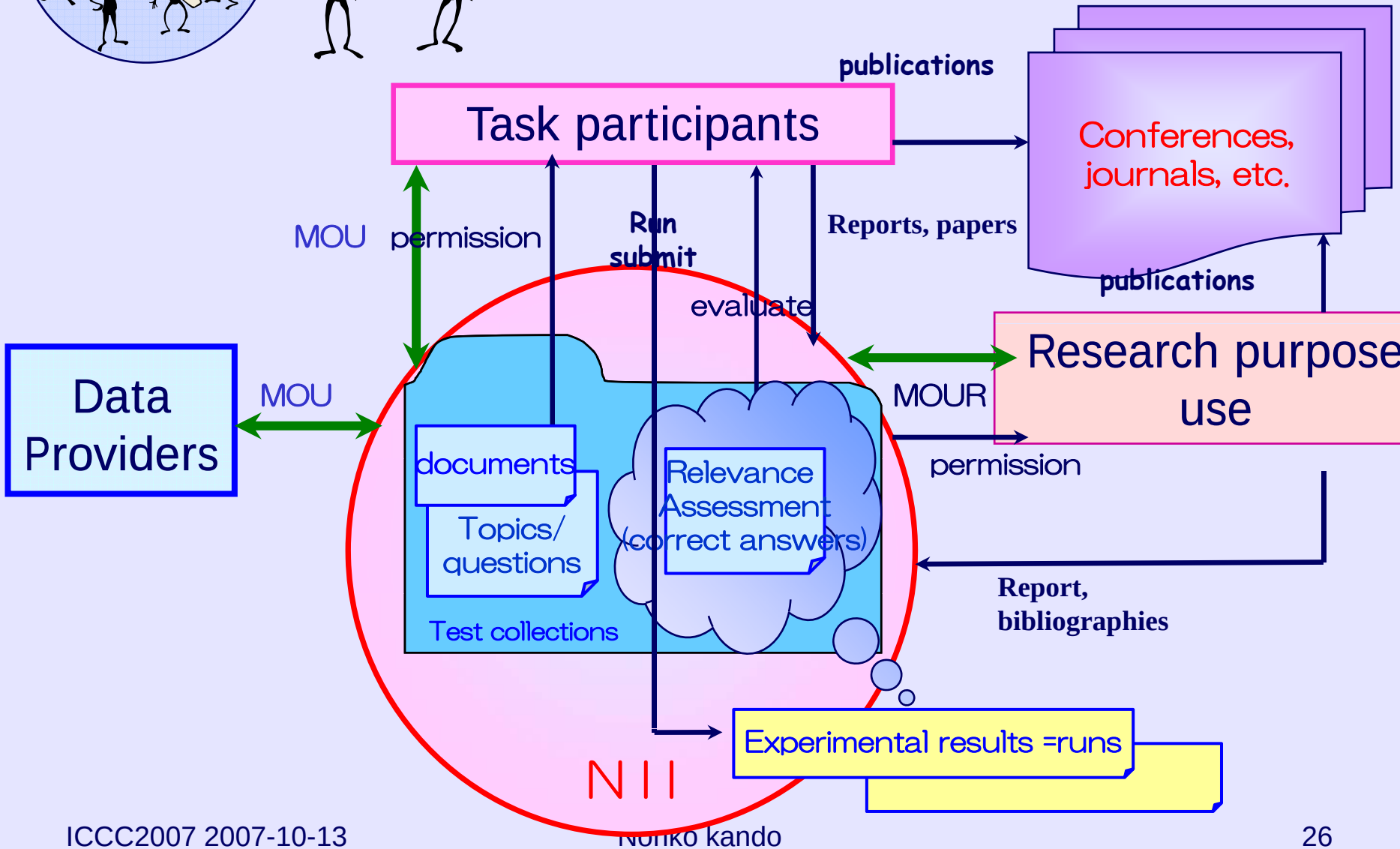
- "evaluation"

  It is not an competition! not an exam!
- Constructs a common data set usable for experiments.
- provides to participants the data sets and unified procedures for evaluation
  - Each participating research group conducts experiments with various approaches and can participate with own purpose.
- Successful examples; TREC, CLEF, DUC, INEX, and TAC, FIRE (new!) Community-based activities
- Implications are various

NTCIR: NII Test Collection for Information Retrieval







Sample document:

<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>

Sample topic:

written statement of user's needs

<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>

Any fields are usable for retrieval. A run using DESC only is mandatory.

Relevance Judgments

- Always Multigrades in NTCIR: 3 or 4 grades
  - [Highly Relvant (S)]
  - Relevant(A),
  - Partial Relevant(B),
  - Irrelevant(C)
- Traditionally "binary" judgments are popular.
- Contains extracted phrases/passages showing the reason that the analyst judged it as "relevant" in NTCIR collections.

NTCIR

IA Systems Evaluation

- Engineering Level: Efficiency
- Input Level: ex. Exhaustivity, quality, novelty of DB
- **Process Level: Effectiveness ex. recall,precision**
- Output Level: Display of output
- User Level: ex. Effort that users need
- Social Level: ex. Importance (Cleverdon & Keen 1966)

Evaluation of IA Effectiveness

- Long tradition of laboratory-typed testing using test collection since Cranfield in 1960s.

- Basic metrics are; … and their variants

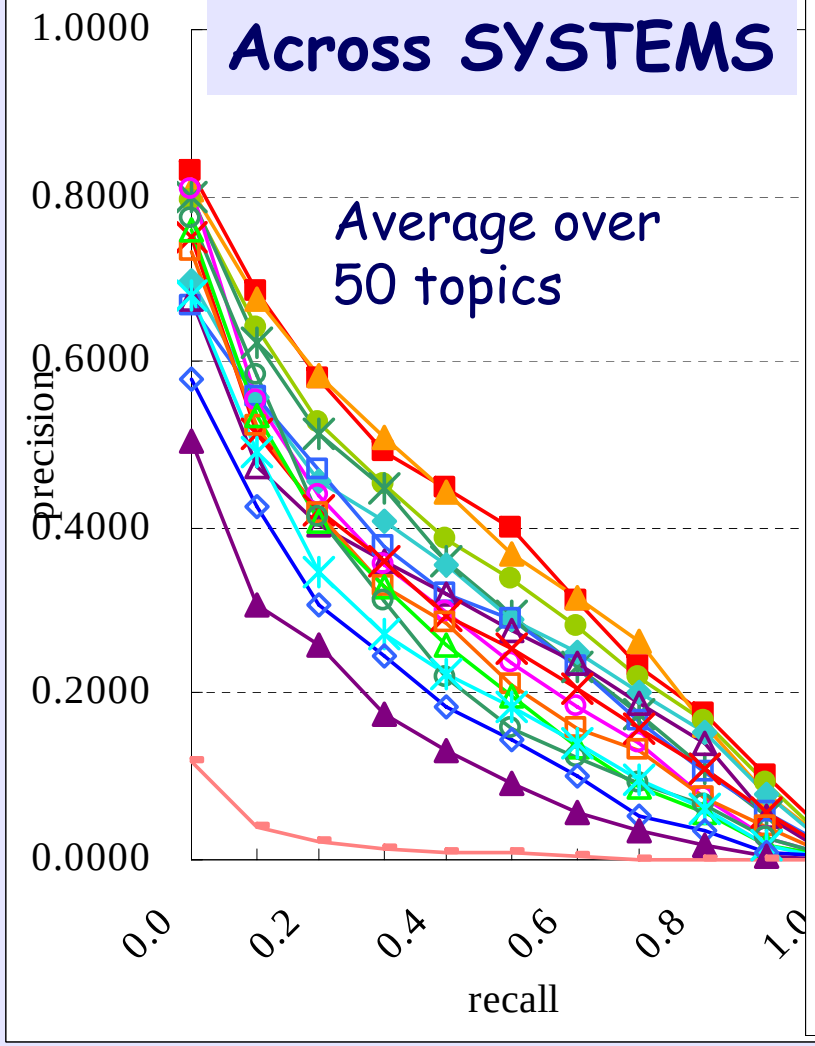
$$\text{Precision} = \frac{\#\text{ of retrieved-relevant}}{\#\text{ of retrieved}}$$

$$\text{Recall} = \frac{\#\text{ of retrieved-relevant}}{\#\text{of all relevant docs}}$$

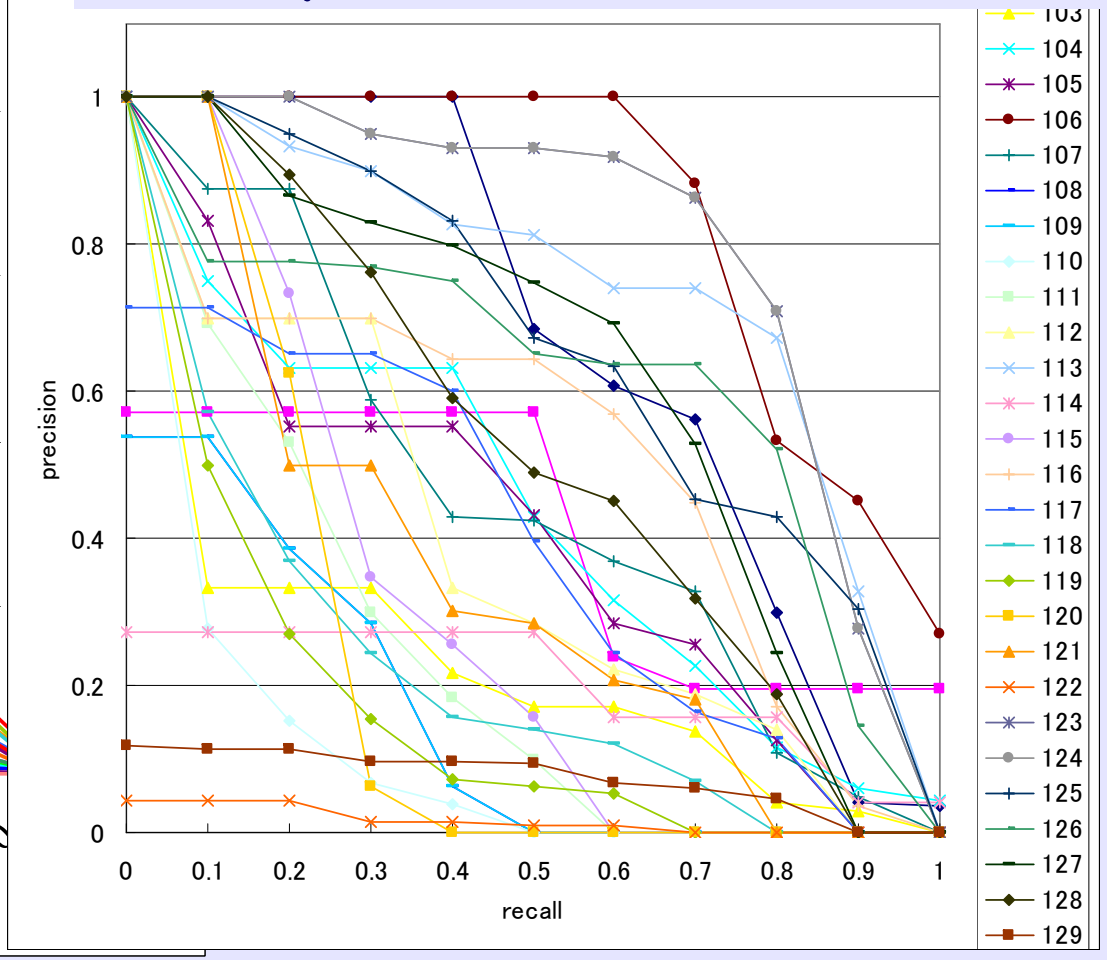
"Recall" can be calculated only in the experimental setting in which all the relevant docs are known.

Retrieval Difficulty Varies with Topics

Effectiveness Across SYSTEMS



Effectiveness Across TOPICS on a System

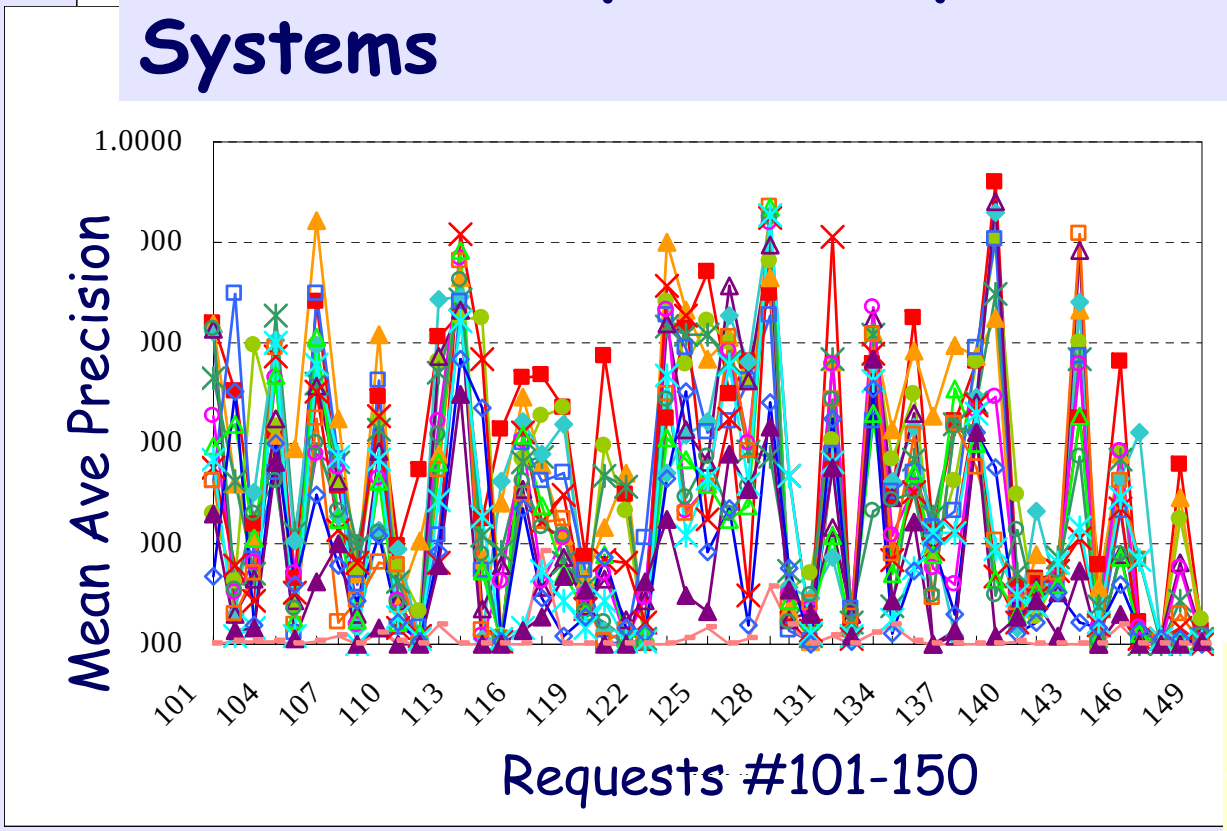


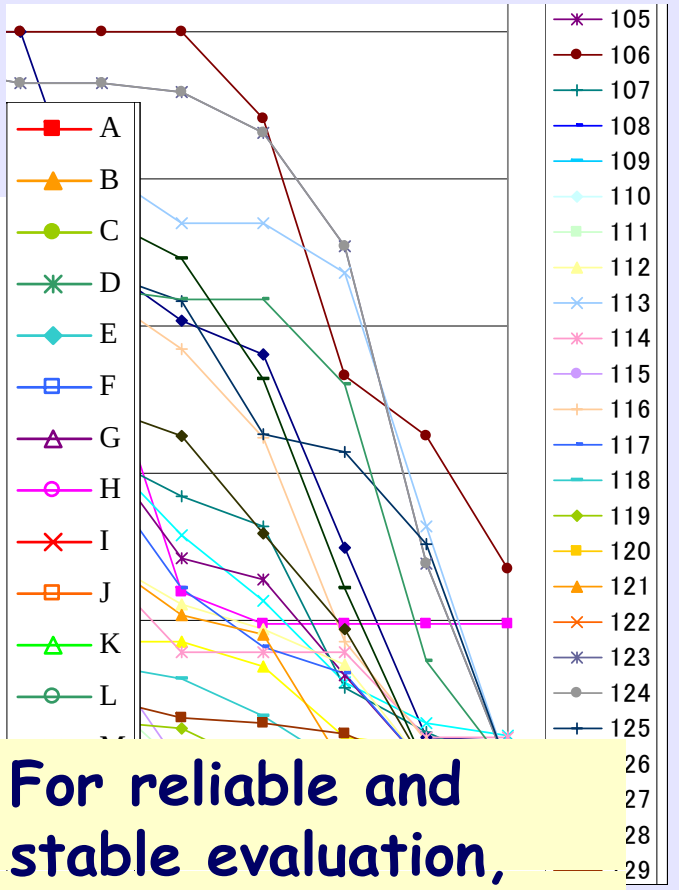
Retrieval Difficulty Varies with Topics

Effectiveness Across SYSTEMS

Effectiveness Across TOPICS on a System

"Difficult Topics" Vary with Systems

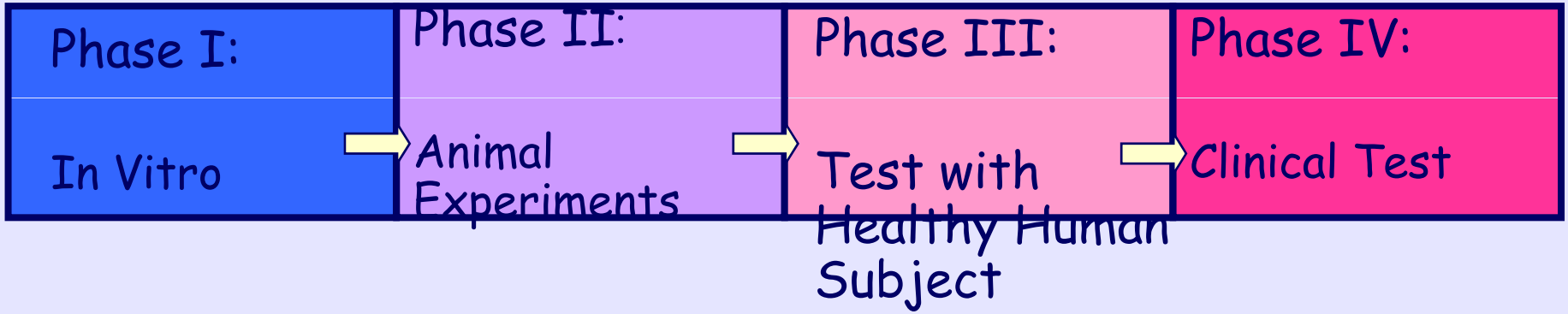




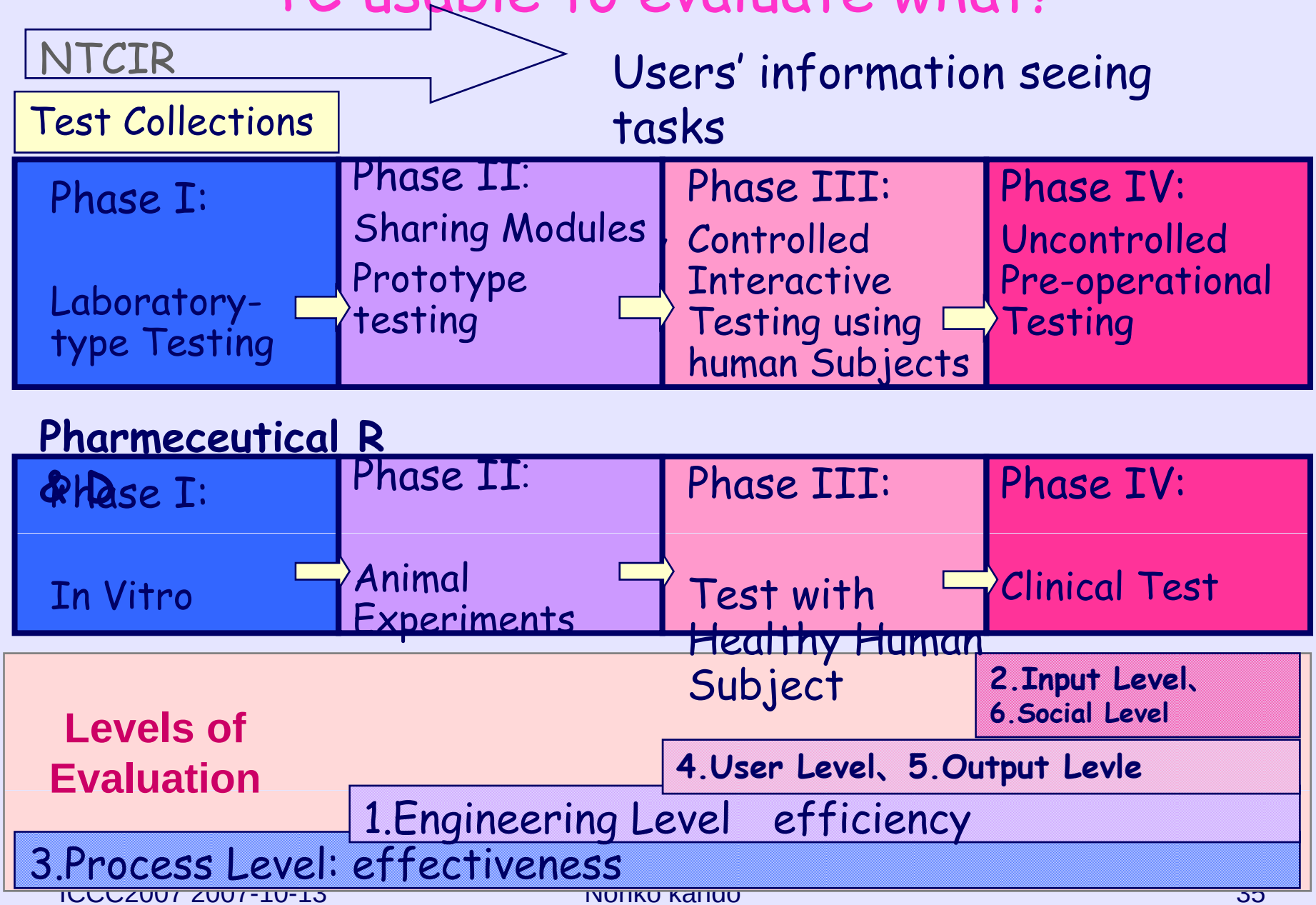
For reliable and stable evaluation, using substantial # topics is inevitable

TC usable to evaluate?

Pharmaceutical R & D



TC usable to evaluate what?



NTCIR

Summary of "What is NTCIR"

- Providing a scientific basis for understanding the effectiveness of automated search systems
- Leveraging the R&D and technology transfer
- Reusable Test collection is a key component
- Evaluating search effectiveness is not easy. A small-scale or carelessly-designed TCs may skew the test results

NTCIR

Road map

- What is NTCIR
- Leason learned from past NTCIRs
- Brief Introction to NTCIR-7
- Conclusion

Lessons Learned from Past NTCIRs

Information Retrieval
Ad hoc

1. **Ad hoc & CLIR**
   - Scientific Abstracts (NTCIR-1 & -2)
   - News (NTCIR-3 thurough -7)
   - Blogs (NTCIR-7)
2. **IR in specific Document Genres**
   - **Patent** (NTCIR-3 through -6)
     → Translation, Mining (NTCIR-7)
   - **WEB** (NTCIR-3 through -5)

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

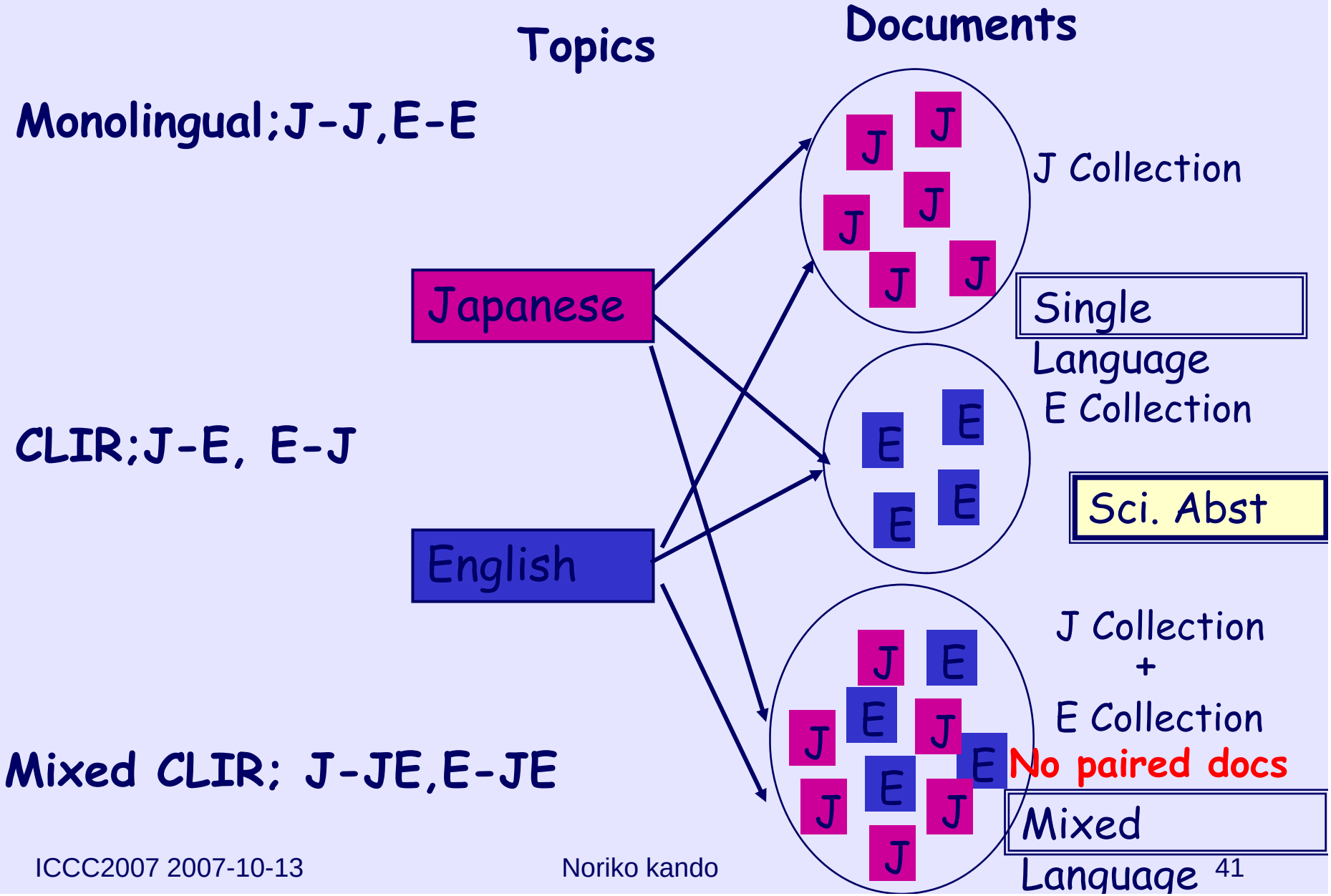
3. **Interest increasing rapidly-**

   Commercial/industrial exchange increased:

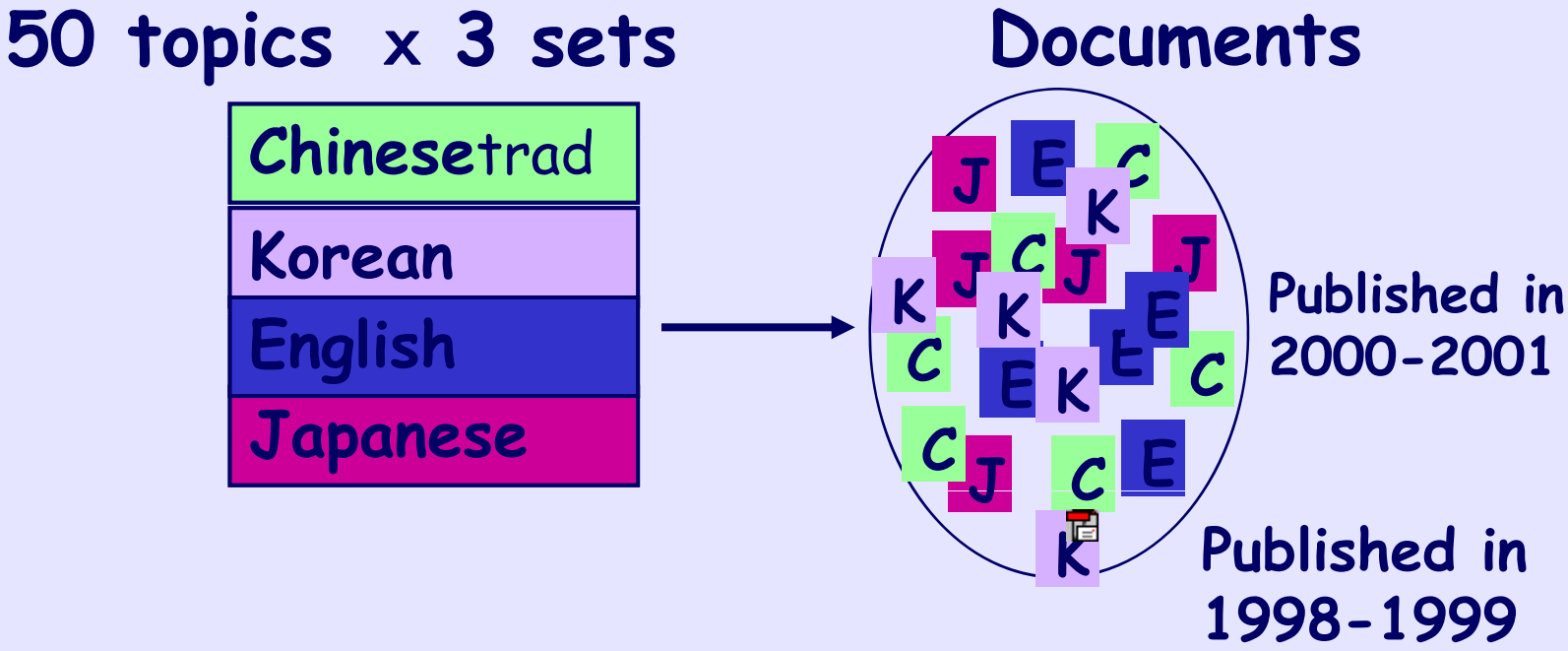
   Cultural/Social interest, Human Exchange,

4. **Languages structures are completely different** Character codes are different

NTCIR -1 &-2: Japanese & English



NTCIR-3 throu -7 CLIR

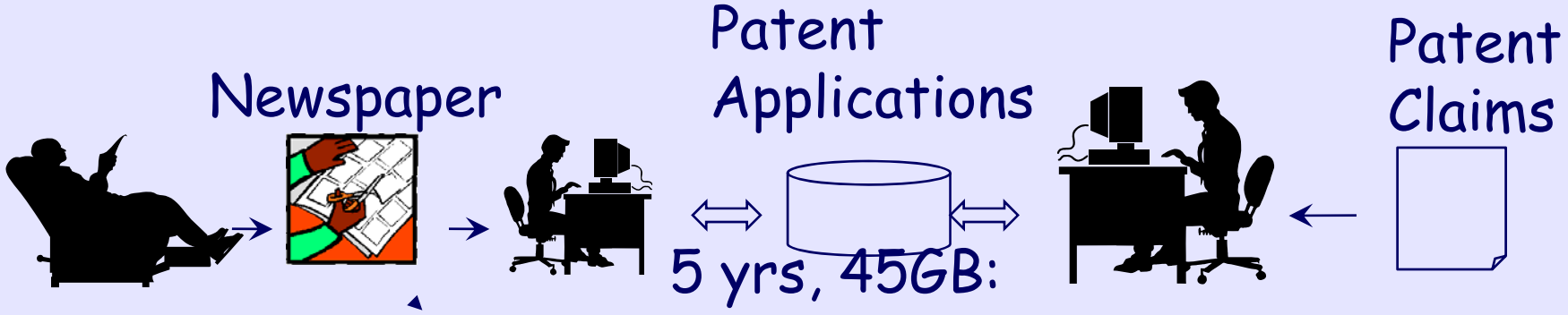


- Forcus: NE, OOV
- proper nouns vs without PN
- domestic/regional/international

CLIR: Lessons Learned

- IR Models: major IR models were worked
- Indexing: bigram vs word vs others, hybrid
- Mostly "Query Trans", but a few "Doc Trans"
- Translation disambiguation w/ WEB w/target doc
- **Out-of-vocabulary (OOV) problem**
  - Transliteration - Cognate
  - **NE identification**
  - **Use of Web**
- Query expansion techniques
  - Selective application PRF, Bounce & Throw
  - Clustering

Patent Retrieval Tasks
situation & users' information seeking task



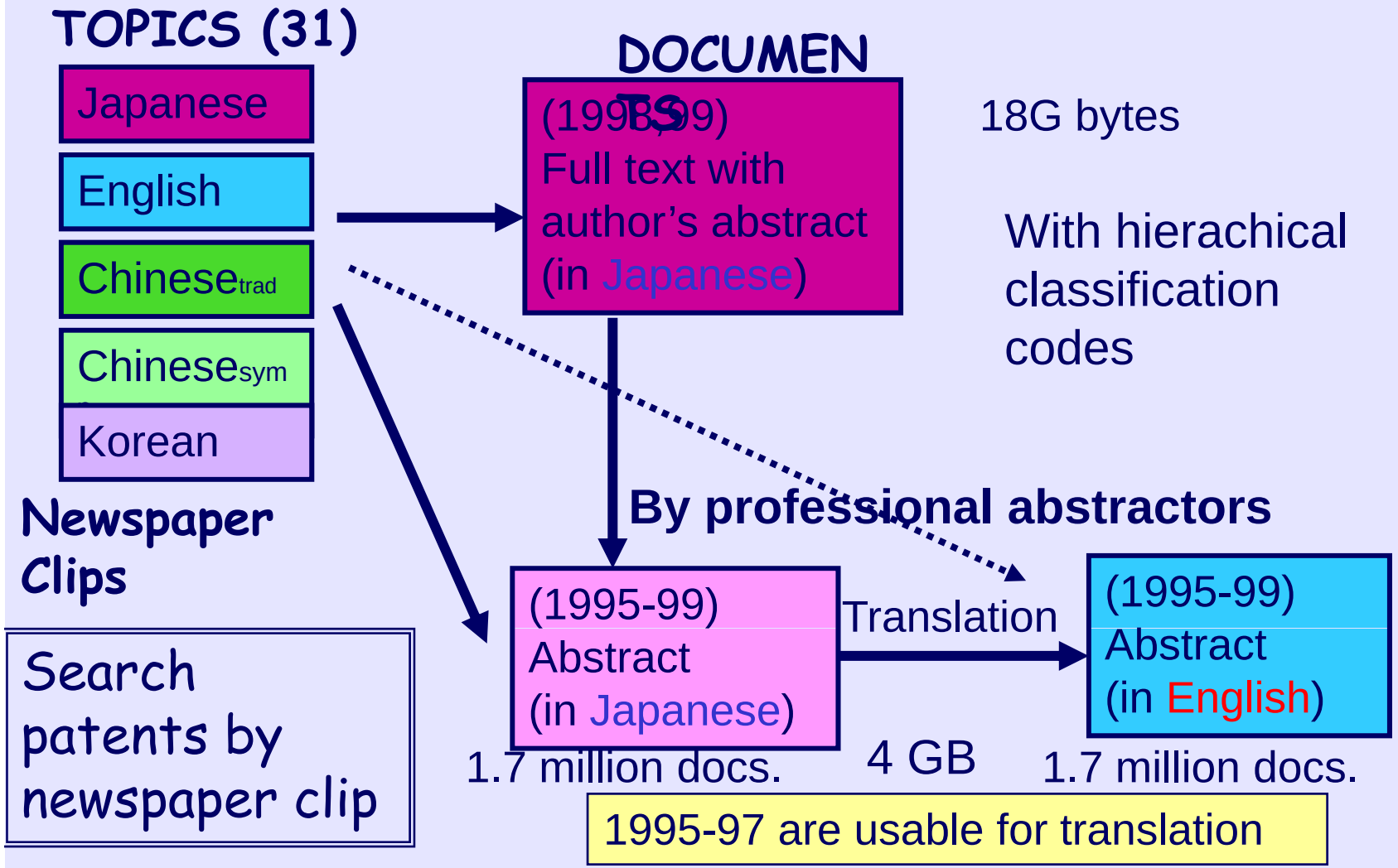
NTCIR-3 PATENT
(2001-2002)

Technological Survey:
Search patents by newspaper
End user: non-experts (ex. Business manager)

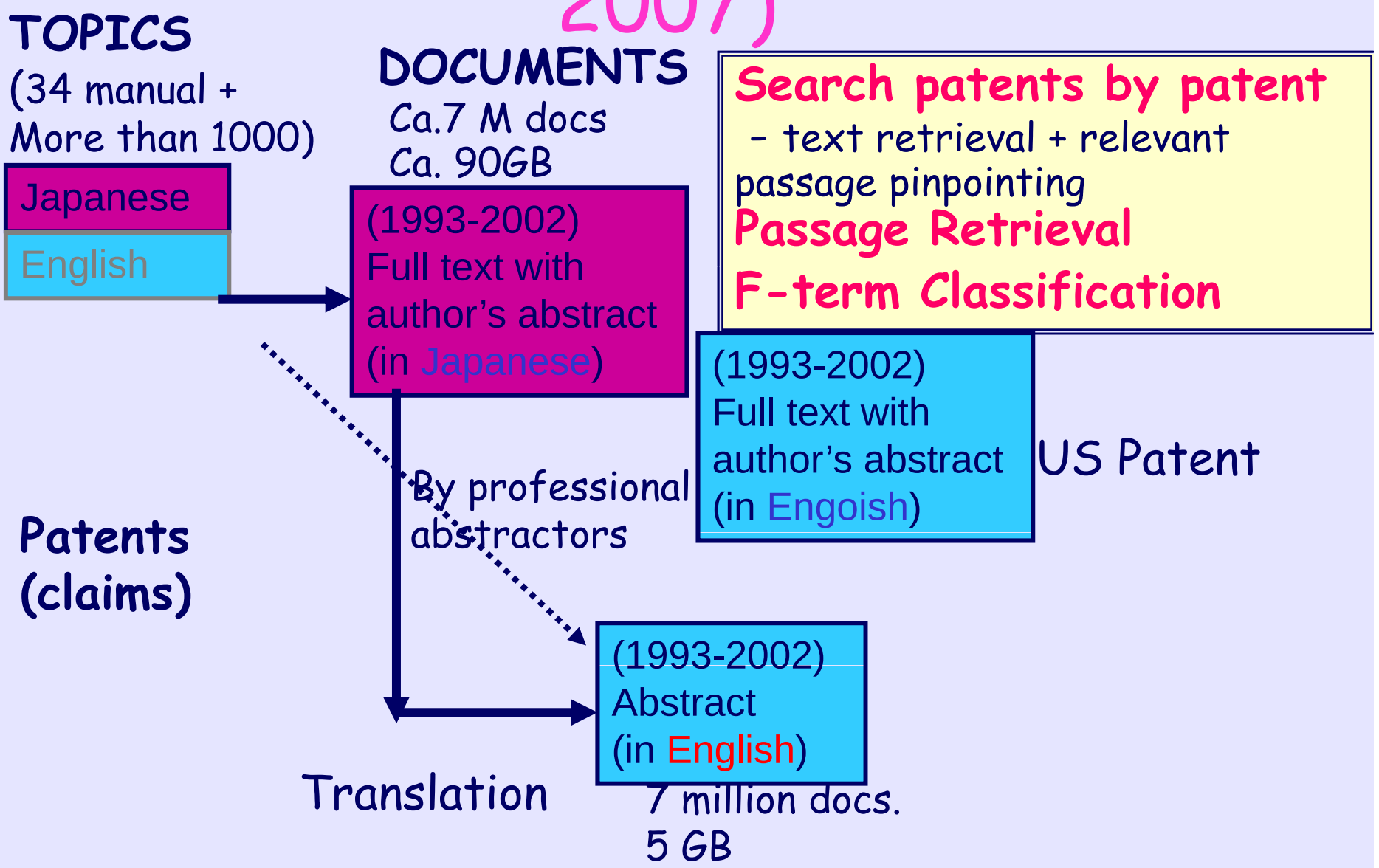
NTCIR-4,-5 PATENT
(2003-2004)(2004-2005)

From a claim of a new patent application, search patents that can invalidate the new patent application.
User: patent experts

NTCIR-3 Patent Collections



NTCIR-4 thro -6 Patent (2004-2007)



automatic patent map generation Example (blue light-emitting diode)

given

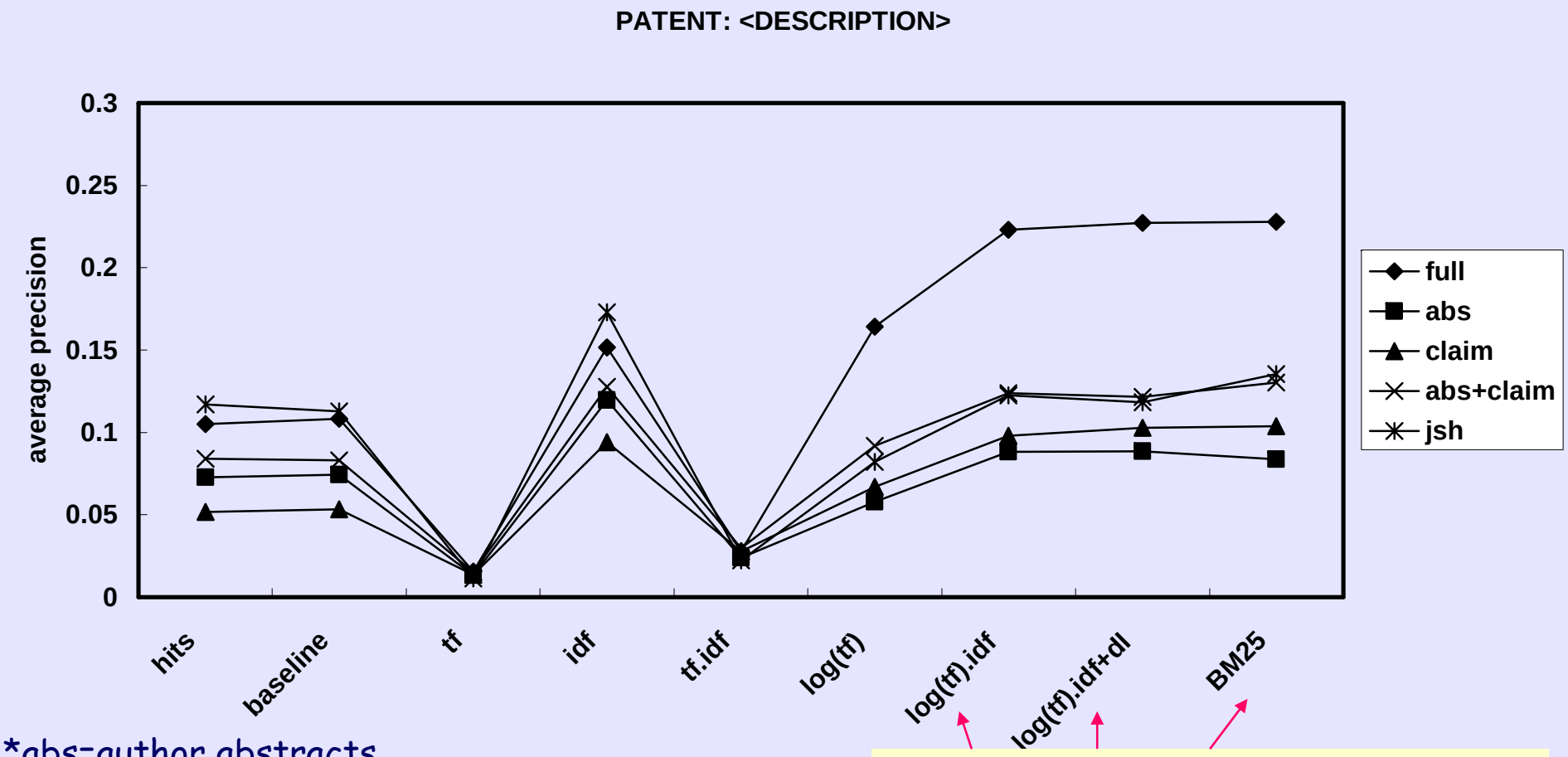
problems to be solved

solutions

	crystalline	reliability	long operating life	emission stability	emission intensity
structure of active layer			1998-145000 1998-233554		
electrode composition		1998-107318		1998-190063 1998-209498	1998-209495
electrode arrangement		1998-215034 1998-223930	1998-242518	1998-173230 1998-209499 1998-256602	1998-242515 1998-270757
structure of light emitting element	1998-135516 1998-242586 1998-247761		1998-135514 1998-256668		1998-012923 1998-247745 1998-256597

participants identify lines and columns

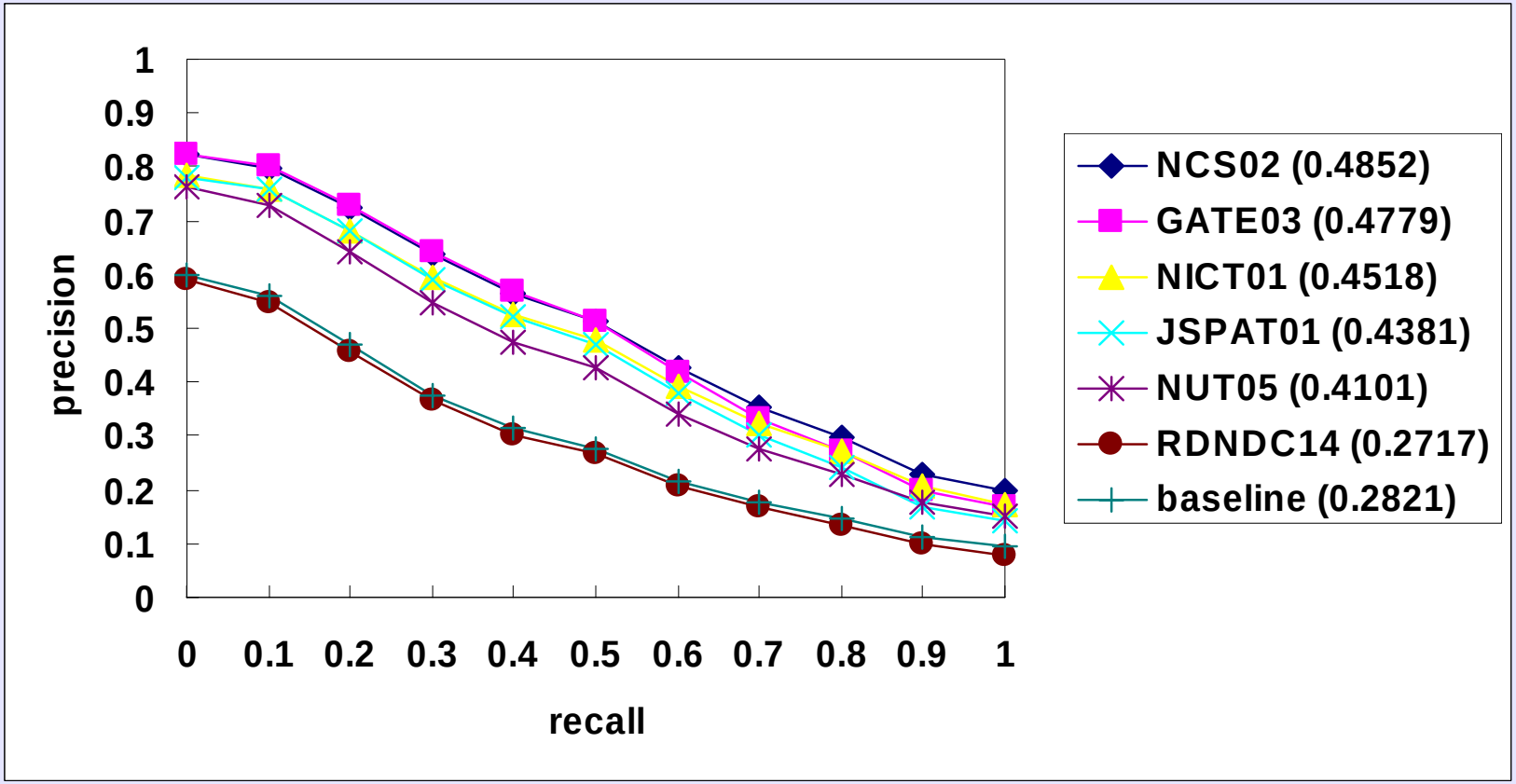
Patent full text vs. abstracts vs. Claims



*abs=author abstracts, jsh=professional abstracts

Search on patent fulltext using sophisticated IR models worked better than any other conditions

Results (exact match)

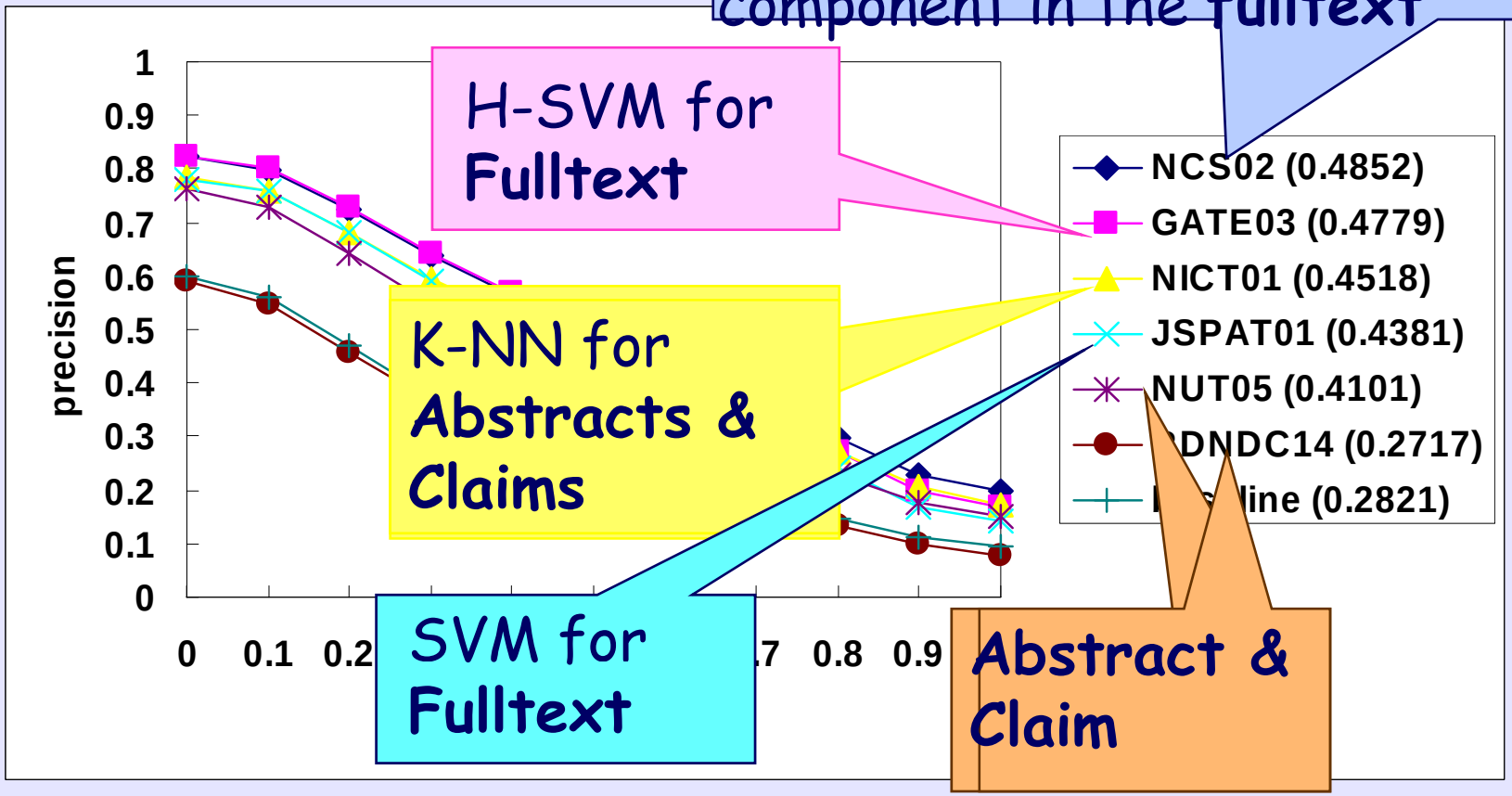


NTCIR

Results (e

Hybrid classifier/Naïve Bayes

Different Classifier for each component in the **fulltext**



WEB Retrieval

(A) Informational Retrieval Task

(B) Navigational Retrieval Task

- "Known Item Search". representative pages

(C) Geographical Task

(D) Topical Classification Task
retrieval result classification, eg.using clustering

Documents:

- 'NW100G-01' (100GB Web pages crawled in 2001 from "*.jp")
- 'NW1000G-04' (1.36TB Web pages crawled in 2004 from "*.jp")

Question Answering

1. **QA Challenge** on a language
   - Information Access Dialogues (NTCIR-3,-4, -5)
   - Natural (Real) Qs (no answer type limitation) (NTCIR-6)
2. **CLQA**
   - Factoid (NE) (NTCIR-5,-6)
   - Complex Questions (NTCIR-7)

Series of Question Situation Settings (User's Task)

1. **Collecting information about a particular topic**
   - One (hidden) global topic and series of Qs on subtopics of the global topic
2. **Browsing along transitive interests**
   - Topic or focus of the Qs are shifting through the interaction of the user and system.
   - Local coherence with the previous Q only

Example of Series of Questions

- What genre does the "**Harry Potter**" series belong to?
- Who is the author?
- Who are the main characters in that series?
- When was the first volume published?
- What title does it have?
- How many volumes were published by 2001?
- How many languages has it been translated into?
- How many copies have been sold in Japan?

Series 02: Gathering Type

Example of Series of Questions

- Where was Wuhan University?
- Which train station is the nearest?
- Who is the actor who visited the university?
- What is the movie he was featured in that was released in the New Year season of 2001?
- What is the movie starring Kevin Costner released in the same season?
- What was the subject matter of that movie?
- What role did Costner play in that movie?

Series 24: Browsing Type

QA: Lessons Learned

Tested for Simulated Interaction
- anaphora resolution, context
- inf gathering >> browsing, but improved

Return one set of All the answer:
- Context
- Answer Granularity
- Level of requiredness : Answer Score
- Answer Set

Complex Questions like asking definition, who, how, etc.
- More needed to investigate for automatic evaluation

NTCIR

Complex QA Evaluation criterion

- Human evaluation measure
  - Level A: System answer has almost the same contents as one of the correct answers.
  - Level B: System answer includes the contents of one of the correct answers.
  - Level C: System answer includes some part (not all one) of the contents of the correct answers.
  - Level D: System answer includes no information of any of the contents of the correct answers.

CLQA : Lessons Learned

- Factoid (esp. NE) QA can be a fundamental module for further CLIA especially among the languages with different scripts
- Major source of the performance drop was poor retrieval modules in QA systems. ➔ need collabolation with IR groups
- OOV

Text Summarization

1. **Text Summarizastion Challenge** on a language
   - Single document (NTCIR-2,-3)
   - Multidocument (NTCIR-3, -4)
2. **Summarization-based metrics used in QA (NTCIR-6, -7)**

Text Summarization Challenge
- Two types of summarization -

- Extraction
  - Extracting important sentences from document sets
    length: # of sentences
- Abstraction
  - Producing summaries from document sets
    length: # of characters

Two lengths: short, long

Automatic Extract Evaluation → Reusable Summarization Test Collection

See. Hirao (COLING 2004)

Opinion Analysis Roadmap

Genre	Subjectivity	Holder	Polarity	Strength
News	NTCIR-6	NTCIR-6	NTCIR-6	
Blog	NTCIR-7	NTCIR-7	NTCIR-7	
Cross-genre	NTCIR-8	NTCIR-8	NTCIR-8	NTCIR-8

Stakeholde	Temporal	Language	Granuality	Application
		C,J,E	single-sent	Summarization
NTCIR-7		C,J,E	clause	QA
NTCIR-8	NTCIR-8		multi-sent	Opinion tracking
		CJE	document	Consistency checking
				Trend

Corpus-Centered Evaluation and Collaborative Research

1. TMREC: Term Recognition (NTCIR-1)
2. MuST: Multimodal Summarization for Trend Information (NTCIR-5, -6)

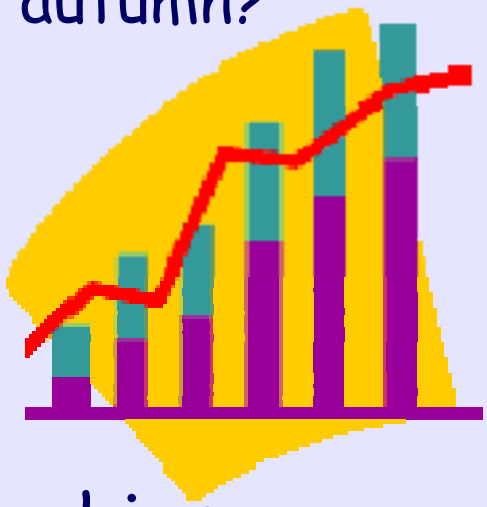
Multimodal summarization for Trend Information

Queries on trends

"How the price of gasoline shifted during the year?"

"What the situation has been in the PC market?"

"How terrible the typhoons were last autumn?"



Concise, plain text

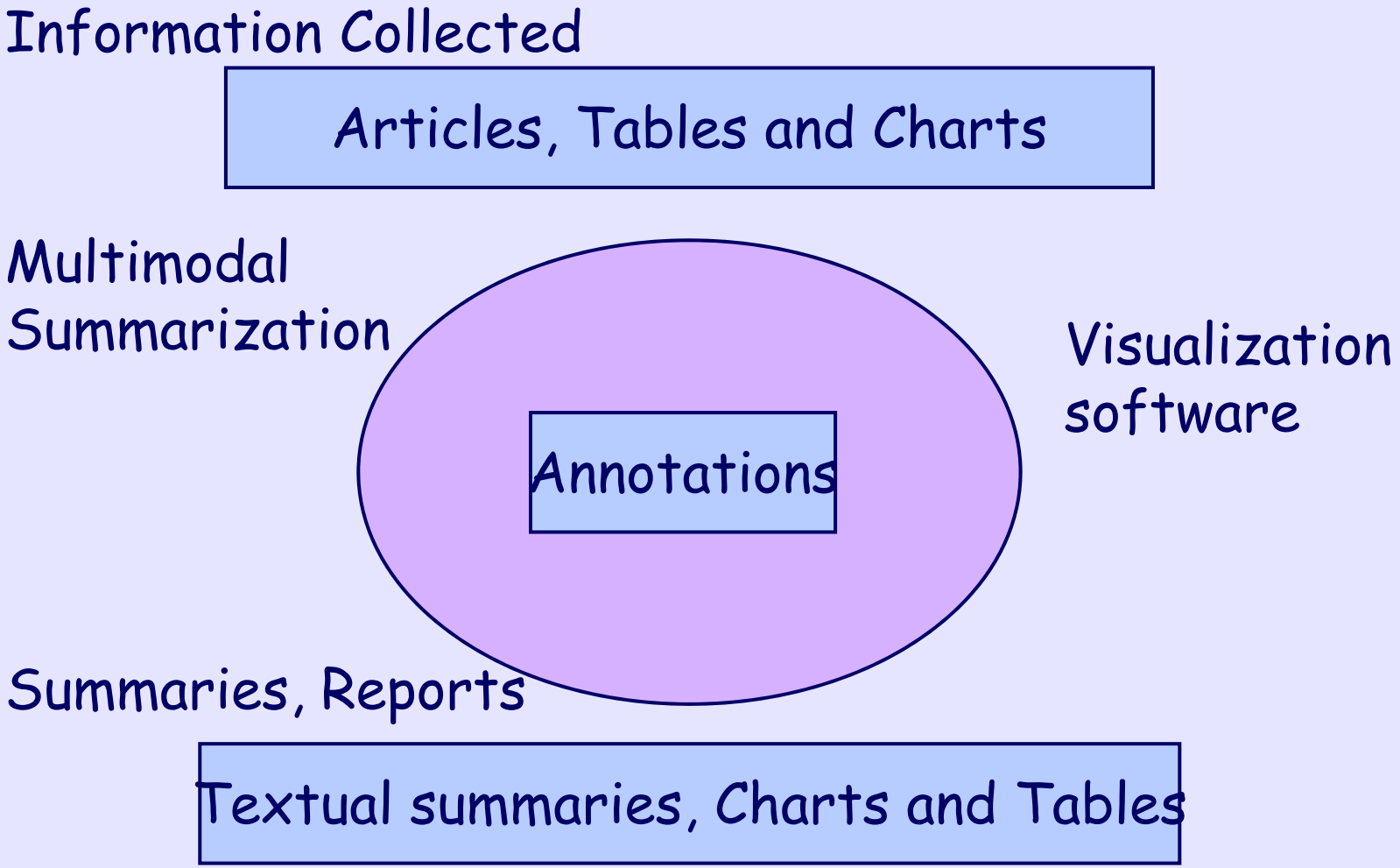
Information graphics

Multimedia presentation

text including references to graphics

graphics annotated with text

The Roles of Data Set

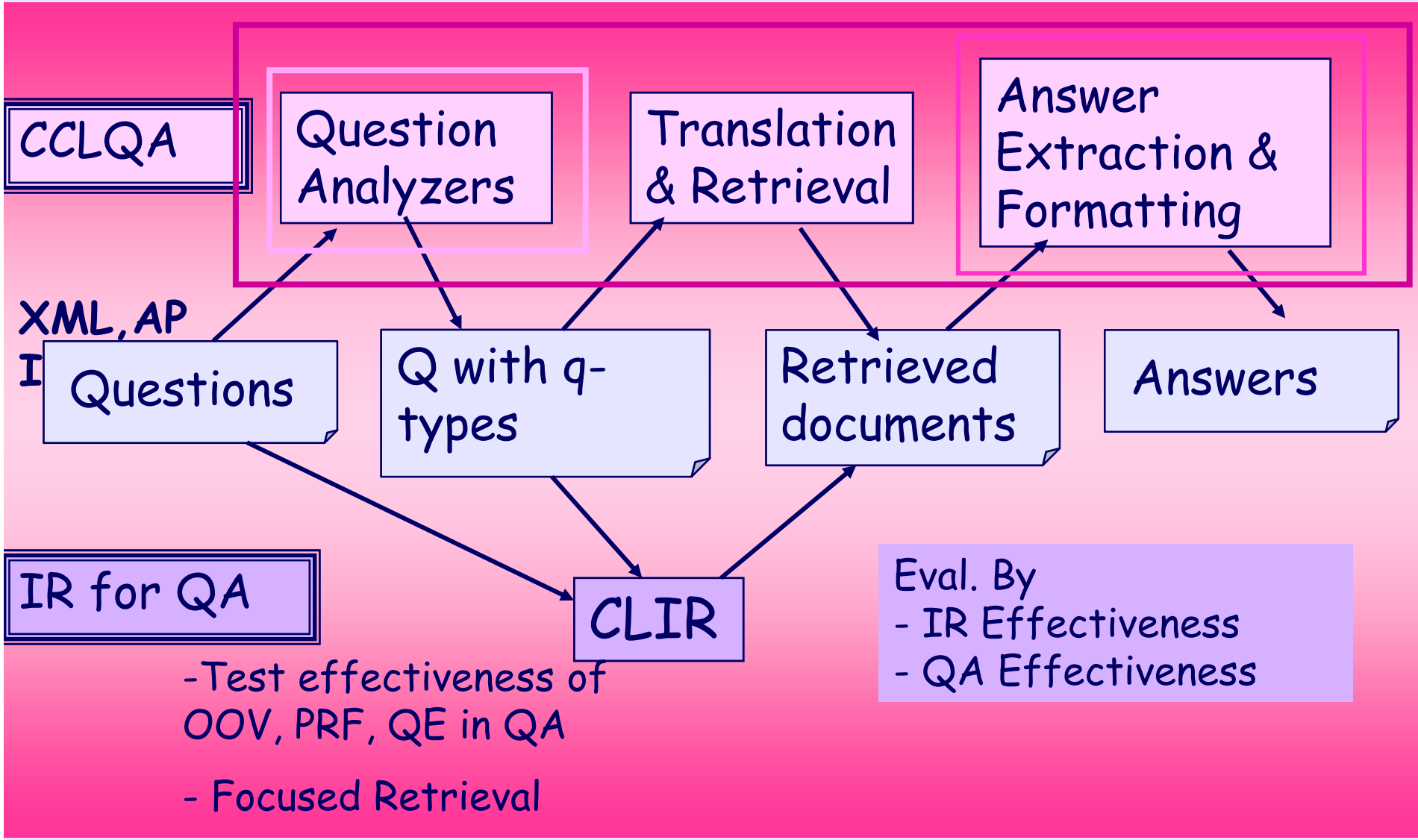


NTCIR

Road map

- What is NTCIR
- Leason learned from past NTCIRs
- Brief Introction to NTCIR-7
- Conclusion

NTCIR-7: Advanced CLIA



ACLIA: Test Collection

Language	Corpus Name	Time Span
CS	Xinhua	1998-2001
CS	Lianhe Zaobao	1998-2001
CT	CIRB020 & CIRB040	1998-2001
JA	Mainichi Shinbun	1998-2001

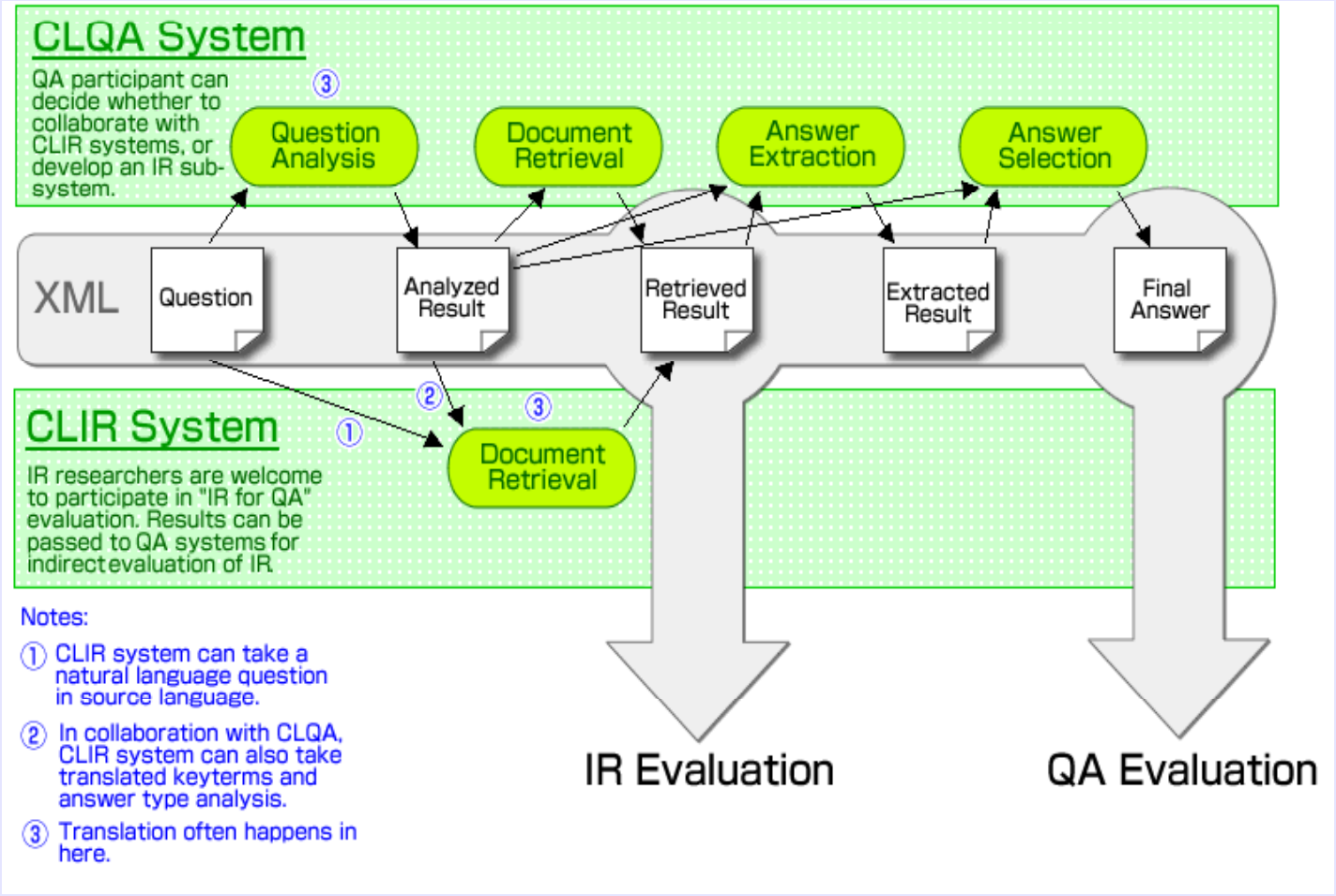
100 Topics in CS, CT, JA and their English translation

DEFINITION, EVENT, BIO, RELATION

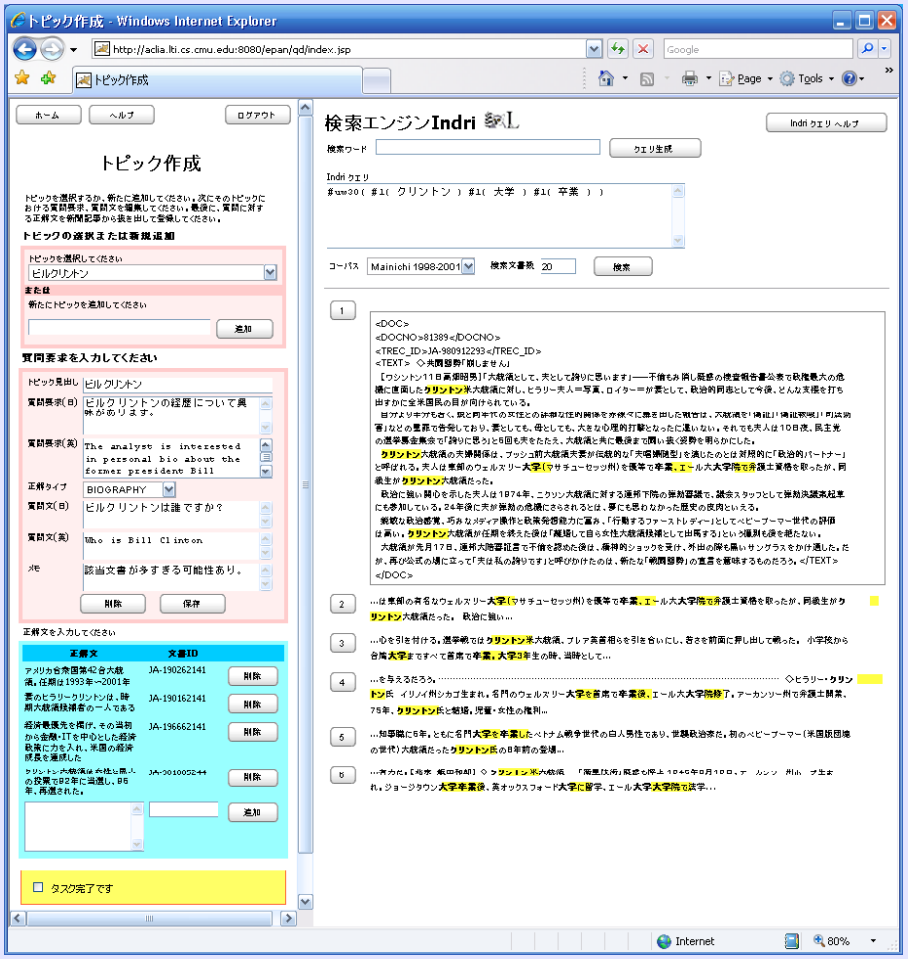
CLQA and IR4QA used the same topics

<TITLE> <Q-TYPE> not released
<QUEstion><NARRATIVE> released

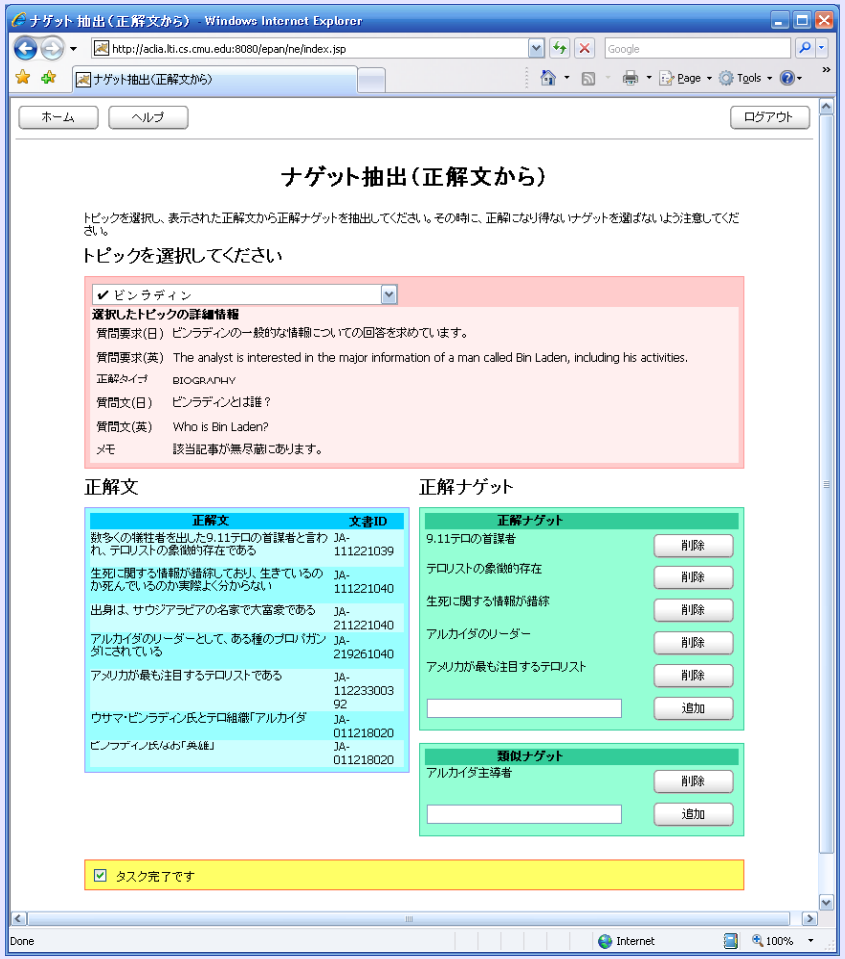
ACLIA: Evaluation



ACLIA: Evaluation EPAN tool



ACLIA: Evaluation EPAN tool



CCLQA:

Nugget Pyramid

IR4QA:

MAP

MS nDCG

Q-Measure

(preference-based)

NTCIR-7: UGC (Blog)

David K Evans (NII -> Amazon Japan)
Yohei Seki (Toyohashi U Tech -> Columbia U)

LunWei Ku (National Taiwan Univ)
Le Sun (Chinese Academy of Science)
Hsin-Hsi Chen (National Taiwan Univ)
Noriko Kando (NII)

Opinion Analysis - Roadmap

Genre	Subjectivity	Holder	Polarity	Strength
News	NTCIR-6	NTCIR-6	NTCIR-6	
Review	NTCIR-7	NTCIR-7	NTCIR-7	NTCIR-7
Blog	NTCIR-8	NTCIR-8	NTCIR-8	NTCIR-8

Stakeholder	Temporal	Language	Granuality	Application
		Chinese	single-sent	Summarization
NTCIR-7		English	clause	QA
NTCIR-8	NTCIR-8	Japanese	multi-sent	Opinion tracking
		CJE	document	Consistency checkir
				Trend

Chinese, Japanese, English

NTCIR-7: UGC (Blog)

- **Documents:**
  Crawled Blog posts + Comments (July – Sept, 2007. 6 weeks) CCKEJ
- **CLIR on Blog (CLIRB) C,C,J,E,(K?) any other q?**
  - Informational Search Task for Opinion-Focused search requests
  - 50 topics, 4 grade relevance judgments
- **Multilingual Opinion Analysis (MOAT)** TraditionalC,J,E
  - selecting relevant documents from ~30 topics used in CLIRB.
  - Following Roadmap, but change the genre
  - Relevant, Opinionated, Polarity (Pos, Neg, Nue), Holder, Stakeholder (Object), ??Strength??

NTCIR-7: MOAT (on News)

- **Documents:**
  NEWS CCEJ
- **CLIR on Blog (CLIRB) Cancelled**
- **Multilingual Opinion Analysis (MOAT)**
  - TraditionalC,Simplifed C, J,E
  - selecting relevant documents from ~25 topics used in ACLIA
  - Following Roadmap, but change the genre
  - Relevant, Opinionated, Polarity (Pos, Neg, Nue), Holder, Stakeholder (Object), ??Strength??

MOAT Participants

Beijing university of posts and telecomunications
Chinese Academy of Sciences(NLPR-IACAS)
City University of Hong Kong
CUHK(The Chinese University of Hong Kong)-PolyU(The Hong Kong Polythechnic University)-Tsinghua(Tsinghua University)
DAEDALUS, S.A.
Dalian University of Technology
Hiroshima City University
Information and Communications University
Keio University
Louisiana State University(University of Maryland College Park)
National Taiwan University
NEC
NEU Natural Language Processing Lab
Peking University
Peking University(ICL)
Pohang University of Science and Technology
SICS - Swedish Institute of Computer Science
Technical University of Darmstadt
The Graduate University for Advanced Studies(SOKENDAI).
Tornado Technologies Co., Ltd., Taiwan.
Toyohashi University of Technology
University of Neuchatel
University of Sussex
Yuan Ze Univ.

80+ registerd, 30+ resigned when docs were changed, 42 registered to News MOAT, 24 sugmitted

NTCIR-7: Focused Domain (Patent)

Atsuhi Fujii (Univ Tsukuba)

Taiich Hashimoto (Tokyo Insti Tech)
Makoto Iwayama (Tokyo Insti Tech/ Hitach)
Hidetsugu Nanba (Hiroshima City Univ)
Masao Utiyama (NICT),
Mikio Yamamoto, U Tsukuba)
Takehito Utsuro (U Tsukuba)

NTCIR-7: Focused Domain (Patent)

Documents:

10 Yrs Japanese Patent Application (NTCIR4-5)

10 Yrs USTPO Patents (NTCIR6)

Parallel Sentence Data (1.8 M sentences JE Pairs)

Scientific Paper Abstracts (NTCIR 1-2)

Patent Translation (PATMT) MT is key for CLIR

Training: 1993-2000, Test: 2001-2002 One Ref Trans good??

Intrinsic Eval. ;BLEU, human assessments

Extrinsic Eval: CLIR task-based

Patent Mining (PATMN) Cross-Genre PAT & Scientific

Classify Paper Abstracts in to IPC Classes

ML approach: Classsify Absts to IPC Class

IR Apprach: use invalidity search system to find relevant Patent, then assign IPCs to Paper Absts.

NTCIR

Patent classification and mining at NTCIR

Organizers:
Makoto Iwayama (Hitachi Ltd/Tokyo Institute of Technology)
Hidetsugu Nanba (Hiroshima City University)
Taiichi Hashimoto (Tokyo Institute of Technology)
Atsushi Fujii (University of Tsukuba)
Noriko Kando (National Institute of Informatics)

NTCIR

Goal: Automatic generation of patent maps.

Example: Blue light-emitting diodes

Given

Problems to be solved

Solutions

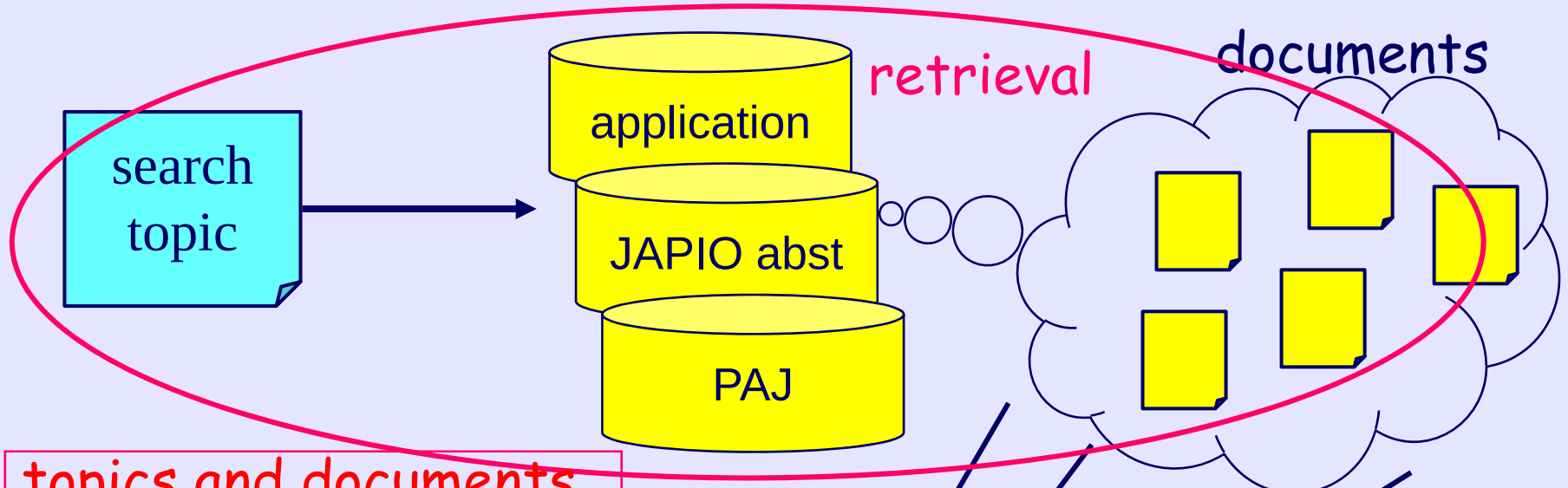
	Crystalline	Reliability	Long operating life	Emission stability	Emission intensity
Structure of active layer			1998-145000 1998-233554		
Electrode composition		1998-107318		1998-190063 1998-209498	1998-209495
Electrode arrangement		1998-215034 1998-223930	1998-242518	1998-173230 1998-209499 1998-256602	1998-242515 1998-270757
Structure of light emitting element	1998-135516 1998-242586 1998-247761		1998-135514 1998-256668		1998-012923 1998-247745 1998-256597

Systems automatically identify rows and columns

History

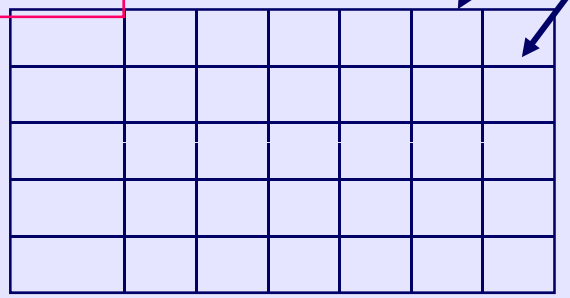
- NTCIR-4 (2003-2004): Patent-map-creation subtask
  - Direct approach to creation of patent maps
  - Hard tasks and insufficient evaluation
- NTCIR-5 (2004-2005): Classification subtask
  - Categorize patents to pre-defined categories called F-terms (multi-faceted and structured)
  - Relatively small number of test documents
  - Evaluate only strict matches in F-term hierarchy
- NTCIR-6 (2006-2007): Classification subtask
  - Increased the number of documents and topics (108 topics)
  - Evaluate partial matches in F-term hierarchy
- NTCIR-7 (2007-2008): Mining subtask

Feasibility Study: automatic patent map generation at NTCIR-4 (2003-2004)



topics and documents in NTCIR-3 collection

Patent-map creation = Multi-faceted patent clustering

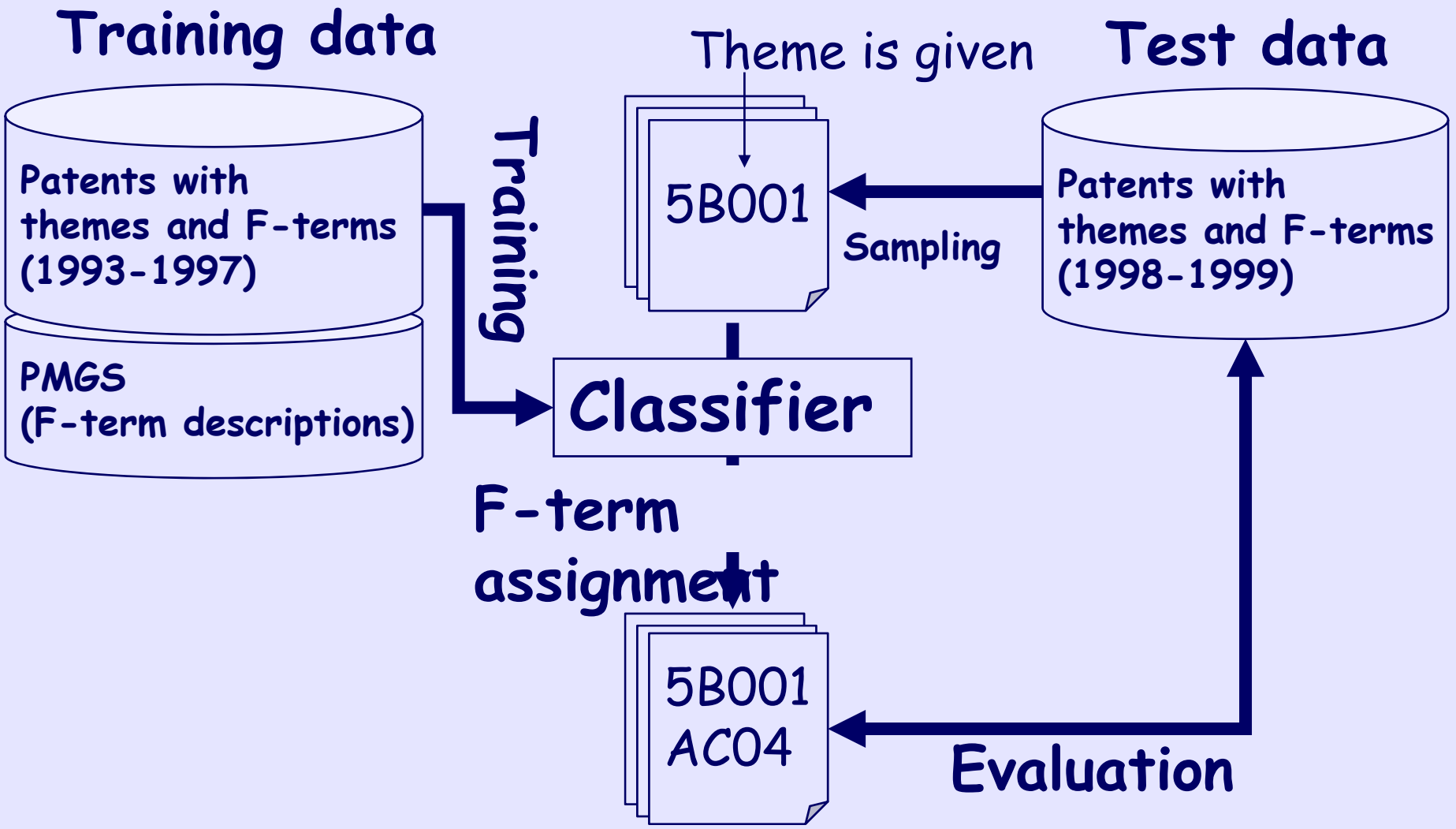


visualization

classification

multi-dimensional matrix

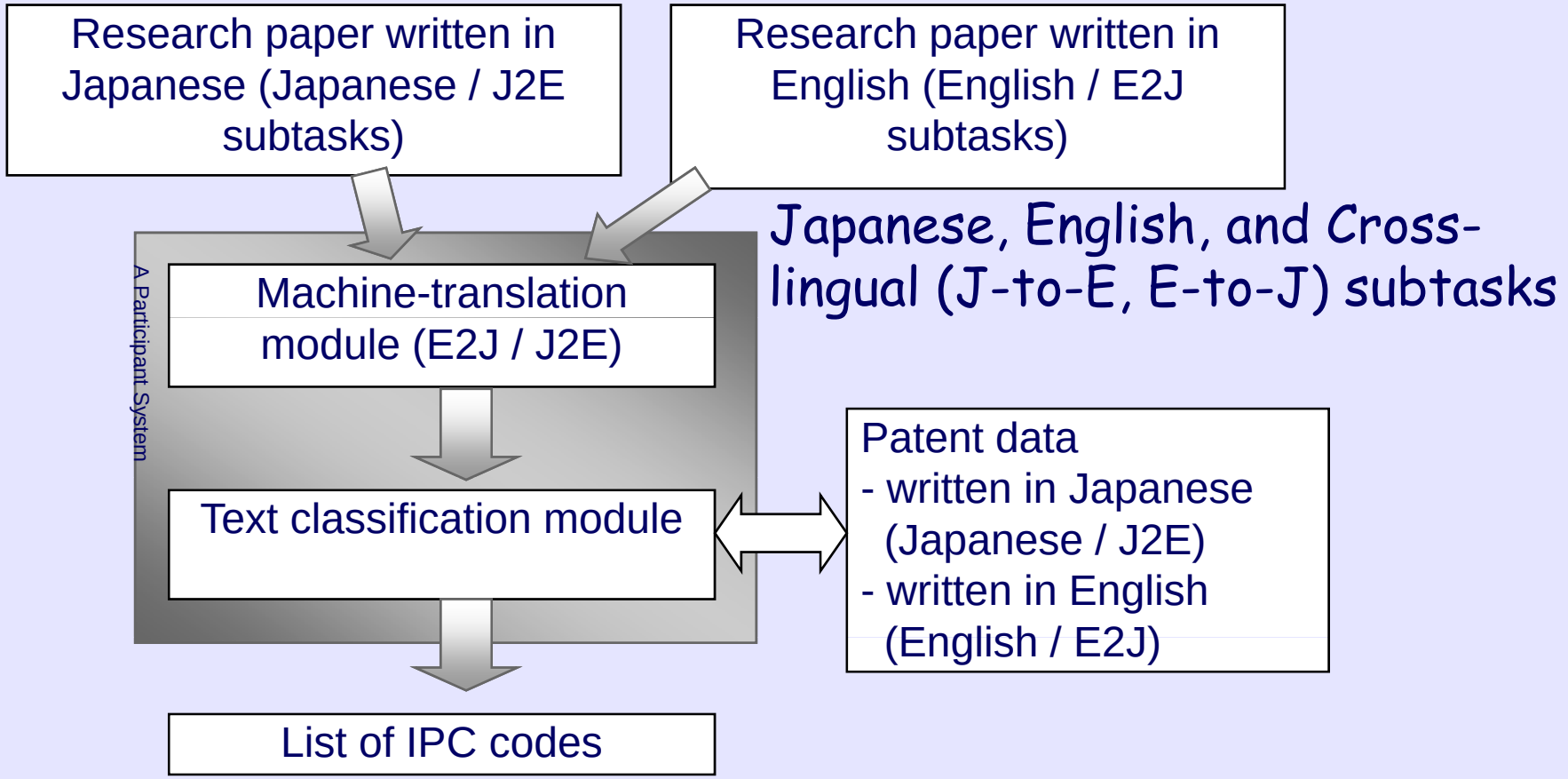
Classification task overview



NTCIR

Patent mining at NTCIR-7 (2007-2008)

Searches and/or classifying patents and scientific papers into IPC



Nanba, Fujii, Iwayama, and Hashimoto. “The Patent Mining Task in the Seventh NTCIR Workshop”, Patent Information Retrieval Workshop at CIKM 2008 (2008)

Summary of patent classification and mining

- Automatic clustering of patents into "problems" and "solutions" are quite feasible, but labeling and controlled evaluation need more investigation.
- Granularity of F-term is appropriate for patent map creation and becoming good.
- Patent minting of scientific papers and patents are practically needed. n-KNN and machine learning have promise
  - The test collections for classification are available for research purpose. The one for mining will be available to the public after Workshop Meeting

Patent machine translation at NTCIR

Organizers:
Atsushi Fujii (University of Tsukuba)
Masao Utiyama (NICT)
Mikio Yamamoto (University of Tsukuba)
Takehito Utsuro (University of Tsukuba)

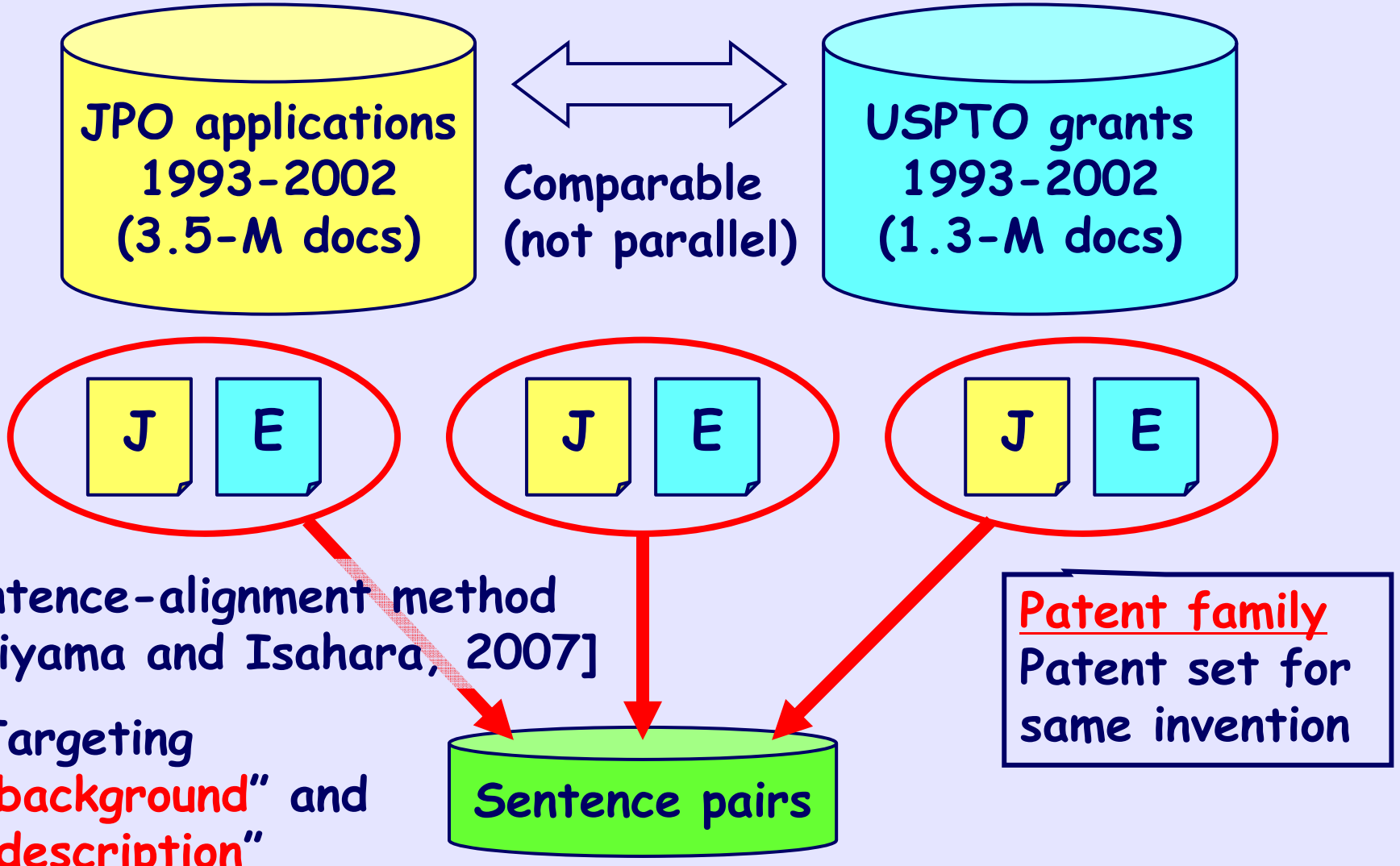
Fujii, Utiyama, Yamamoto, and Utsuro. “Toward the Evaluation of Machine Translation Using Patent Information”, AMTA 2008

NTCIR

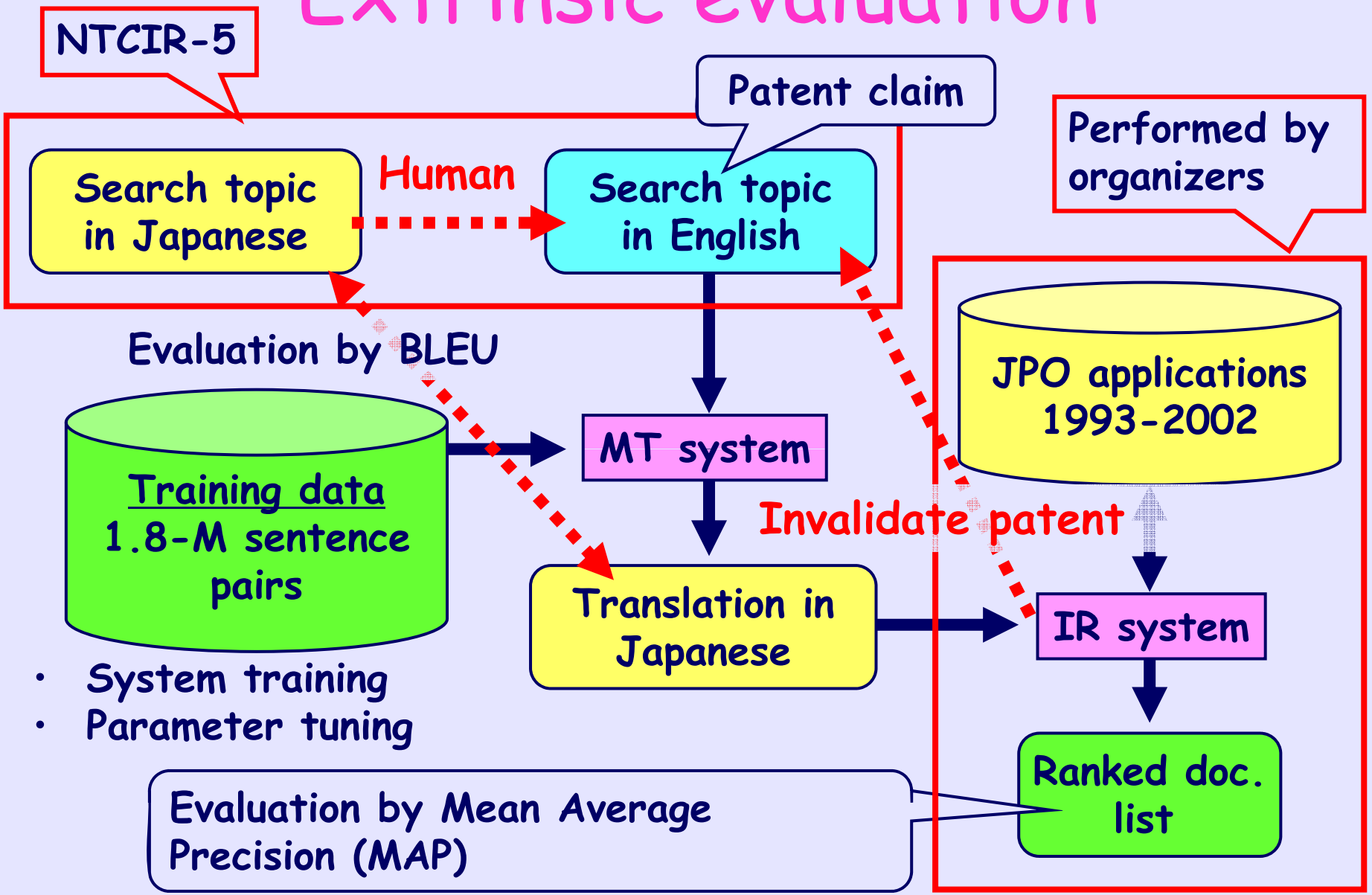
Patent machine translations at NTCIR-7 (2007-2008)

- Patent Machine Translation (MT) is realistic
  - Parallel corpora can potentially be produced from JPO/USPTO patent-document sets
  - Decoders for statistical MT (SMT) are available
- Two types of players
  - Organizer = Authors of this paper
    - Providing data, and evaluating participating MT systems
  - Participants = Research groups
    - They can use e.g., SMT and rule-based MT.
- Utility of patent MT
  - Cross-lingual patent retrieval
  - Filing patent applications in foreign countries

Producing parallel corpora



Extrinsic evaluation



NTCIR

Patent machine translation

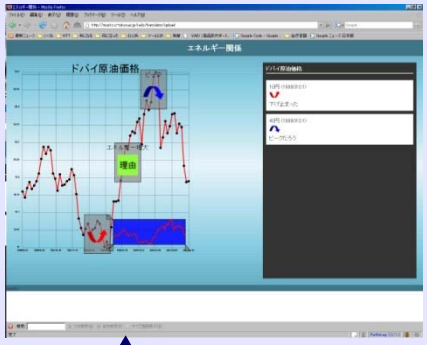
- Constructed a large test collection for J/E MT: USTPO and JPO with 10 years of full texts
- Large-scale sentence-alignment dataset (E-J sentence pairs)
- Statistical MT (SMT)* vs. rule-based MT
- Results demonstrated:
  - SMT is much better for CLIR
  - Rule-based MT is good for human evaluations
  - Human evaluations and creation of reference translations must be carefully done (in the real world, professional patent translators do use MT).
- Test collection will be available for research purpose after the workshop meeting

*SMT : a system automatically learns the translation rules from the given large-scale sentence pairs.

Multimodal summarization for Trend Information

Queries on trends
"How the price of gasoline shifted during the year?"
"What the situation has been in the PC market?"
"How terrible the typhoons were last autumn?"

Concise, plain text
Information graphics
Multimedia presentation
  text including references to graphics
  graphics annotated with text



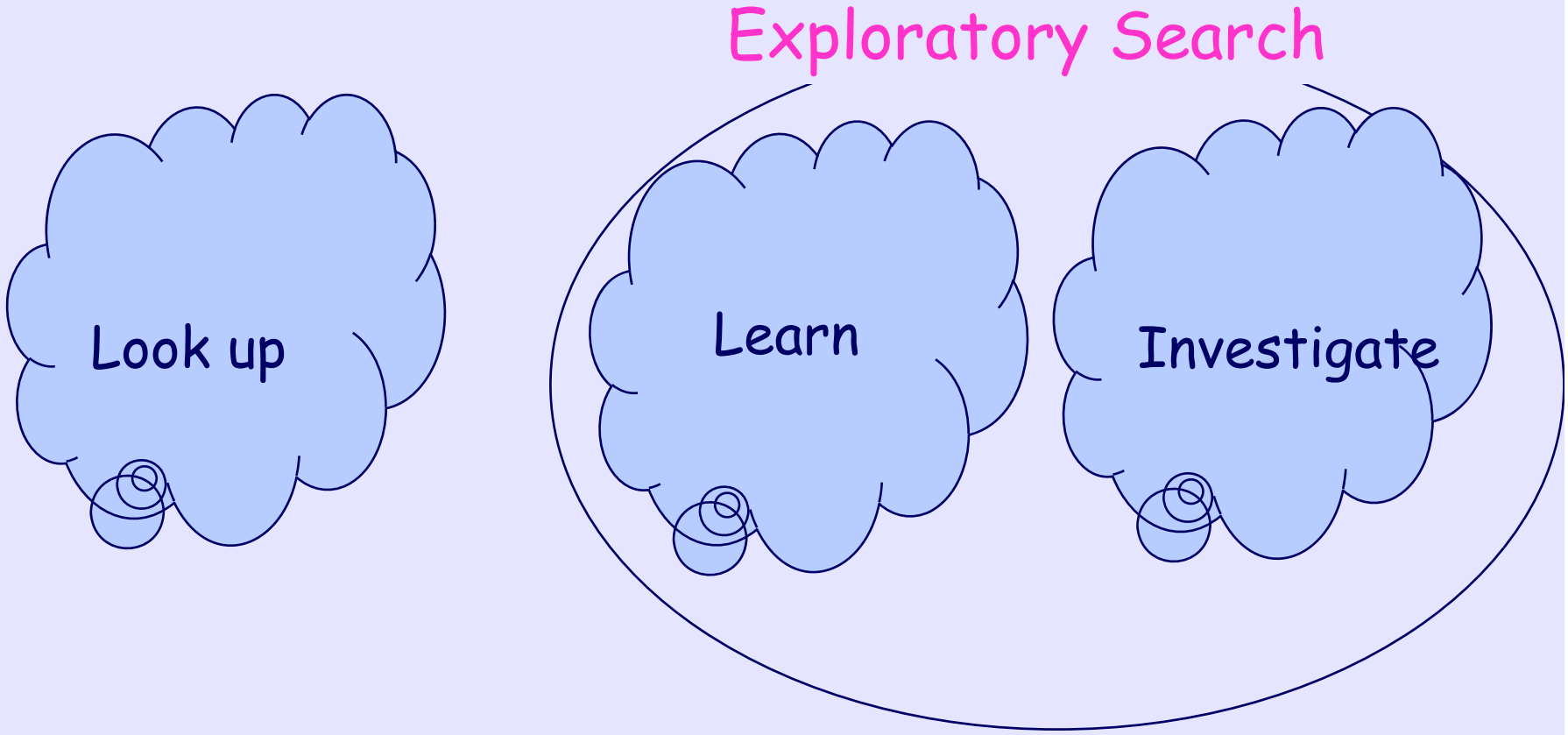
Visualization Platform

NTCIR-7 Workshop Meeting December 16-19, 2008 @ Tokyo

http://research.nii.ac.jp/ntcir/ntcir-ws7/meeting/

Past data: http://research.nii.ac.jp/ntcir/data/data-en.html
Proceedings: http://research.nii.ac.jp/ntcir/publication1-en.html

Types of Information Access



NTCIR

Call for NTCIR-8 task proposals

- Let's work together to construct a better infrastructure to encourage information-access research to move forward. Resources constructed in past NTCIRs are also available.

- Due to 30th November 2008
  - Write to Noriko Kando

Acknowledgments

- Japan Intellectual Property Association (JIPA)
- Industrial Property Cooperation Center, Japan
- Japan Parent Office
- Japan Patent Information Organization (JAPIO)
- Mainichi Newspaper
- NRI Cyber Patents
- PATOLIS
- Task organizers
- Participants and test-collections' users
- Information Retrieval Facility

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

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