Twitter: #ntcir10
Ust: ntcir10-kickoff

NTCIR-10 Kick-Off Event

2011.03.08

日本語セッション: 13:30-

English Session: 15:30-

Program

- About NTCIR
- About NTCIR-10
- Accepted Tasks
- Why participate?
- How to participate
- Important Dates
- Q & A

About NTCIR

Introduction

NTCIR: NII Testbeds and Community for Information access Research

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.

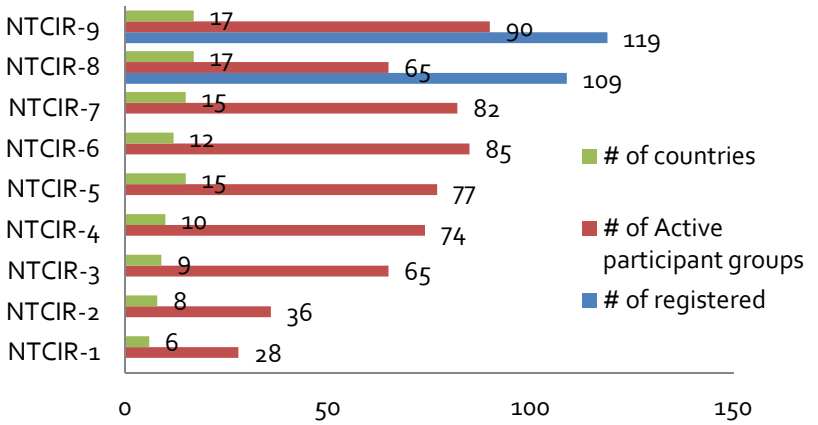
- Data sets, evaluation methodologies, and forum

Project started in late 1997

- Once every 18 months

Data sets (Test collections or TCs)

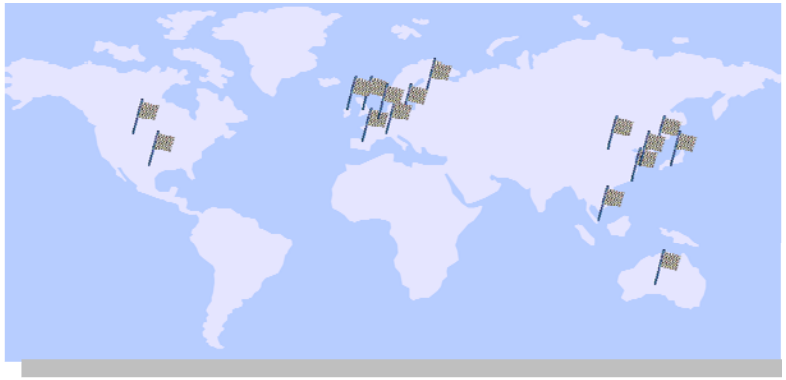
- Scientific, news, patents, web, CQA, Wikipedia, Entrance Exams
- Chinese, Korean, Japanese, and English



Tasks (Research Areas)

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

Community-based Research Activities



NTCIR

Information retrieval (IR)

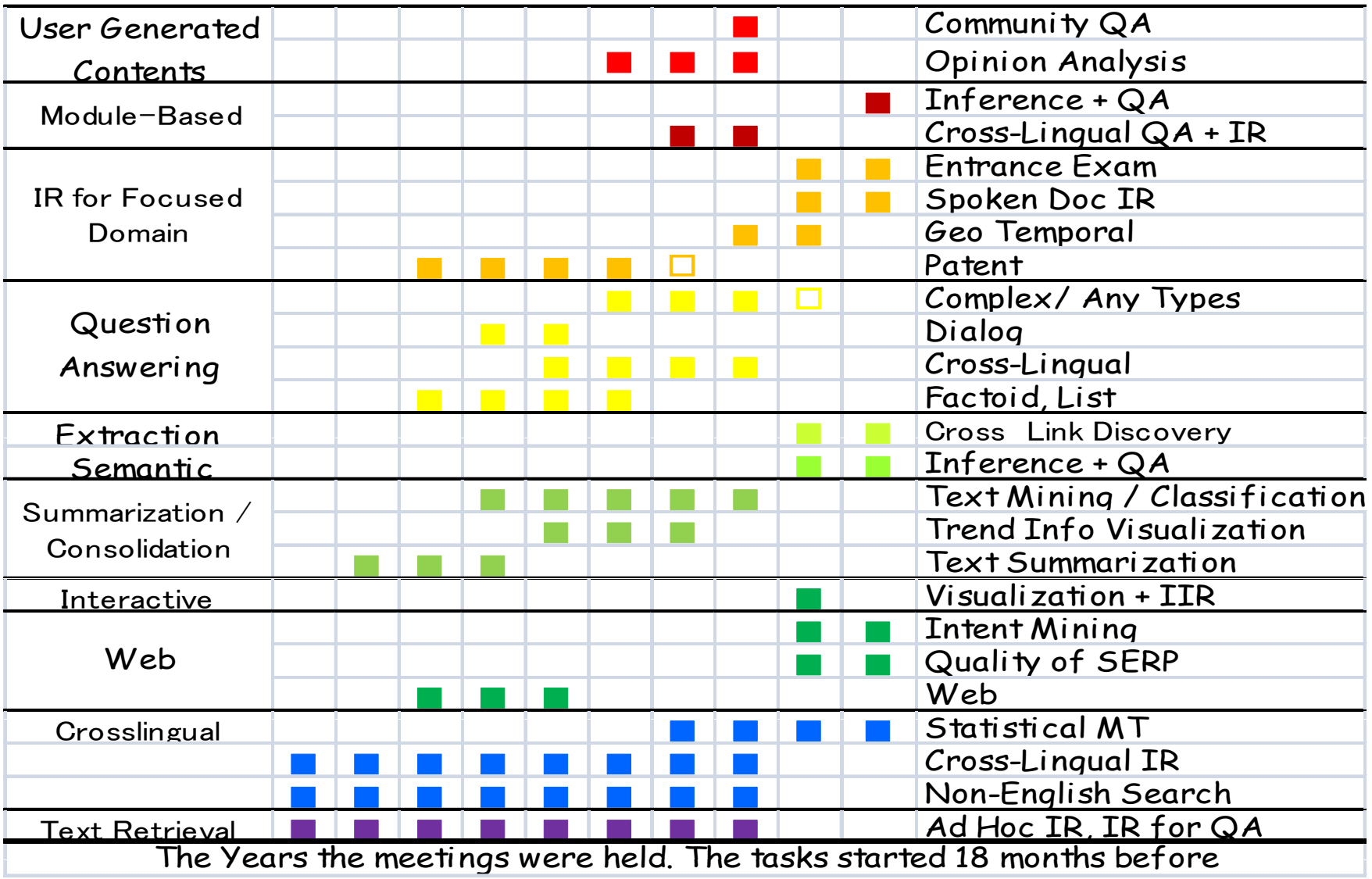
- Retrieve RELEVANT information from vast collection to meet users' information needs
- Using computers since the 1950s
- First CS uses human assessments as success criteria
  - Judgments vary
  - Comparative evaluations on the same infrastructure

Information access (IA)

Whole process to make information usable by users.

ex.: IR, text summarization, QA, text mining, and clustering

Tasks at Past NTCIRs

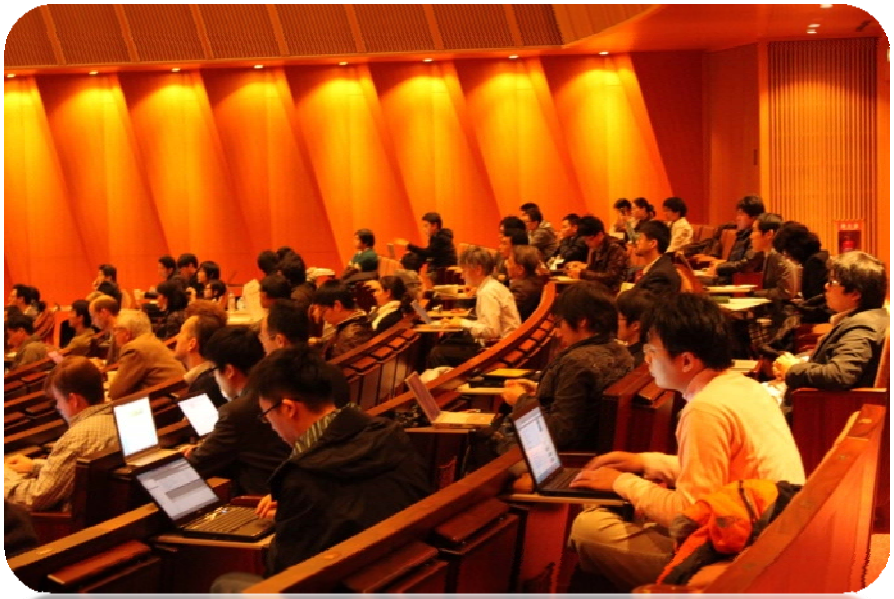


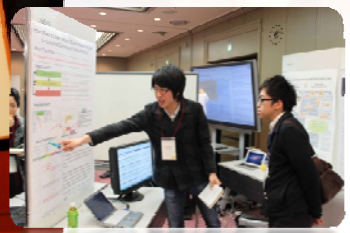
Procedures in NTCIR Workshops

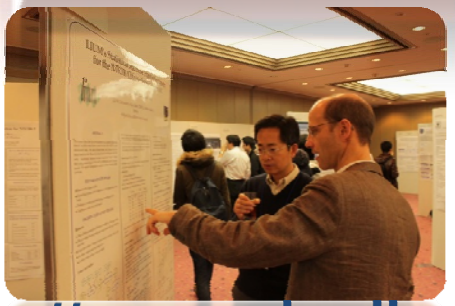
- Call for Task Proposals
- Selection of Task Proposals by Committee
- Discussion about Experimental Designs and Evaluation Methods (can be continued to Formal Runs)
- Registration to Task(s)
  - Deliver Training Data (Documents, Topics, Answers)
    - Experiments and Tuning by Each Participants
  - Deliver Test Data (Documents and Topics)
    - Experiments by Each Participants
- Submission of Experimental Results
- Pooling the Answer Candidates from the Submissions, and Conduct Manual Judgments
- Return Answers (Relevance Judgments) and Evaluation Results
- Workshop Meeting Discussion for the Next Round

NTCIR

NTCIR: Workshop Meeting







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

NTCIR

NTCIR-10 Organization

<NTCIR-10 Program Co-Chairs>
- Hideo Joho （University of Tsukuba, Japan）
- Tetsuya Sakai （Microsoft Research Asia, China)

<EVIA2013 Co-Chairs>
-Ruihua Song (Microsoft Research Asia, China)
-William Webber (University of Maryland, USA)

<NTCIR-10 General Co-Chairs>
- Noriko Kando (NII, Japan)
- Tsuneaki Kato (The University of Tokyo, Japan)
- Douglas W. Oard (University of Maryland, USA)
- Mark Sanderson (RMIT University, Australia)
-ex-general chair: Eiichiro Sumita (NICT)

-48 Task Organizers from all over the world

NTCIR

NTCIR Test Collections

Test Collections = Docs + Topics/Questions + Answers

	Ad Hoc/ CLIR [Scientific Abstracts] (日本語・英語検索)	Chinese IR (中国語検索)	CLIR [News] (言語横断検索)	CLQA (言語横断・質問応答)	MuST (「動向情報の要約と可視化」)	OPINION (意見分析)	PATENT (特許検索・分類)	QAC (質問応答)	TMREC (用語抽出)	TSC (要約)	WEB
			ACLIA (高度言語横断情報アクセス)			MOAT (多言語意見分析)	PATMN (特許マイニング) / PATMT (特許翻訳)				
NTCIR-1	NTCIR-1 Ad Hoc/ CLIR	-	-	-	-	-	-	-	NTCIR-1 TMREC	-	-
NTCIR-2	NTCIR-2 Ad Hoc/ CLIR	CIRB010	-	-	-	-	-	-	-	NTCIR-2 SUMM	-
NTCIR-3	-	-	NTCIR-3 CLIR	-	-	-	NTCIR-3 PATENT 特許検索	NTCIR-3 QA	-	NTCIR-3 SUMM	NTCIR-3 WEB
NTCIR-4	-	-	NTCIR-4 CLIR	-	-	-	NTCIR-4 PATENT 特許検索	NTCIR-4 QA	-	NTCIR-4 SUMM	NTCIR-4 WEB
NTCIR-5	-	-	NTCIR-5 CLIR	NTCIR-5 CLQA	-	-	NTCIR-5 PATENT 特許検索 分類	NTCIR-5 QA	-	-	NTCIR-5 WEB
NTCIR-6	-	-	NTCIR-6 CLIR	NTCIR-6 CLQA	NTCIR-6 MuST	NTCIR-6 OPINION	NTCIR-6 PATENT 特許検索 分類	NTCIR-6 QA	-	-	-
NTCIR-7	-	-	NTCIR-7		NTCIR-7	NTCIR-7	NTCIR-7 / NTCIR-7	-	-	-	-

Available to Non-participants for Research Purpose

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

Interactive/Exploratory search

QA types at topic crea

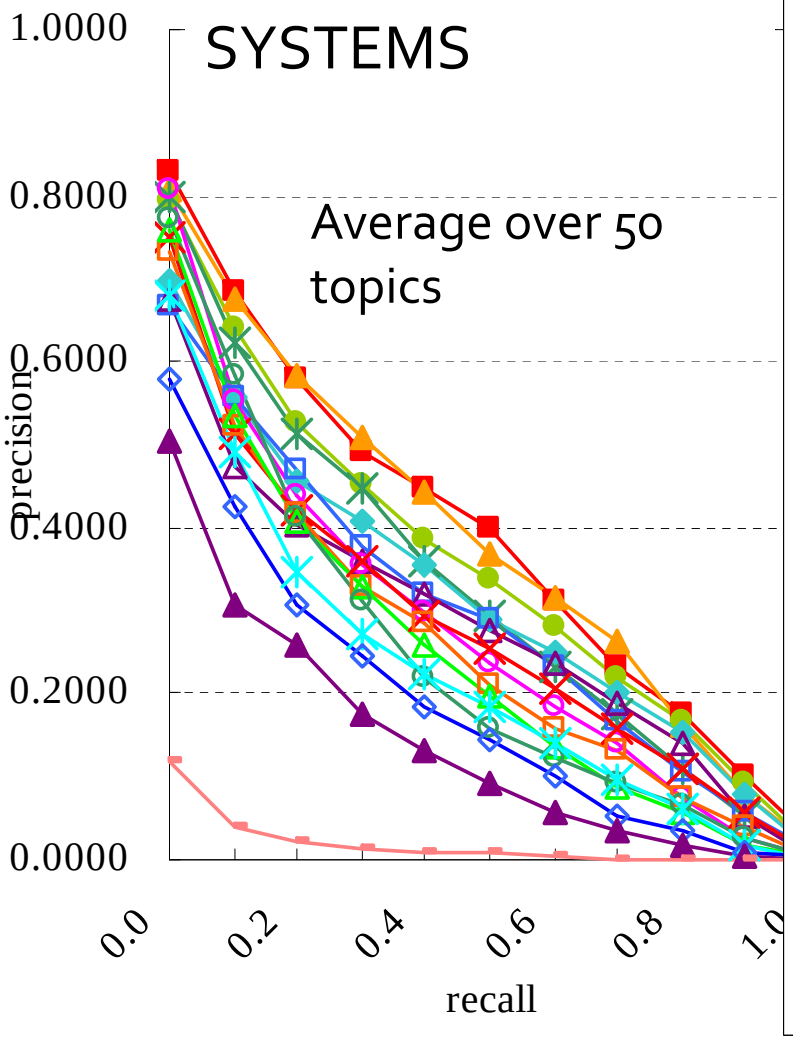
Forum for Researchers and Other Experts/users

Idea Exchange

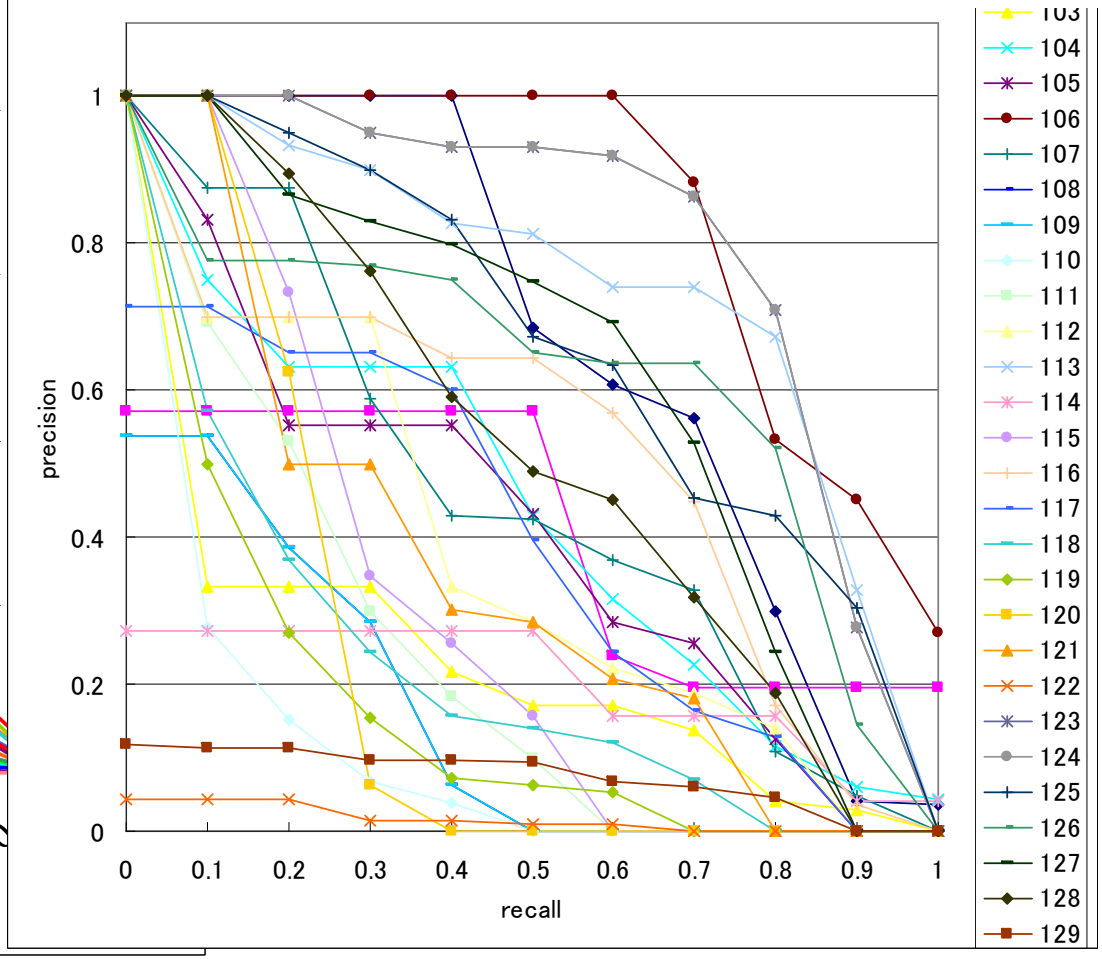
Discussion/Investigation on Evaluation methods/metrics

Difficulty of retrieval varies with topics

Effectiveness across SYSTEMS



Effectiveness across TOPICS on system

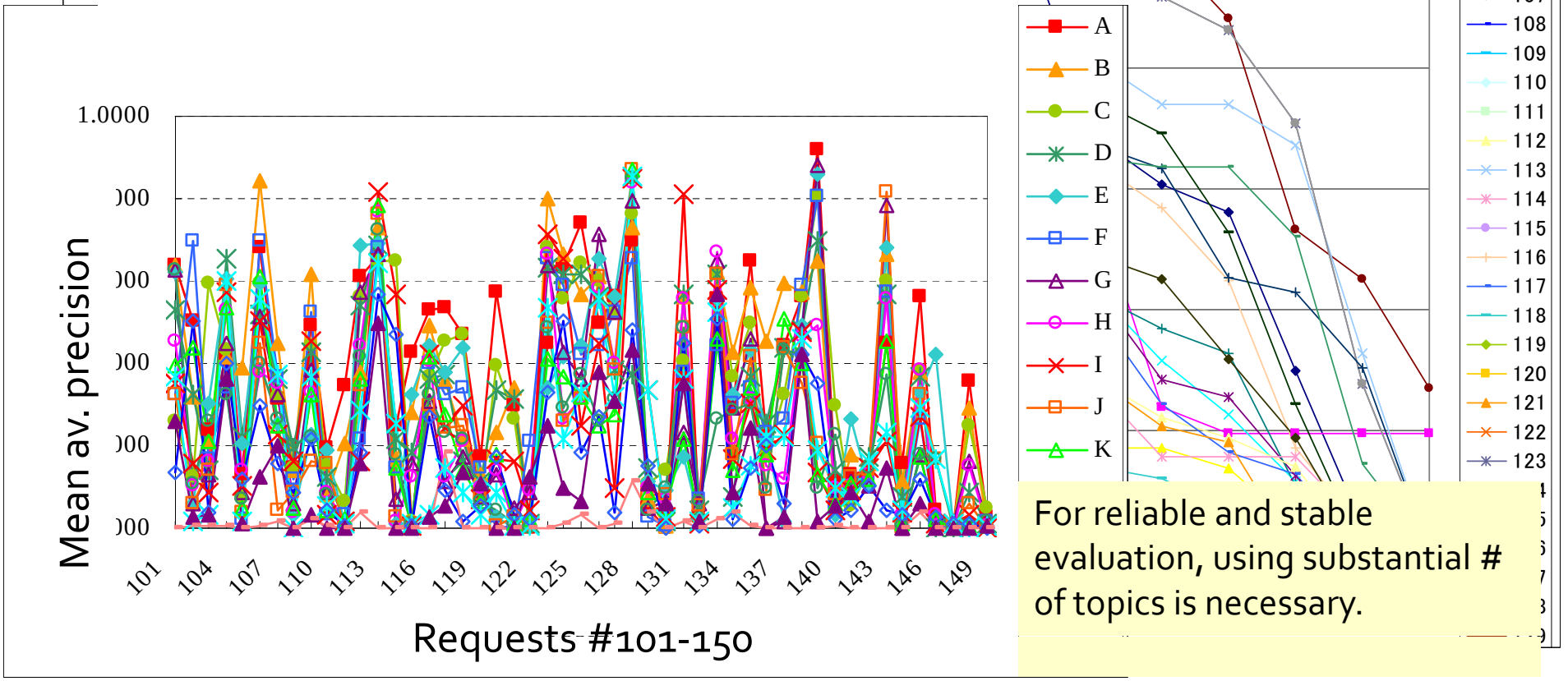


Difficulty of retrieval varies with topics

Effectiveness across SYSTEMS

Effectiveness across TOPICS on system

"Difficult topics" vary with systems

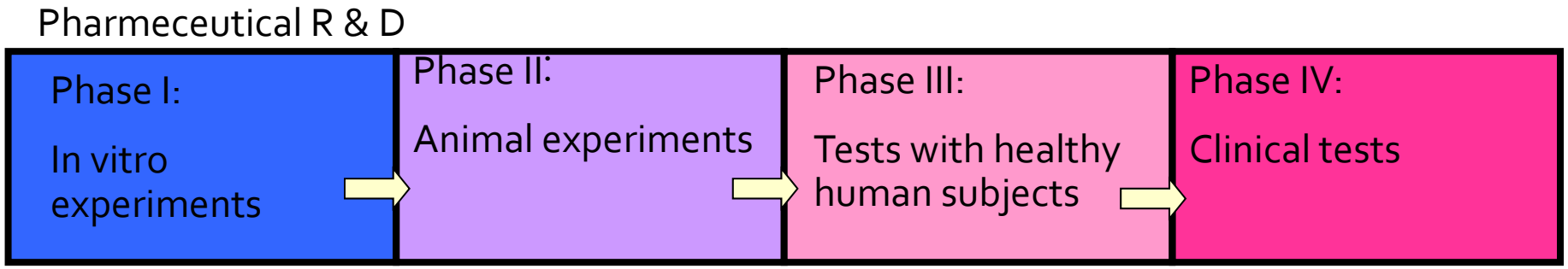


For reliable and stable evaluation, using substantial # of topics is necessary.

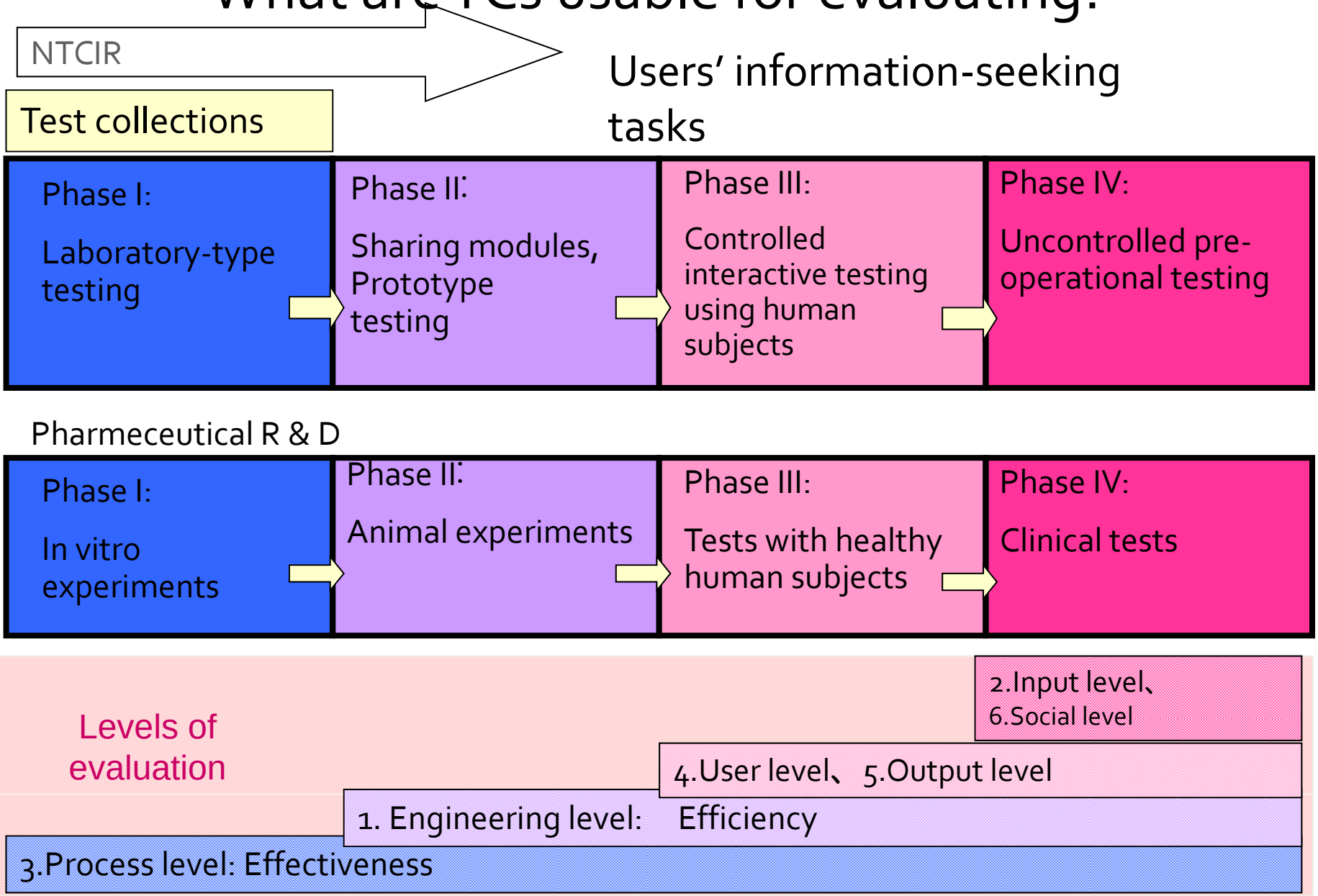
IR Systems Evaluation

- 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)

What are TCs usable for evaluating?



What are TCs usable for evaluating?



- Information Seeking Task
  - document types + user community
  - user's situation, purpose of search, realistic

Experiments are Abstraction of the Real World Tasks.
Trade-off between "reality" and "contorable"

- Testing & Bench marking

To learn how and why the system works better (worse) than others

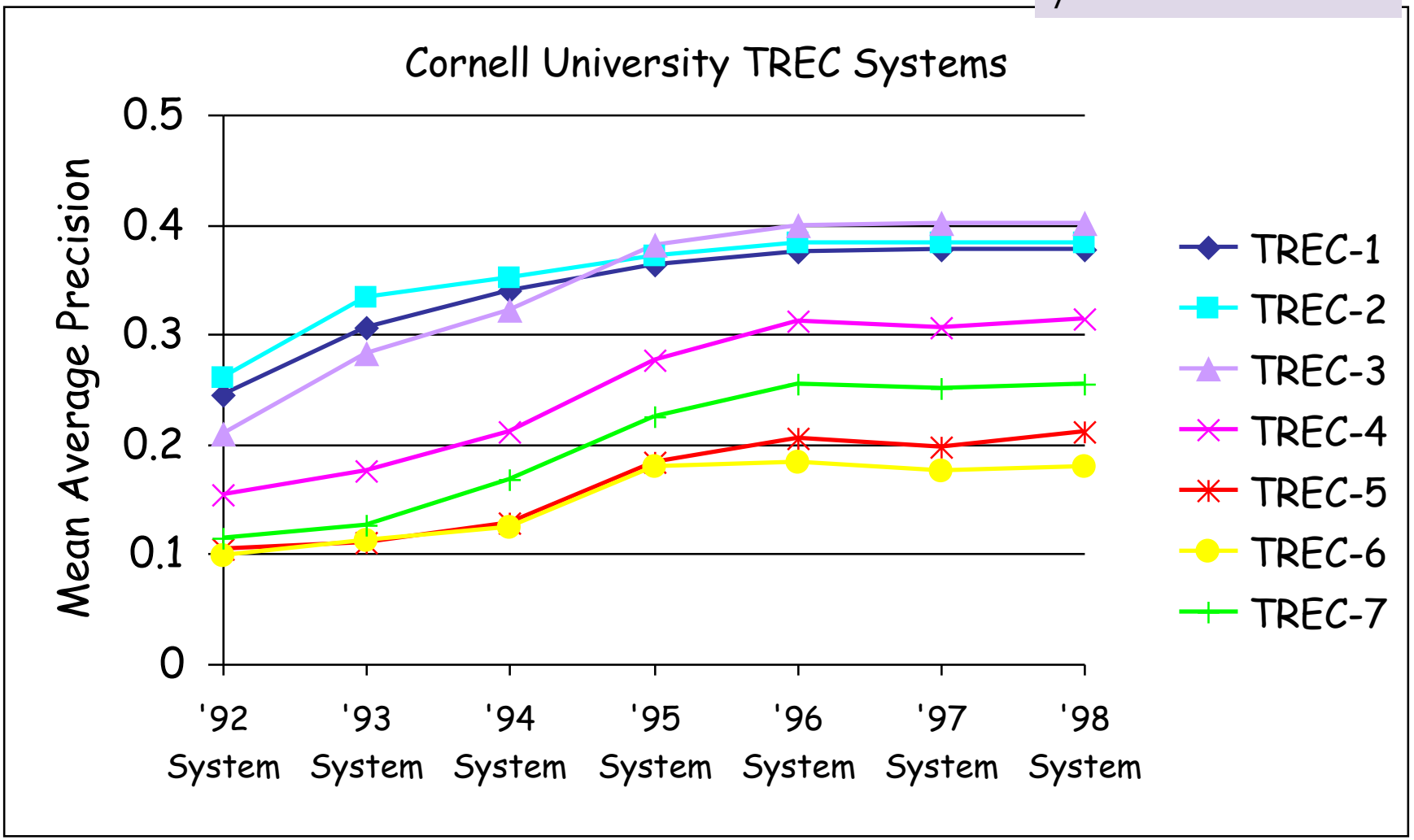
To learn how it can be improved

Scientific Understanding of the effectiveness

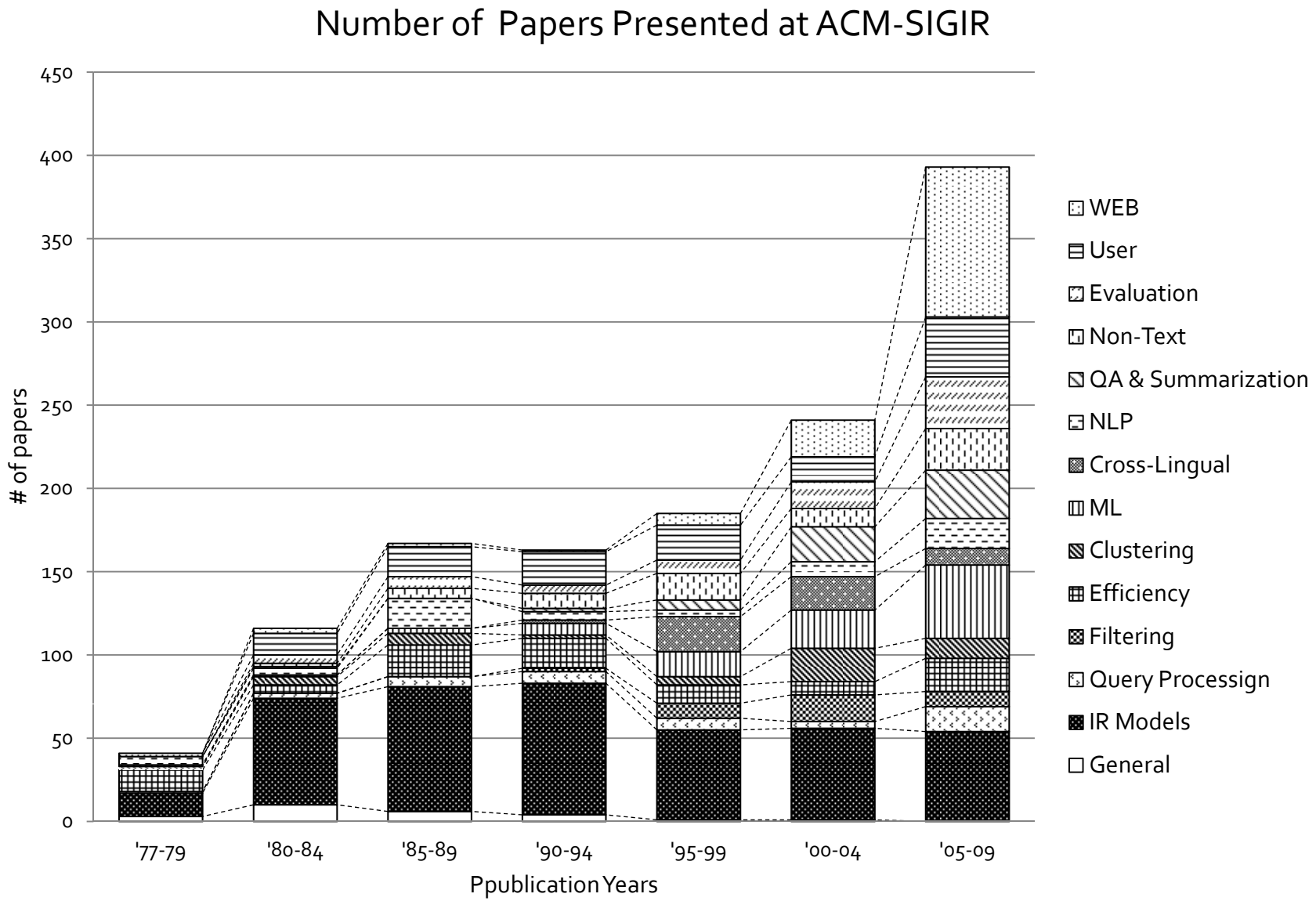
NTCIR

Improvement of Effectiveness by Evaluation Workshops

1.5 – 2 times in 3 years



Research Trends



NTCIR

Some Thoughts on Future

- Requirements for Evaluating Individual Applications etc.
- Interactive and Exploratory Information Access
  - Users' Intention, Diversity
  - Collaborative Search
  - Expertise, Knowledge, Literacy, etc.
- Ex. Key Topics discussed at SWIRL 2012
  - Conversasional Retrieval
  - MobileIR
  - Zero Query (or Less)
  - SmarteIR
  - Query by walking around
  - Structure, Entity, Relation

About NTCIR-10

What's new?

NTCIR-10: Objectives

- Solid foundation
  - Even stronger structure
- Task diversity
  - Covers a wide context in Information Access
  - Studies rich media types
- Community-led task organisation
  - Sustainability of research
- Promotion of research resources
  - Show case in the NTCIR-10 Meeting

NTCIR-10: Structure

- General Co-Chairs
  - Noriko Kando (NII)
  - Tsuneaki Kato (Tokyo University)
  - Douglas W. Oard (University of Maryland)
  - Mark Sanderson (RMIT)
- Program Co-Chairs
  - Hideo Joho (Tsukuba University)
  - Tetsuya Sakai (MSRA)
- Task Organisers
  - 48 researchers worldwide
  - Participants (You!)
- EVIA 2013 Co-Chairs
  - Ruihua Song (MSRA)
  - William Webber (University of Maryland)

http://research.nii.ac.jp/ntcir/ntcir-10/organizers.html

NTCIR-10: Development so far

Oct 2011	New structure formed for NTCIR-10
Nov 2011	Call for task proposal announced and 10 proposals were submitted
Dec 2011	NTCIR-9 Workshop Meeting
Jan 2012	7 proposals were accepted by the program committee and co-chairs
Feb 2012	Calls for task participation prepared
Mar 2012	NTCIR-9 Kick-Off Event

NTCIR-10 Program Committee

- Charles Clarke (University of Waterloo, Canada)
- Kalervo Järvelin (University of Tampere, Finland)
- Hideo Joho (Co-chair, University of Tsukuba, Japan)
- Gareth Jones (Dublin City University, Ireland)
- Noriko Kando (NII, Japan)
- Tsuneaki Kato (The University of Tokyo, Japan)
- Douglas W. Oard (University of Maryland)
- Tetsuya Sakai (Co-chair, Microsoft Research Asia, PRC)
- Mark Sanderson (RMIT, Austraria)
- Ian Soboroff (NIST, US)

NTCIR-10 Evaluation Tasks

Calls for task participation

Tasks accepted for NTCIR-10

CORE TASKS

- [Intent-2] Search intent and diversification
- [1Click-2] One-Click Access
- [RITE-2] Recognizing Inference in Text
- [SpokenDoc-2] IR for Spoken Documents
- [PatentMT-2] Cross-lingual access to Patent Docs
- [CrossLink-2] Cross-lingual Link Discovery

PILOT TASKS

- [Math] Access to mathematical contents

CrossLink-2

Crosslink Introduction

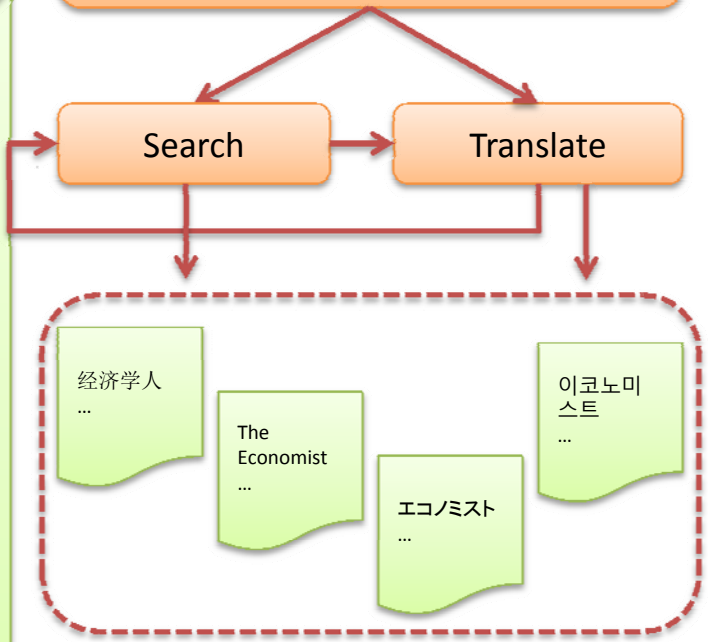
- Cross-lingual link discovery (Crosslink or CLLD) is concerned with automatically finding potential links between documents in different languages.
- CLLD algorithms actively recommend a set of meaningful anchors in the context of a source document and establish links to documents in an alternative language.
- CLLD is helpful for complimentary knowledge discovery in different language domains.

Cross-lingual Link Discovery

Article: Australia

...
Ranked third in the Index of Economic Freedom (2010),[178] Australia is the world's thirteenth largest economy and has the ninth highest per capita GDP; higher than that of the United Kingdom, Germany, France, Canada, Japan, and the United States. The country was ranked second in the United Nations 2010 Human Development Index and first in Legatum's 2008 Prosperity Index.[179] All of Australia's major cities fare well in global comparative livability surveys;[180] Melbourne reached first place on **The Economist**'s 2011 World's Most Livable Cities list, followed by Sydney, Perth, and Adelaide in sixth, eighth, and ninth place respectively.[181] Total government debt in Australia is about $190 billion.[182] Australia has among the highest house prices and some of the highest household debt levels in the world.
...

No link was created for this term, for finding articles in languages we prefer traditionally we do:



How to automatically create cross-lingual links for a document if no links existing yet?

- ❑ Links in other languages?
- ❑ New articles?
- ❑ Missing links?
- ❑ Not what we are looking for?
- ❑ What about other relevant links?

Cross-lingual Link Discovery

- ❖ Cross-lingual Links
- ❖ New Links
- ❖ Better Links
- ❖ More options

✓

- All about ***multi-lingual knowledge discovery*** in knowledge bases (e.g. Wikipedia)
- All about ***easy*** and ***efficient*** information access

The Goal of Crosslink Task

- It is aimed to create a reusable resource for evaluating automated cross language link discovery approaches. The results of this research will be used in building and refining systems for automated link discovery.

Crosslink task at NTCIR-9

- Cross-lingual link discovery (Crosslink) as a pilot task of NTCIR-9 has been successfully held in 2011
- Below is a list of participating teams with submissions:

GROUP	ORGANISATION
DUIIS	Daegu University
HITS	Heidelberg Institute for Theoretical Studies
IISR	Yuan Ze University
ISTIC	Institute of Scientific and Technical Information of China
KMI	The Open University
kslab_nut	Nagaoka University of Technology
KSLP	Kyungsung University
nthuisa	Academia Sinica
QUT	Queensland University of Technology
UKP	TU Darmstadt
WUST	Wuhan University of Science and Technology

Crosslink task at NTCIR-9 (Cont.)

- There were in total 57 runs from 11 teams were received.

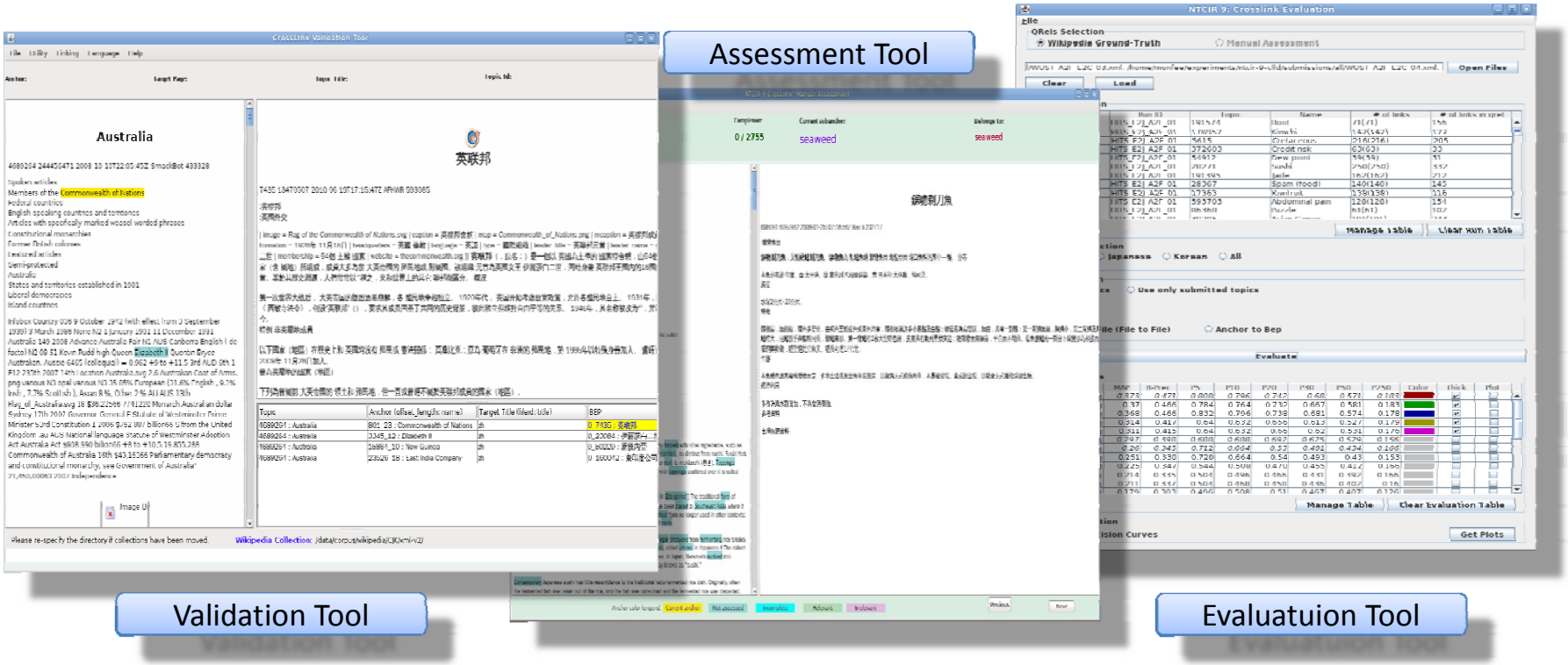
Group	En-2-Zh	En-2-Ja	En-2-Ko
DUIIS	0	0	2
HITS	3	3	3
IISR	0	0	5
ISTIC	1	0	0
KMI	4	0	0
kslab_nut	0	1	0
KSLP	0	0	5
nthuisa	3	0	0
QUT	5	2	1
UKP	5	5	5
WUST	4	0	0
Sub-total	25	11	21
Total	57		

Crosslink Task at NTCIR-10

- CLLD is a core task of NTCIR-10
- New Subtasks
  - Chinese to English CLLD (C2E)
  - Japanese to English CLLD (J2E)
  - Korean to English CLLD (K2E)
- New Document Collections
  - ECJK Wikipedia collections
- New Topics
  - 25 topics for each language

Evaluation Framework

- Assessment Types: Automatic (Wikipedia Ground Truth), Manual (human in the loop)
- Test data, training data, gold standard, ECJK Wikipedia collections, validation tool, assessment tool and evaluation tool (with system evaluation metrics).
- Snapshots of the validation, assessment and evaluation tools:



System Evaluation Metrics

- $LMAP = (\sum_{t=1}^{n} \frac{\sum_{k=1}^{m} p_{kt}}{m})/n$

- $R - Prec = \sum_{t=1}^{n} P_t @ R /n$

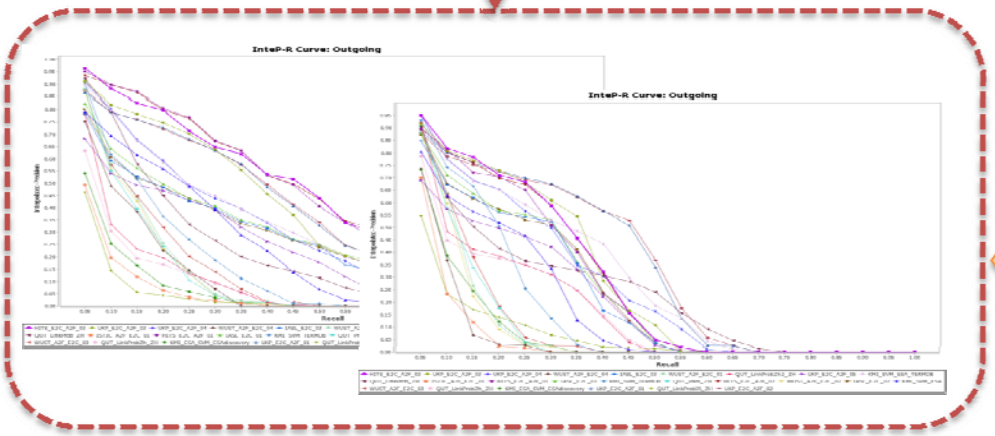
where *n* is the number of topics; *m* is the number of identified items (links or anchors); P_{kt} is the precision at top *k* items (links or anchors) for topic *t*; $P_t@R$ (= *number of correct items (links or anchors) / number of items (links or anchors) in qrel*) is the precision calculated using number of links / anchors in *qrel* as denominator for topic *t*.

- *Precision-at-N* is computed using the average precision for all topics (source articles) at a pre-defined position *N* in the results list. Values of *N* were chosen as: 5, 10, 20, 30, 50, and 250.

Expected Outcomes

- More good submissions
- More original and innovative approaches can be seen in identifying meaningful anchors and suggesting high quality cross-lingual links
- The research results can really help the cross-lingual knowledge discovery in knowledge bases

Evaluation Framework



- The evaluation framework will be further refined
- The evaluation methods will be further perfected to distinguish the good and the bad CLLD algorithms for the new subtasks

Expected Participants

- Previous active participants of NTCIR-9 Crosslink task
- The registered participants of NTCIR-9 Crosslink task.
- Participants of CLIR or IR4QA task in previous NTCIR workshops.
- Participants of previous INEX Link-the-Wiki track
- Other researchers in the CLIR field

Contacts

- Organisers:

Shlomo Geva	Queensland University of Technology, Australia
Andrew Trotman	Universityof Otago, New Zealand
Yue Xu	Queensland University of Technology, Australia
Eric Tang	Queensland University of Technology, Australia
In-Su Kang	Kyungsung University, South Korea
Fuminori Kimura	Ritsumeikan University, Japan
Haitao Mi	Chinese Academy of Sciences, China

- Mailing list: crosslink@lists.otago.ac.nz

INTENT-2





INTENT-2@NTCIR-10 Kickoff (English version)

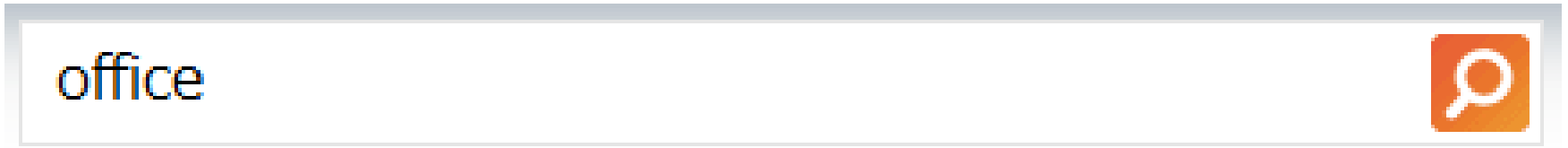
TASK ORGANISERS: Tetsuya Sakai, Ruihua Song, Zhicheng Dou (Microsoft Research Asia) Min Zhang, Yiqun Liu (Tsinghua University) Takehiro Yamamoto, Makoto Kato (Kyoto University) Mayu Iwata (Osaka University)

http://research.microsoft.com/en-us/people/tesakai/intent2.aspx



March 8, 2012@NII, Tokyo

What do you expect from a search engine?



Ambiguous query

office

Microsoft software!

Microsoft Office

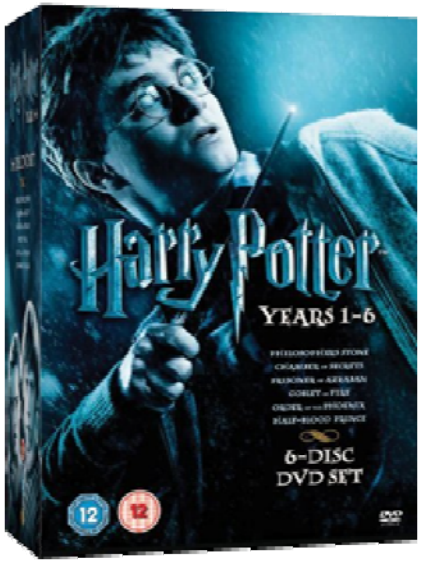
Workplace!

What do you expect from a search engine?

Underspecified query

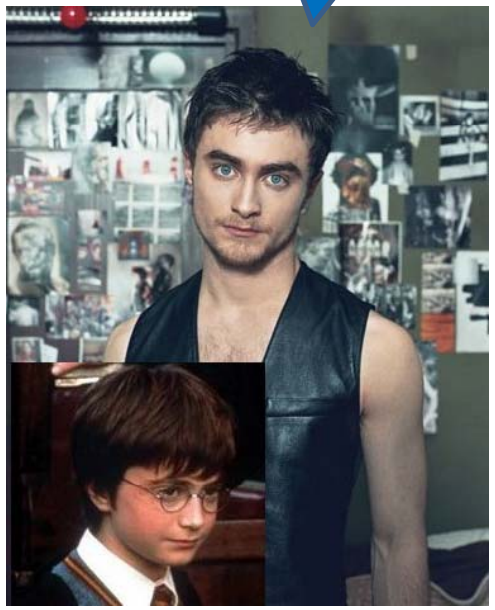
harry potter

Harry Potter films!



Harry Potter books!

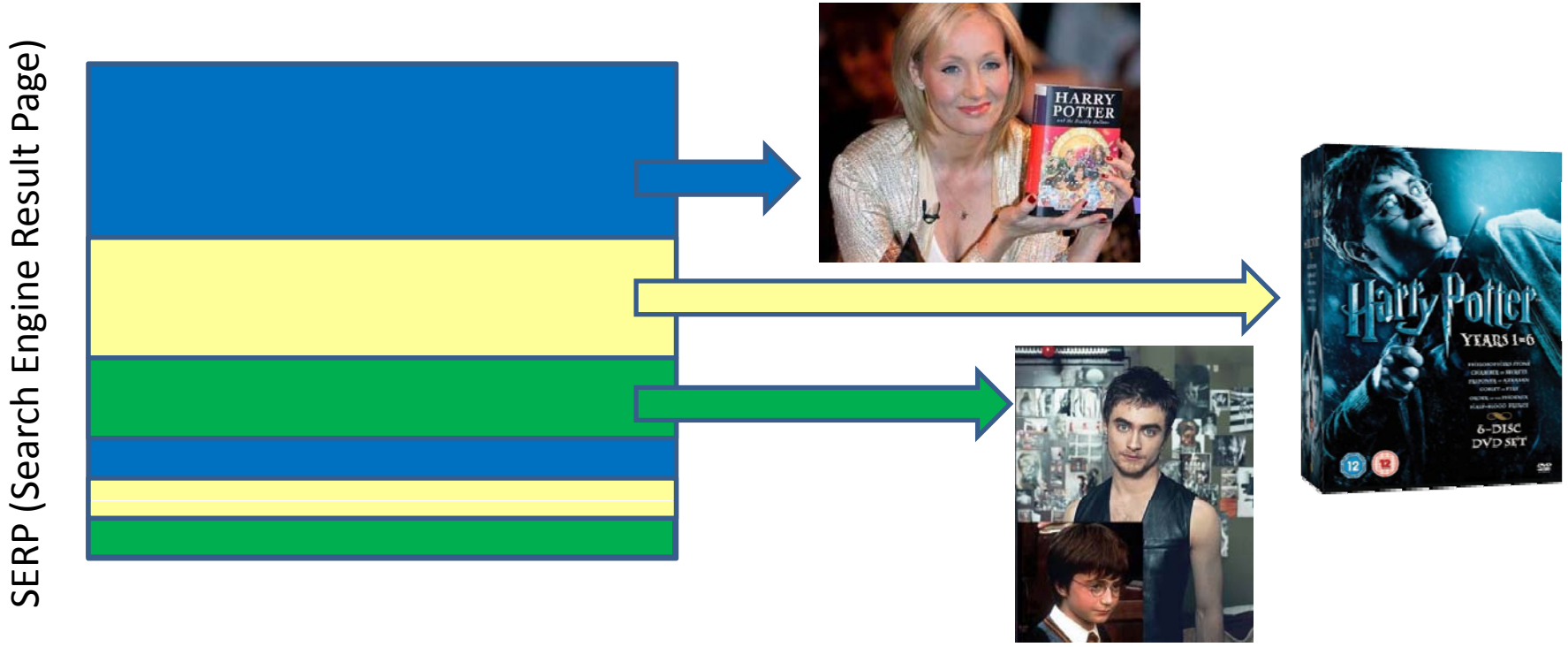
Harry Potter the character!



Search Result Diversification

Given an ambiguous/underspecified query,

produce a single Search Engine Result Page that satisfies different user intents!



INTENT-2 Subtasks

- Subtopic Mining (Chinese, Japanese, English)

INPUT: query (e.g. “harry potter”)

OUTPUT: ranked list of subtopic strings

(e.g. “harry potter book, harry potter film, harry potter the character…”)

- Document Ranking (Chinese, Japanese)

INPUT: query (e.g. “harry potter”)

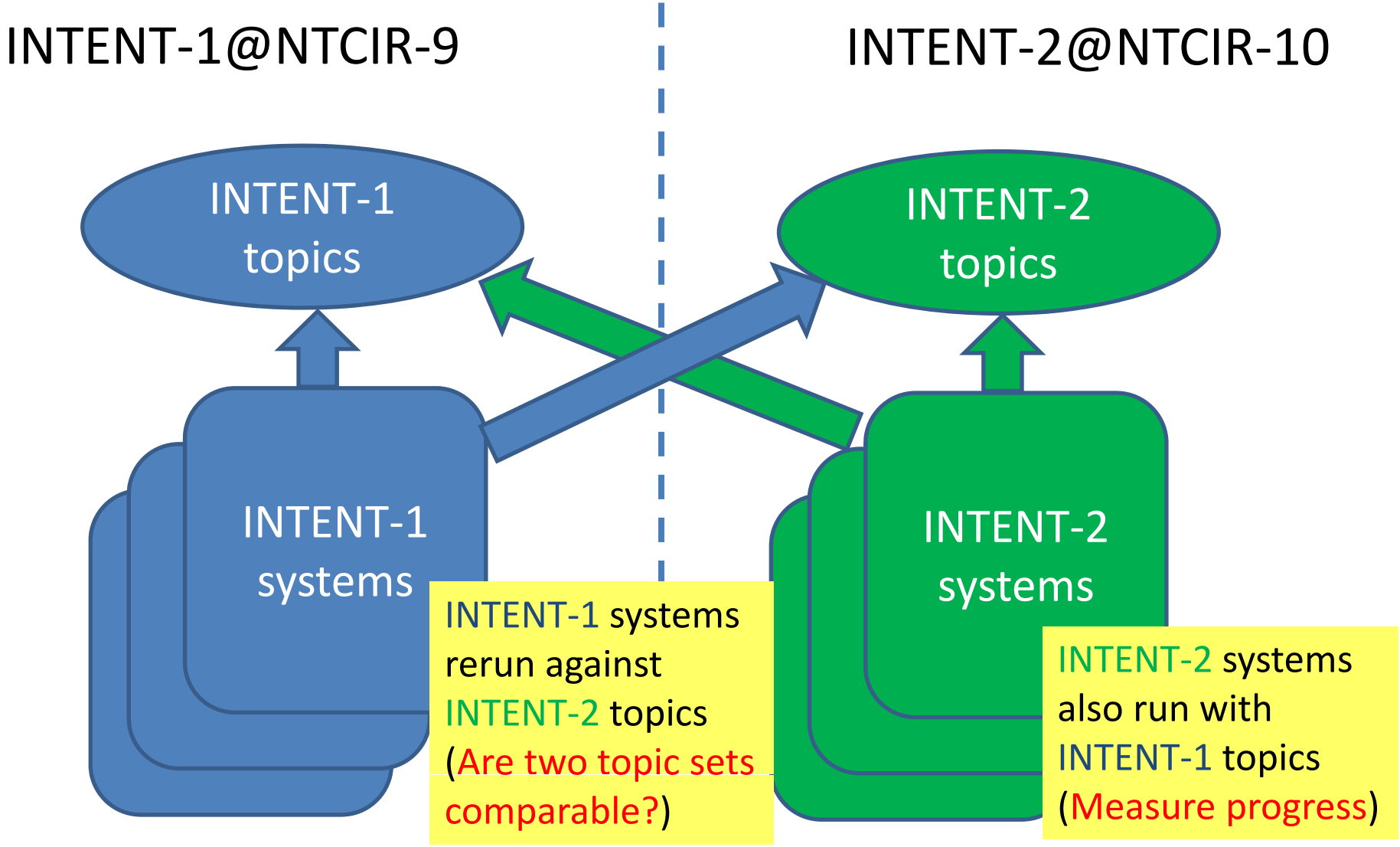
OUTPUT: diversified ranked list of web pages

Main challenge: Balancing relevance and diversity

Departures from INTENT-1 (1)

- English Subtopic Mining (new!) will share the topic set with TREC 2012 diversity task
- Inclusion of queries that do not require diversification (one item search) – systems have to selectively diversify
- Organisers will provide query suggestions and baseline search results to participants – researchers who do not have a search engine can easily participate

Departures from INTENT-1 (2)



IMPORTANT DATES

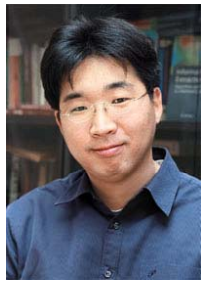
PARTICIPANTS ORGANISERS

May 2012	Topics and baselines released
July 2012	All submissions due
Aug-Dec 2012	Identify intents from subtopics -> per-intent relevance assessments
Jan 2013	Evaluation results released
March 2013?	Draft papers due
May 2013?	Camera ready papers due
June 18-21	NTCIR-10

INTENT-1 had 17 research teams from 7 countries/regions!
But we'd like to see more!

1Click-2

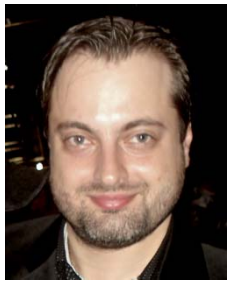



1CLICK-2@NTCIR-10 Kickoff (English version)

TASK ORGANISERS: Makoto Kato, Takehiro Yamamoto (Kyoto U),
Tetsuya Sakai, Young-In Song, Zhicheng Dou (MSRA),
Virgil Pavlu, Matthew Ekstrand-Abueg, Shahzad Rajput (Northeastern U)
Mayu Iwata (Osaka U)

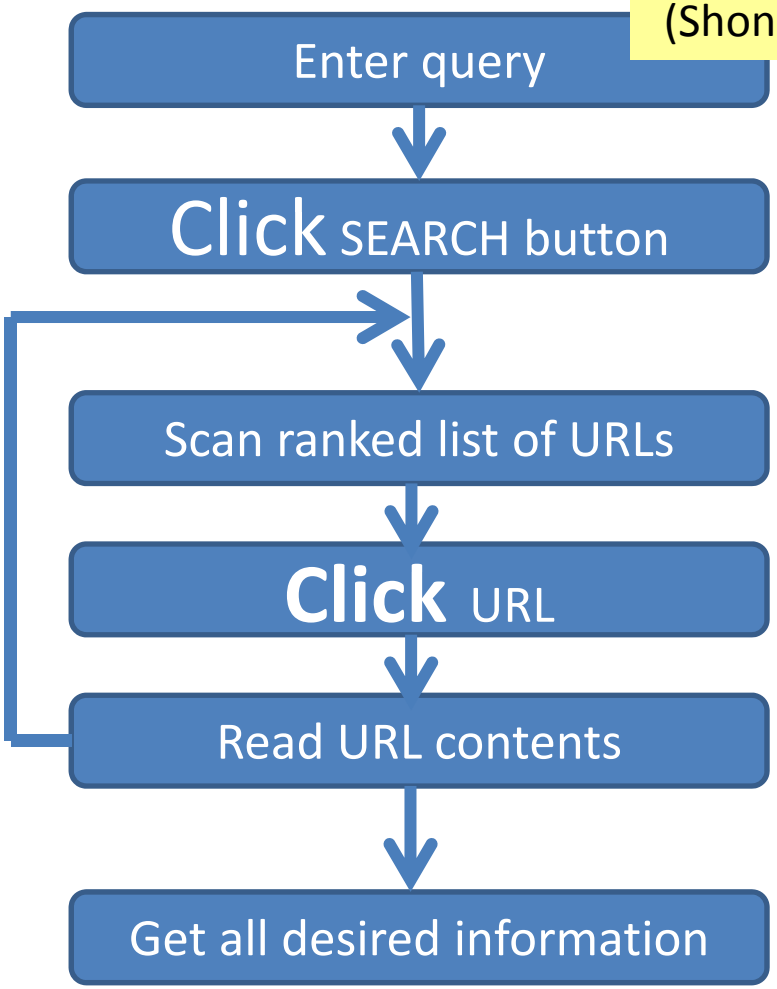
http://research.microsoft.com/en-us/people/tesakai/1click2.aspx




March 8, 2012@NII, Tokyo

Traditional Search
= More-than-One Click Access

湘南厚木病院
(Shonan Atsugi Hospital)



One Click Access

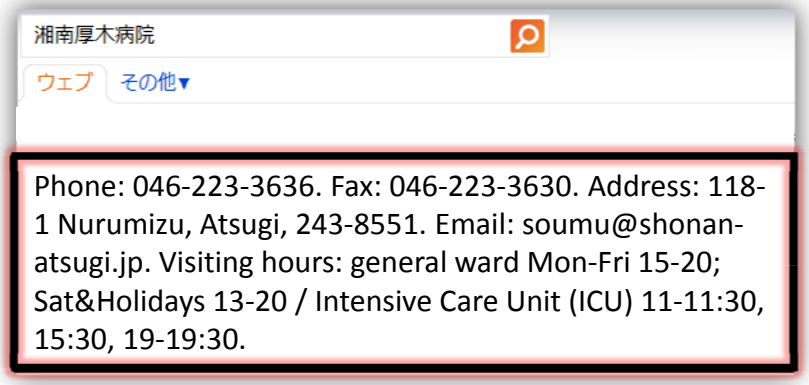
湘南厚木病院
(Shonan Atsugi Hospital)

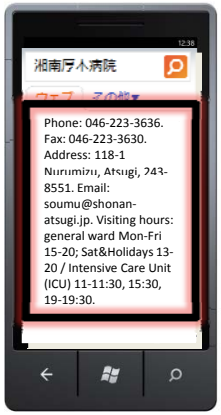
Enter query

Click SEARCH button

Get all desired information

The system outputs ***X-string***





Particularly important for

mobile search

Go beyond the "ten-blue-link" paradigm in Web search

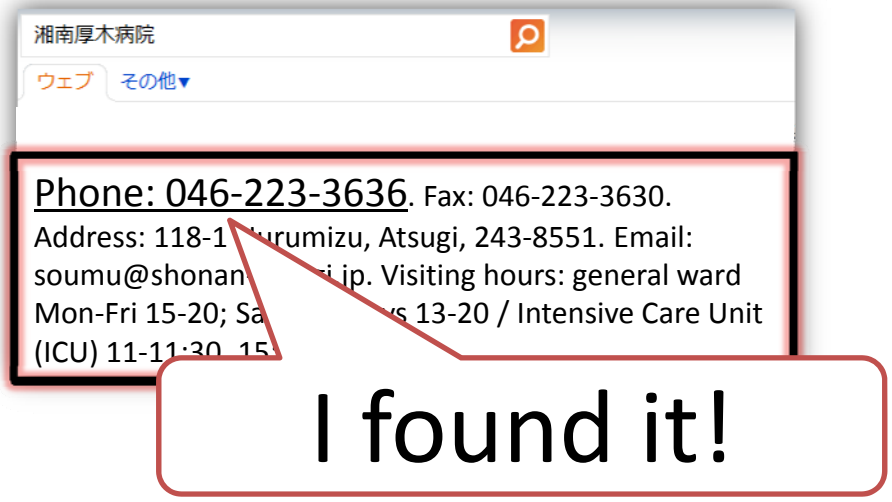
Benefit of 1CLICK Access

Immediate & Direct Access

When you have to make a call to the Shonan Atsugi Hospital as soon as possible

Traditional search engine

1CLICK access system



Queries

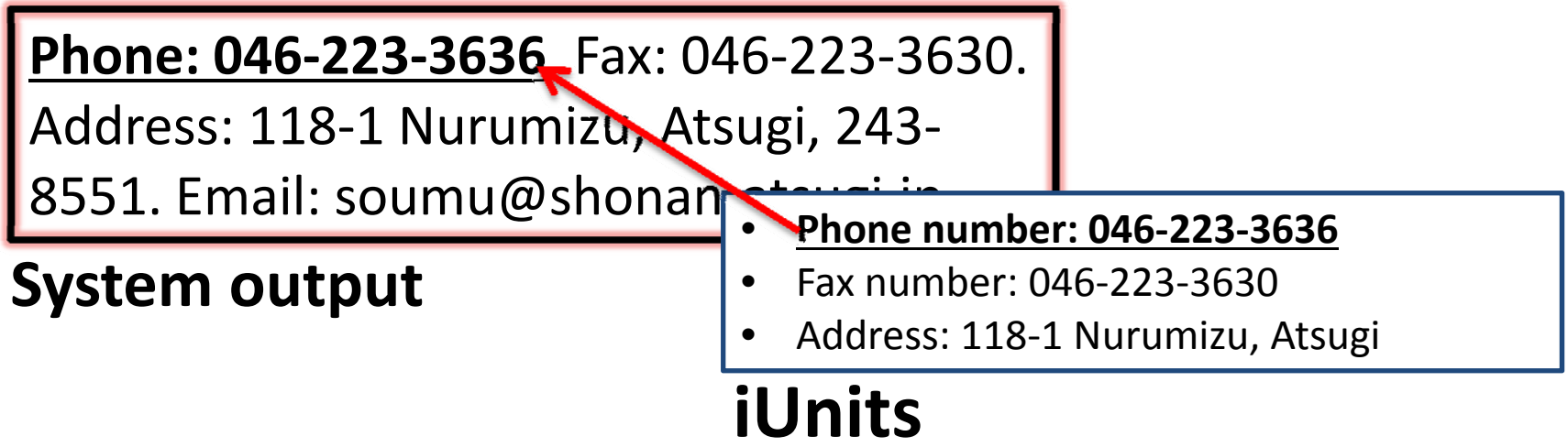
- **8 query types based on [Li et al., SIGIR09]**
  (for which the user's information need can be satisfied by the search results page)
  - **ARTIST**, **ATHELETE**, **POLITICIAN**, **ACTOR**
  - **FACILITY**, **LOCAL**
  - **DEFINITION** (of a term)
  - **QA** (Natural language questions)

- **Queries were selected out of a large query log of a commercial search engine**

J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and PC internet search. ACM SIGIR 2009

Evaluation with iUnits

- Manual/automatic matching between the **system output** and **iUnits** (atomic pieces of information)



System output

iUnits

- Systems are evaluated from viewpoints of **INFORMATION** retrieval rather than document retrieval

1CLICK-2 Subtasks

Participants must submit at least one run to

- **Main Tasks**
  - **Mandatory Runs**: generate X-string based on *baseline web search results*
  - **Oracle Runs** (optional): generate X-string based on Webpage URLs supporting the gold standard iUnits
  - **Open Runs** (optional): generate X-string by searching the live Web on their own

OR

- **Query classification subtasks**
  - Given a query, return its query type (ARTIST, QA, etc.)

Departures from 1CLICK-1

- The language scope was expanded:
  Japanese & **English**
- More query types (4 → **8 types**)
- A new info. unit **iUnit** will be used for evaluation
- **Automatic**/manual matching between
  the X-string and iUnits
- **Diversified tasks**
  - Fair comparison by a common dataset
    (**mandatory runs**)
  - You can join us
    even only with **query classification tasks**!

IMPORTANT DATES

PARTICIPANTS | ORGANISERS

Apr 2012	Sample queries and nuggets released
Aug 2012	Test queries released
Oct 2012	Run submissions due
Nov 2012-Jan 2013	Nugget match evaluation
Feb 2013	Evaluation results released
Mar 2013	Draft papers due
Jun 2013	NTCIR-10

Join us if you are interested in **IR**, **IE**, **MOBILE SEARCH**, **SUMMARISATION**, or **the next search paradigm**

PatentMT-2

The Patent Machine Translation Task (PatentMT)

Isao Goto (NICT)
Bin Lu (City Univ. of Hong Kong / Hong Kong Institute of Education)
Ka Po Chow (Hong Kong Institute of Education)
Eiichiro Sumita (NICT)
Benjamin K. Tsou (Hong Kong Institute of Education / City Univ. of Hong Kong)

Motivation

- There is a significant **practical need** for patent translation.
  - to understand patent information written in foreign languages.
  - to apply for patents in foreign countries.
- Patents constitute one of the **challenging domains**.
  - Patent sentences can be quite **long** and contain **complex structures**.
  - Translation between **languages with largely different word order** is difficult for **long** sentences.

Goals of PatentMT

- To develop **challenging** and **significant practical** research into patent machine translation.
- To **investigate** the **performance** of state-of-the-art machine translation systems in terms of patent translations involving Japanese, English, and Chinese.
- To **compare** the effects of **different methods** of patent translation by applying them to the same test data.
- To **create** publicly-available **parallel corpora of patent documents** and human evaluations of MT results for patent information processing research.
- To **drive machine translation research**, which is an important technology for cross-lingual access of information written in unknown languages.
- The ultimate goal is **fostering scientific cooperation**.

Findings of PatentMT at NTCIR-9

- **SMT** was the **best** system for **Chinese to English** and **English to Japanese** patent translation.
  - This is the **first time** for **SMT** to be **demonstrated equal or better** quality than that of the top-level RBMT for **English to Japanese** patent translation.
  - The **pre-ordering** method of NTT-UT for SMT is very effective for English to Japanese patent translation.
- **80%** of patent sentences could be understood in the best system for **Chinese to English** patent translation.
- **RBMT** was the best system for **Japanese to English** patent translation.

Remaining Issues of NTCIR-9

- Practical evaluation
  - The quality of translated sentences was evaluated at NTCIR-9.
  - More practical evaluations are also expected.

Outline of the Plans for NTCIR-10

- Three subtasks:

Subtasks	Training data
Chinese to English	**1 million** sentence pairs
Japanese to English	Approximately **3.2 million** sentence pairs
English to Japanese	

(Subtasks and training data are the same as at NTCIR-9)

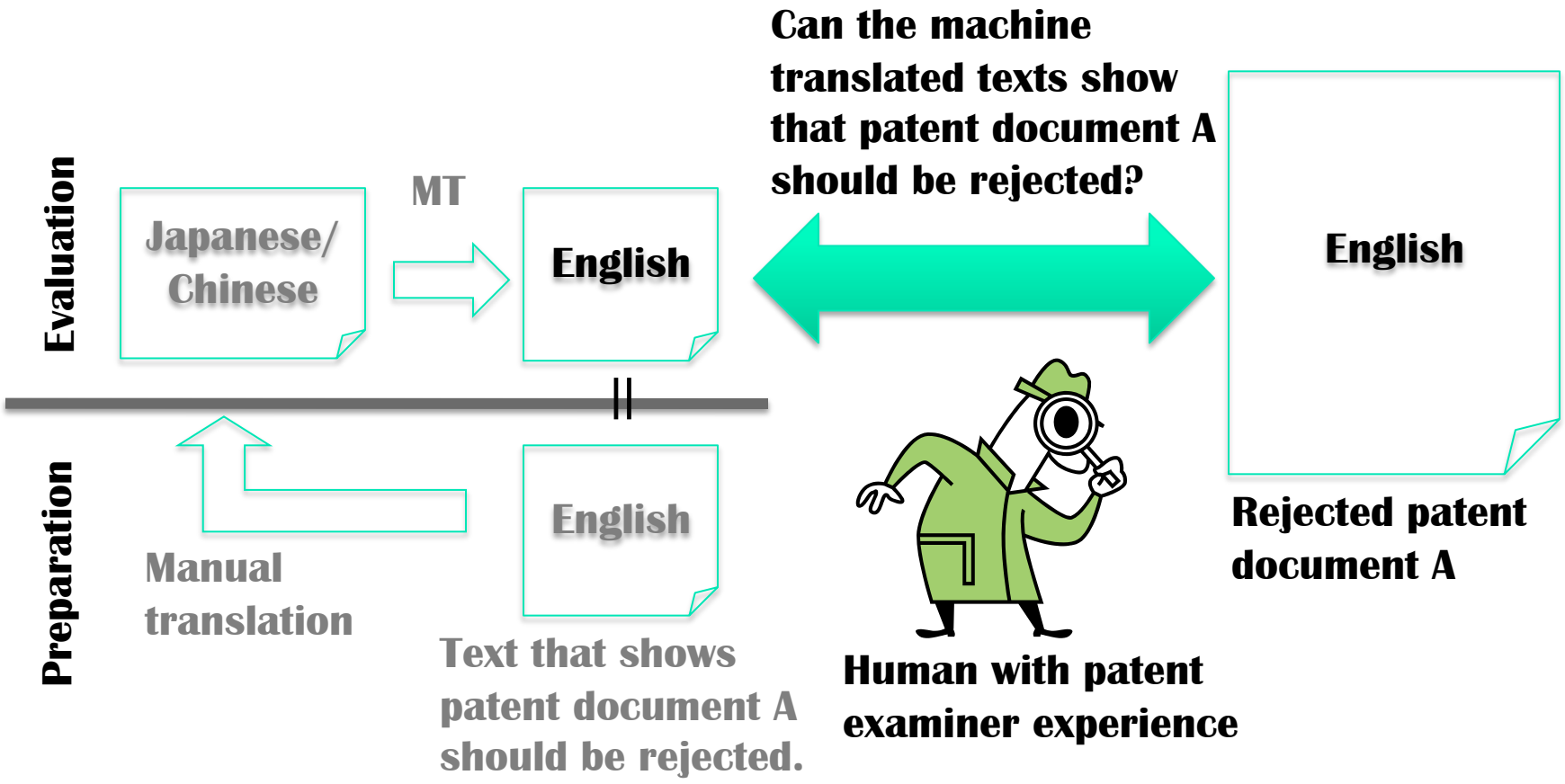
- Participants select subtasks in which they wish to participate.
- Large scale **parallel corpora** and **new test sets** will be **provided**.
- **Practical evaluation** will be added (under consideration).
- **Human evaluation** will be carried out.

Differences from NTCIR-9

<table>
<tr><td>Practical Evaluation (under consideration)</td><td colspan="2">New: To explore practical MT performance in appropriate fields for patent machine translation.</td></tr>
<tr><td rowspan="3">Intrinsic Evaluation</td><td colspan="2">Similar to the NTCIR-9 evaluation.
Quality of translated sentences will be evaluated.

Additions:</td></tr>
<tr><td>Chronological evaluation</td><td>Comparison between NTCIR-10 and NTCIR-9 to measure progress.</td></tr>
<tr><td>Multilingual evaluation</td><td>Comparison of CE and JE translations using the same English reference will be added.</td></tr>
</table>

Possible Approach to Practical Evaluation



(The feasibility of this is under investigation.
We are working hard to make necessary arrangements.)

Chronological Evaluation

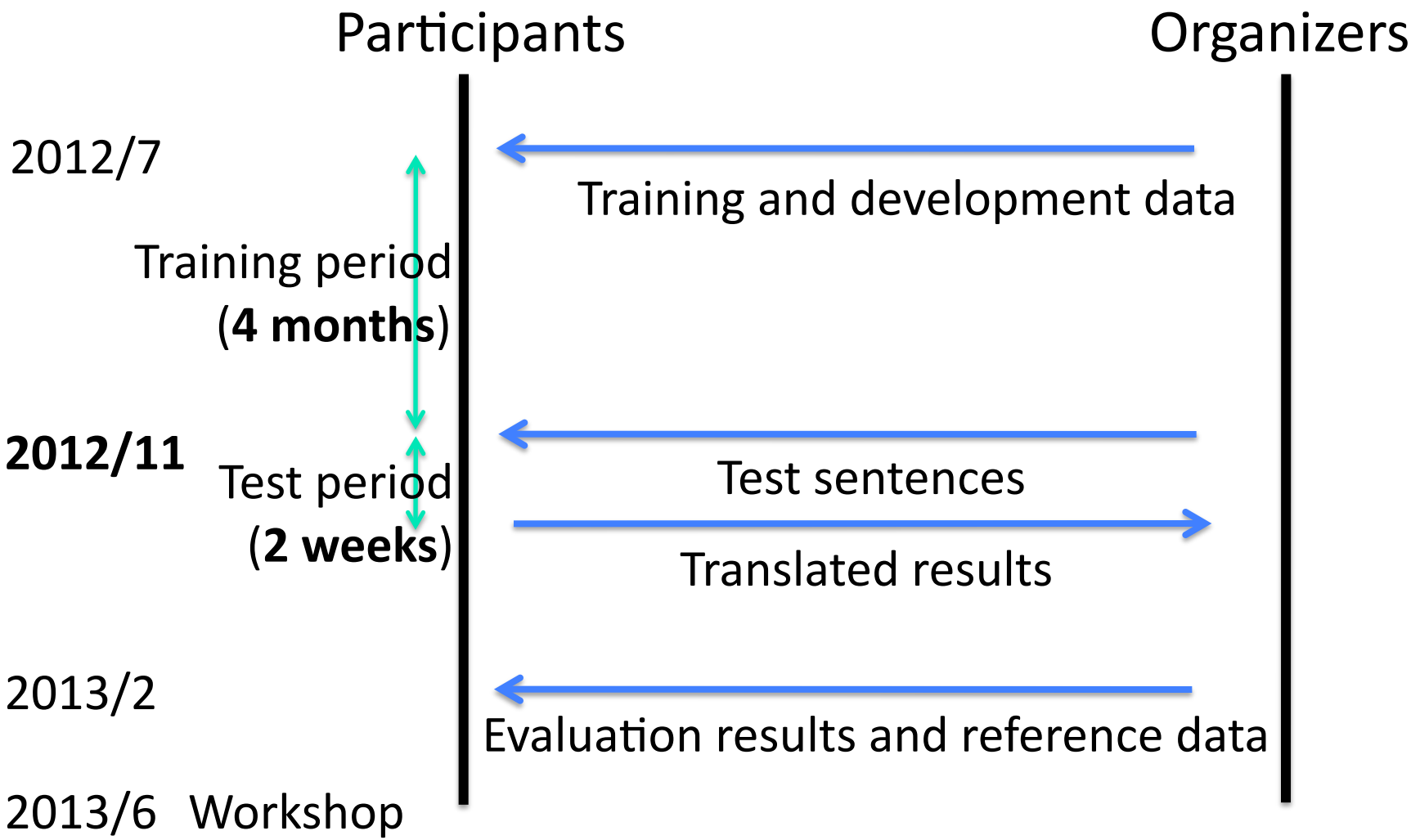
- In addition to the new NTCIR-10 test sets, the NTCIR-9 test sets will be also translated.
- Translations of the NTCIR-9 test sets at NTCIR-10 will be compared to the NTCIR-9 submissions.
- This allows **measurement** of **the progress from NTCIR-9**.

Multilingual Evaluation

- We will produce a **C-J-E multilingual** test set.

- A C-J-E multilingual test set enables **comparison** of the **CE** translations and the **JE** translations using the **same English reference** data.

The Flow and Tentative Schedule of the Task



Why is it so exciting to participate in?

- **Patents** are one of the **challenging domains** for MT.
  - Patent sentences could be quite **long** and contain **complex structures.**
  - Translation between **languages with largely different word order** is difficult for **long** sentences.
- Participants will receive **evaluation results** for their MT quality.
- Participants can use **large-scale patent parallel** and **monolingual corpora**.
- Participants can choose subtasks from three language directions, including **the language direction of Chinese to English**.

- We look forward to many groups participating in PatentMT at NTCIR-10!

SpokenDoc-2

NTCIR-10
SpokenDoc-2

IR (Information Retrieval) for Spoken Documents

Kiyoaki Aikawa[1], Tomoyosi Akiba[2] , Seiichi Nakagawa[2],
Hiroaki Nanjo[3], **Hiromitsu Nishizaki[4]**, Yoichi Yamashita[5]

[1]Tokyo University of Technology, Japan
[2]Toyohashi University of Technology, Japan
[3]Ryukoku University, Japan
[4]University of Yamanashi, Japan
[5]Ritsumeikan University, Japan

What is “SpokenDoc-2”?

- Second round of the IR for spoken documents
- Finding the information related to given a query from too much speech data

Speech data ≅ Transcribed text by automatic speech recognition

“information retrieval”

Information retrieval (word-unit transcription)
In for ma tion re triev al (sub-word-unit)

Problem Including a lot of errors!!

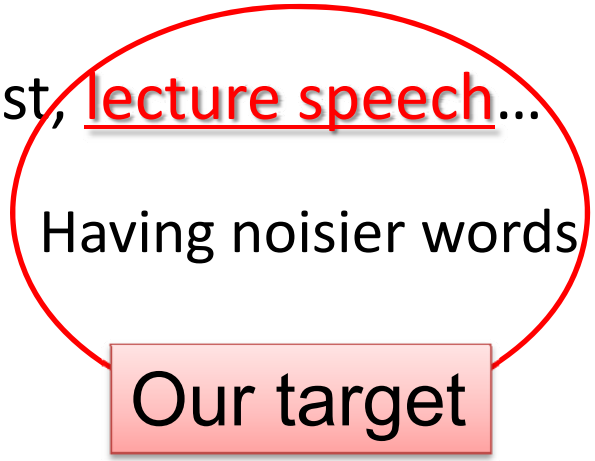
Participants of SpokenDoc-2 will challenge

Information retrieval from very noisy text data

Techniques for SpokenDoc may be used for OCR or Machine Translated text retrieval

Target Speech Data

- Type of speech data
  - Broadcast news speech, podcast, lecture speech…



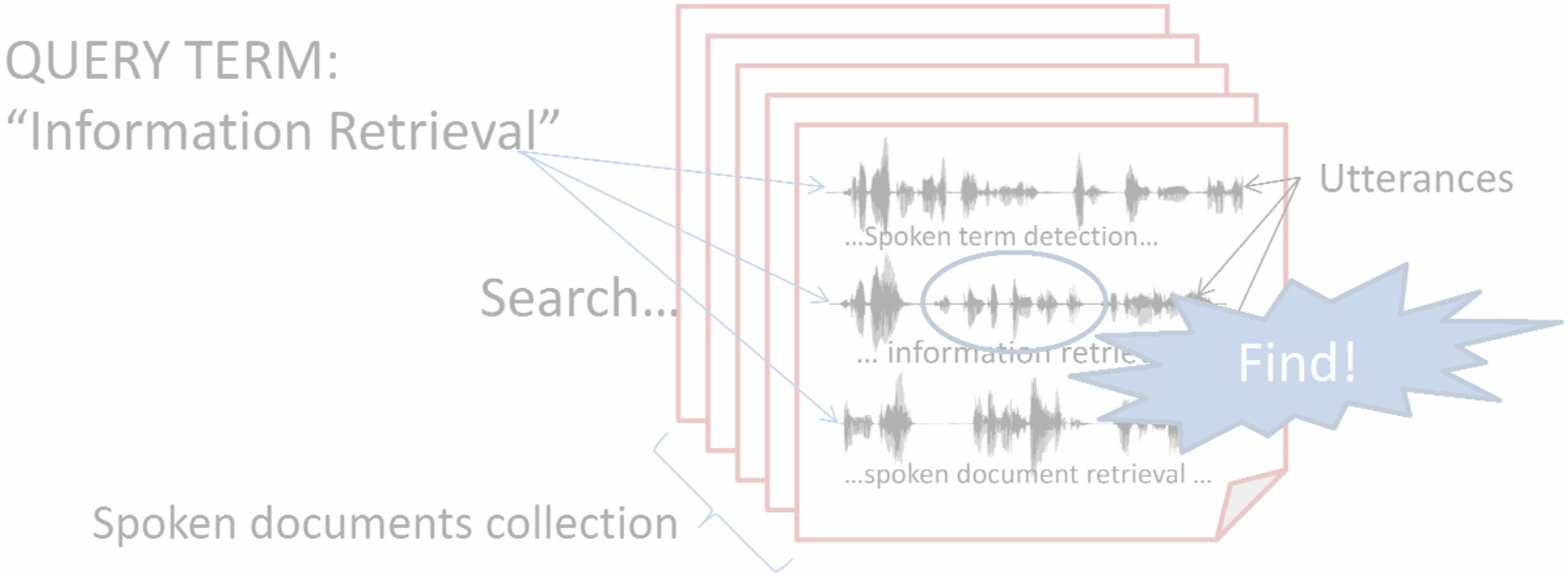
- Databases
  - CSJ (Corpus of Spontaneous Japanese)
    - 2,702 lecture speeches, 628 hours
  - New target! Real academic meeting lectures collection
    - Over 70 speeches from the spoken document processing workshops

SpokenDoc-2 Task Overview

- Sub-task1: Spoken Term Detection
- Sub-task2: Spoken Document Retrieval

Spoken Term Detection

1. Finding utterances including a query term



2. A query term is included or not included in a spoken document?

Spoken Document Retrieval

- Ad-hoc Information Retrieval from lecture speeches
- Finding the passages including the relevant information related to a given query topic
- Query
  - Text query
  - Spoken Query (optional)

Merits for Task Participants

- The organizers will provide the rich transcriptions by an automatic speech recognition system
  - Participants can focus on developing IR technique for noisy data, although you have not any techniques for speech recognition
- Participants will get the real lecture speeches:
    - Academic meeting lectures
  - All lectures have reference transcriptions
  - ~~Some lectures have slide data and digitalized papers~~

All data is Japanese

We are welcome to join SpokenDoc-2

- Schedule
  - Mar. 2012: release of the task description
  - June 2012: release of the reference automatic transcription
  - Sept. 2012: dry-run
  - Nov. 2012: formal-run evaluation
  - Nov. 2012-Feb.2013: relevance judgment
  - Feb. 2013: release of the evaluation results
  - May 2013: camera ready submission due
  - June 2013: NTCIR-10 workshop meeting
- Contact
  - SpokenDoc-2 organizers:
    - E-mail: ntcadm-spokendoc2@cl.ics.tut.ac.jp
  - Web (coming soon!)
    - http://www.cl.ics.tut.ac.jp/~sdpwg/index.php?ntcir10
  - Twitter: @spokendoc2

RITE-2

RITE-2
Recognizing Inference in TExt@NTCIR10

Introduction to NTCIR-10 RITE-2 Task
(Recognizing Inference in TExt)

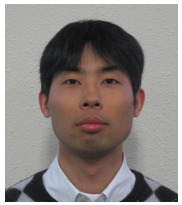
Yotaro Watanabe[1]

Junta Mizuno[1]

[1]Tohoku University

Yusuke Miyao[2]

[2]National Institute of Informatics



Tomohide Shibata[3]

[3]Kyoto University

Cheng-Wei Lee[4]

[4]Academia Sinica

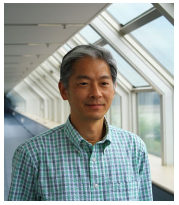
Chuan-Jie Lin[5]

[5]National Taiwan Ocean University

Shuming Shi[6]

[6]Microsoft Research Asia

Hiroshi Kanayama[7]



Koichi Takeda[7]

[7]IBM Research

Hideki Shima[8]

Teruko Mitamura[8]

[8]Carnegie Mellon University

NTCIR-10 Kick-off Event March 8th, 2012

RITE-2
Recognizing Inference in TExt@NTCIR10

Overview of RITE-2

- RITE is a benchmark task (not a competition!) for automatically detecting semantic relations between two sentences
  - Entailment, paraphrase and contradiction in text

- Given t_1(text), can a computer infer that
  t_2(hypothesis) is most likely true?
  - t_1: Yasunari Kawabata won the Nobel Prize in Literature for his novel "Snow Country"
  - t_2: Yasunari Kawabata is the writer of "Snow Country"

- Target languages
  - Japanese, Simplified Chinese, Traditional Chinese

RITE-2
Recognizing Inference in TExt@NTCIR10

RITE-2 Subtasks

- **Binary-Class (BC) Subtask**
  - Given a text pair (t_1, t_2), identify whether t_1 entails t_2 or not
- **Multi-Class (MC) Subtask**
Entail	t_1 entails t_2	t_1 has the same meaning as t_2 (paraphrase)
Not Entail	t_1 contradicts t_2	otherwise (independence)

- **Entrance Exam Subtasks**
  - **BC Subtask**
    - Same as the BC subtask in terms of input and output. All the data is created from actual college-level entrance exams.
  - **Search Subtask**
    - Given a hypothesis t_2 and resources (Wikipedia, textbook), identify whether t_2 is true or not

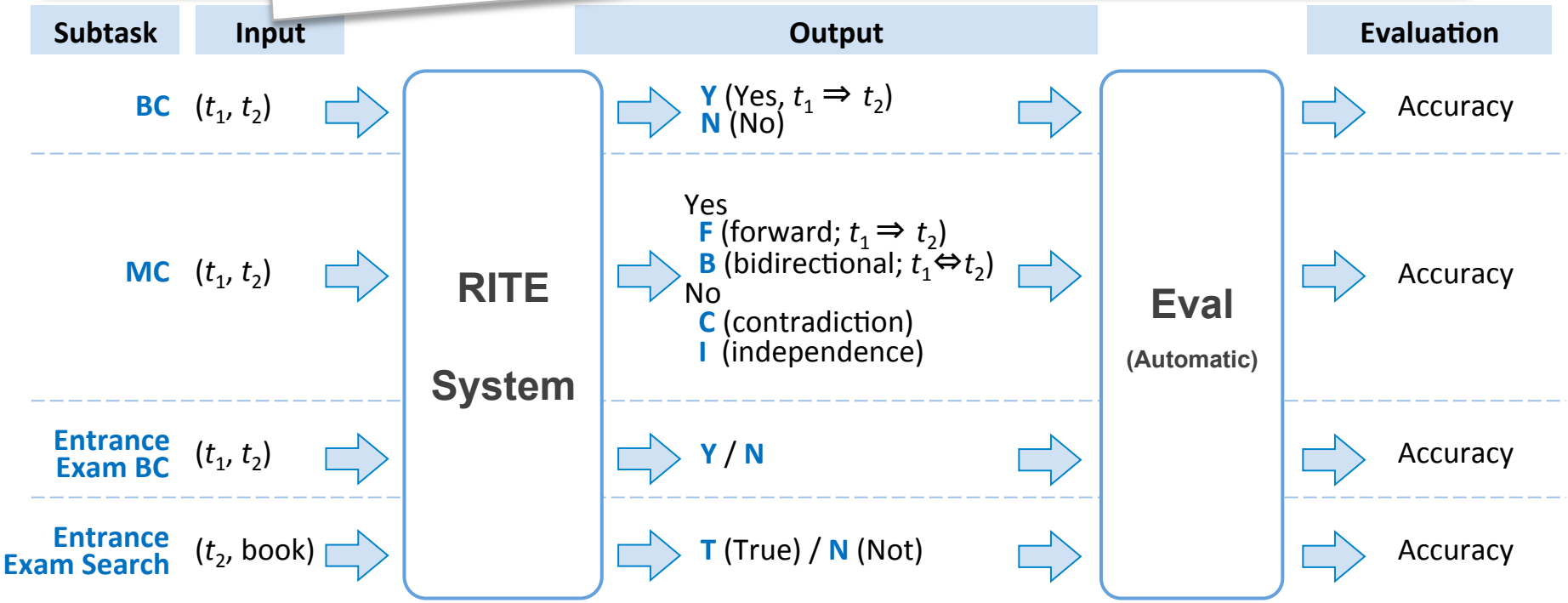
RITE-2
Recognizing Inference in TExt@NTCIR10

RITE-2 Overview

Does t_1 entail (infer) t_2?

t_1: Yasunari Kawabata won the Nobel Prize in Literature for his novel "Snow Country".

t_2: Yasunari Kawabata is the writer of "Snow Country".



RITE-2
Recognizing Inference in TExt@NTCIR10

Why you should participate

In addition to researchers in entailment and paraphrases, various research fields can benefit from RITE:

- **Core technologies**: Semantic processing, Lexical acquisition, Machine learning, etc.
- **Applications**: Information retrieval, Question answering, Document summarization.

We try hard to welcome wide variety of participations – from undergraduate students to industry researchers, from all over the world.

- **Resource pool**: will be available to help you build a prototype system quickly or participate in collaboration by sharing useful resource and tools with others.

Website: http://www.cl.ecei.tohoku.ac.jp/rite2

Math

NTCIR-10 kickoff

March 8, 2012

NTCIR-10 MATH PILOT TASK

The Goal of NTCIR-10 Math Task

- NTCIR Math Task aims at exploring methods for mathematical content access through its task design and the construction of the evaluation dataset.

INFORMATION ACCESS TO MATHEMATICAL CONTENT ?

[Formula]
a mathematical relationship or rule expressed in symbols
(Oxford Dictionary)

In science, a formula is a concise way of expressing information, or a general relationship between quantities.
(Wikipedia)

Math Information Access

Representations

Embedded image (png, gif, ...)

$$\log(z_1) + \log(z_2) == \log(z_1\, z_2)\ /;\ z_1 + z_2 \geq 0$$

Character sequence (latex source)

```
log(z_1)+log(z_2) == log(Z_1,
Z_2)¥¥; z_1+z_2 ¥geq 0
```

Web-browsable XML

<math xmlns='http://www.w3.org/1998/Math /MathML' mathematica:form='TraditionalForm' xmlns:mathematica='http://www.wolfram.com/XML /'> <semantics> <mrow> <mrow> <mrow> <mrow> <mi> log </mi> <mo> ⁡ </mo> <mo> (</mo> <msub> <mi> z </mi> <mn> 1 </mn> </msub>

XML for math semantics

<annotation-xml encoding='MathML-Content'> <apply> <ci> Condition </ci> <apply> <eq /> <apply> <plus /> <apply> <ln /> <apply> <ci> Subscript </ci> <ci> z </ci> <cn type='integer'> 1 </cn> </apply> </apply> <apply> <ln /> <apply> <ci> Subscript </ci>

Resources

mathematical knowledge-base and math ontology

OpenMath ⇔ Strict Content MathML (W3C recommendation)

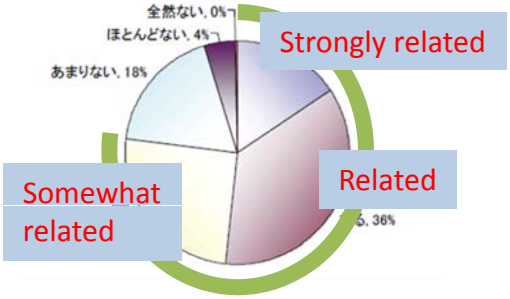
Wolfram MathWorld :
13,081 entries (Sep. 13, 2011)
Wolfram Functions site :
307,409 formulas (Sep. 15, 2011)

Wikipedia: 26,566 mathematics articles

Requirement

NISTEP Policy Study
Mathematics as deserted science in Japanese S&T policy
- Current situation on mathematical sciences research in major countries and need for mathematical sciences from the science in Japan (2006.5)

Q. Is mathematics related to your research?



77% researchers across diversity of disciplines answered 'YES'.

Task Overview

- **[Math Retrieval Subtask]**
  - Given a document collection, retrieve relevant mathematical formulae or documents for a given query.

- **[Math Understanding Subtask]**
  - Extract natural language definitions of mathematical expressions in a document for their semantic interpretation.

Math Retrieval Subtask

- **Dataset (scheduled)**
  - Scientific Articles from ArXiv e-print server http://arxiv.org/
    - Converted into XML+MathML by arXMLiv project http://kwarc.info/projects/arXMLiv/
    - 10,000 docs for a dry run, additional 100,000 docs for a formal run
- **Search Types**
  - The Math retrieval task uses the above 110K docs and can be envisaged in three different search scenarios
    - Formula Search
      - Search for formula queries within the formulae database of the used dataset
    - Full-Text Search
      - Search the document collection using formula queries. Combinations of keywords and formulae.
    - Open Information Retrieval
      - Search the document collection using free textual queries.

Math Retrieval Subtask : Dataset Example

```
            <m:mi id="id57134">S</m:mi>
            <m:mi id="id57136">I</m:mi>
            <m:mo id="id57138">&prime;</m:mo>
          </m:msubsup>
        </m:mrow>
        <m:mo id="id57141">&ne;</m:m
        <m:msup id="id57143">
          <m:mi id="id57144">GK</m:m
          <m:mo id="id57146">&prime;
        </m:msup>
      </m:mrow>
      <m:mo id="id57149">,</m:mo>
    </m:mrow>
    <m:annotation-xml id="id57151"
      <m:apply id="id57154">
        <m:neq id="id57155"/>
        <m:apply id="id57156">
          <m:apply id="id57157">
            <m:csymbol id="id57158" cd="ambiguous">superscript</m
            <m:apply id="id57163">
```

math representation (MathML)

Let $\{G_\gamma \mid \gamma \in \Gamma\}$ be a family of abelian groups. If G_γ is not a proper union of
then $G = \oplus_{\gamma\in\Gamma} G_\gamma$ is also not a proper union of finitely many cosets.

Demonstration Proof

To prove it by transfinite induction we have two cases to distinguish. If Γ is
with some Γ' and for Γ' the statement is true. Then we get $G = G_\gamma \oplus G'$, wi
coset of G with respect to a subgroup H such that $b + G_\gamma \subseteq S$ with some b
form $S = G_\gamma + S'$, where S' is a proper coset of G'.

Suppose that G is a proper union of the cosets $S_1, \cdots, S_n$. If S_l contains a co
written as $G_\gamma + S_l'$; otherwise, S_l' is the empty set. By induction

$$\bigcup_{l=1}^{n} S_l' \neq G',$$

therefore, there is a $d \in G'$, such that $d + G_\gamma$ is not contained in any S_l'. Mo
then it contains an $r_l + d$ and $S_l = r_l + d + G_l$, where $r_l \in G_\gamma$ and G_l is a subg

$$S_l \cap (d + G_\gamma) = (r_l + d + G_l) \cap (r_l + d + G_\gamma) = (r_l + d) + G_l \cap G_\gamma and$$

imply that G_γ is a proper union of some of the cosets $r_l + (G_l \cap G_\gamma)$, which c

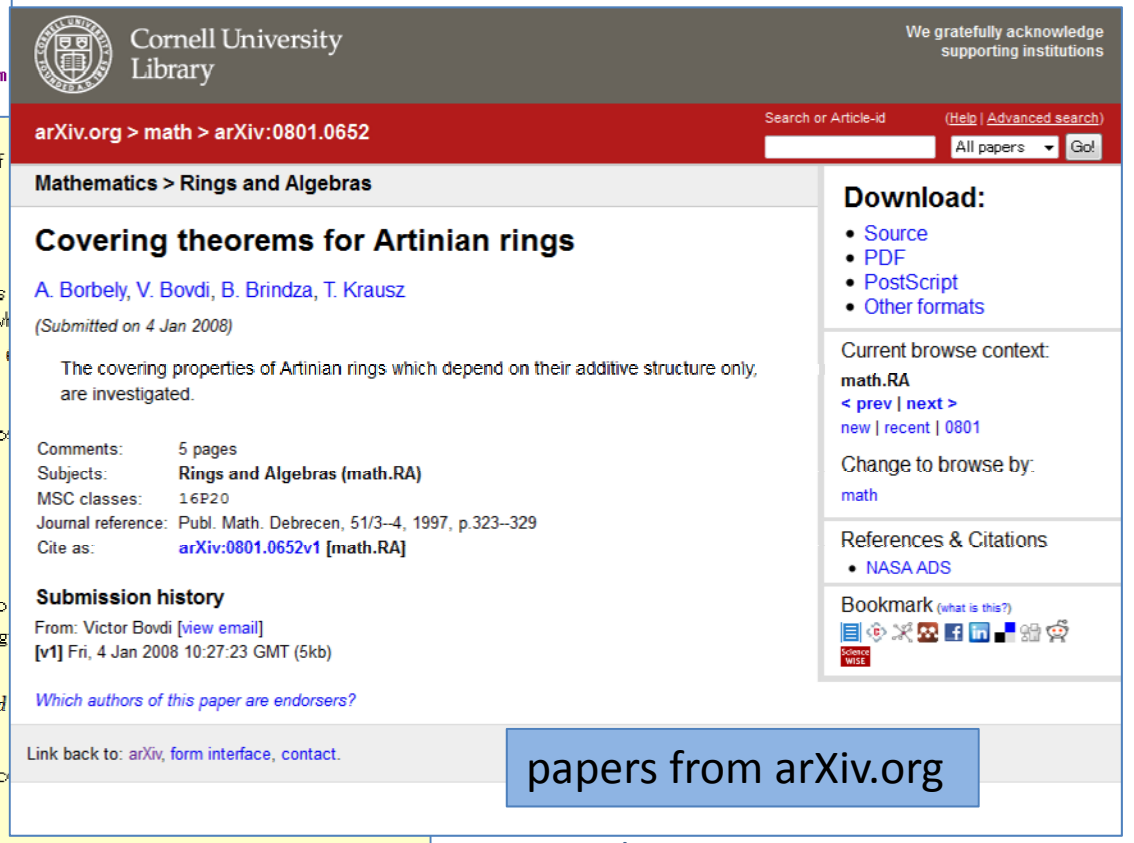
In the second case Γ is a limit ordinal. For a $\Gamma' < \Gamma$ set

$$G_{\Gamma'} = \bigoplus_{\alpha\in\Gamma'} G_\alpha.$$

Assuming G is a proper union of the cosets $T_1, \cdots, T_k$ we obtain

$$G_{\Gamma'} = \bigcup_{l=1}^{k} (G_\Gamma \cap T_l).$$

xhtml/xml



papers from arXiv.org

Math Retrieval Subtask : Query example

Formula search

1. $\int_{\boxed{l}}^{\boxed{h}} \boxed{f}(x)^2\,dx$

2. $\boxed{a} + \boxed{b} = \boxed{b} + \boxed{a}$

3. $\lim_{\boxed{x}\to\infty} \boxed{a} + \boxed{b} = \boxed{c}$

Full-text search

1. Pythagorean Theorem

2. Bell curve in the form of $\frac{1}{\boxed{\sigma}\sqrt{2\pi}} \exp^{-\frac{(\boxed{x}-\boxed{\mu})^2}{2\boxed{\sigma}^2}}$

3. NOT Bayes Theorem, but written as $P(\boxed{X} \mid \boxed{Y})$

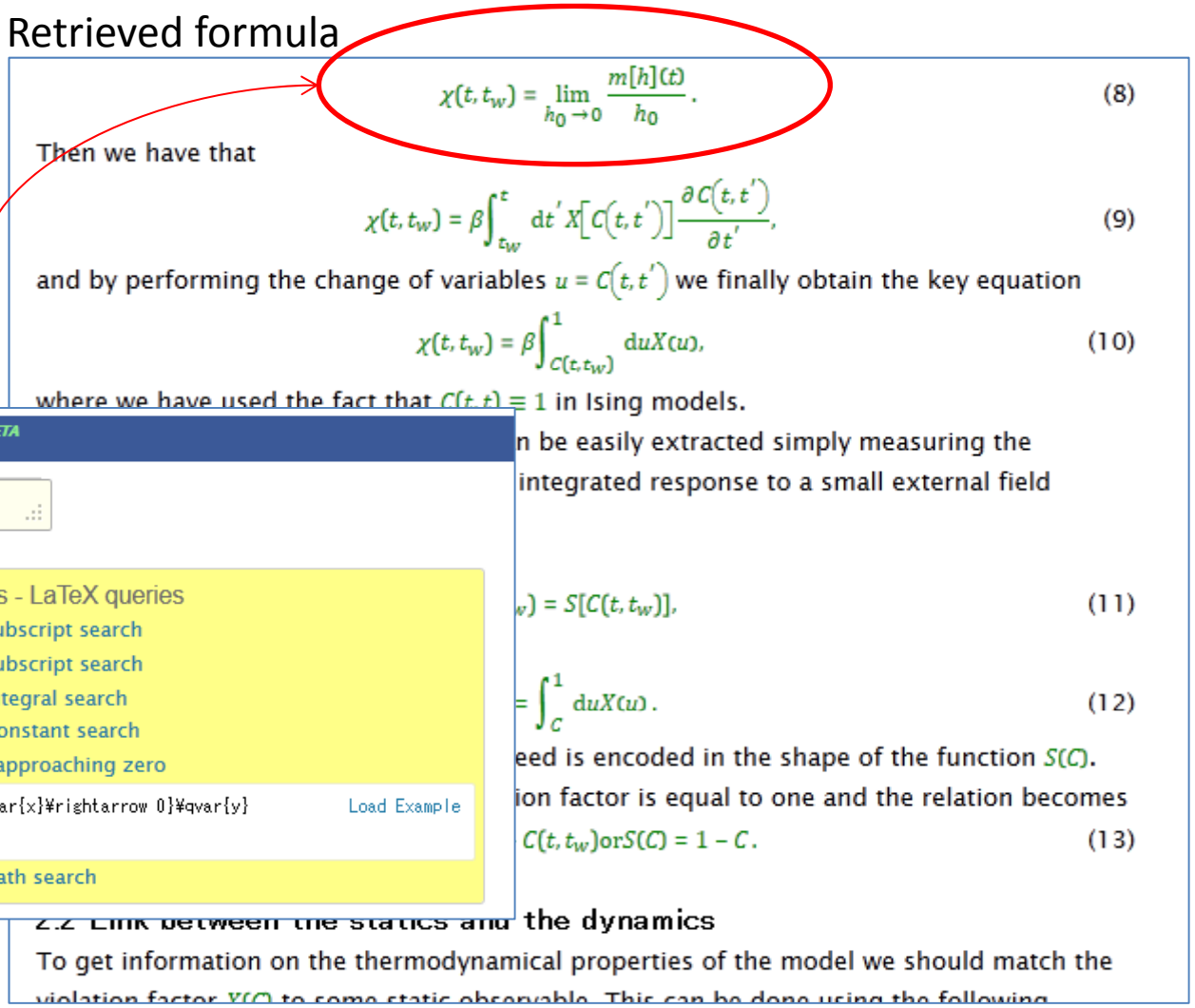
Open IR

1. What is the fifth summand in the Taylor expansion of sinus hyperbolicus?

2. For which n and k is $PSL(n,k)$ not commutative.?

Math Retrieval Subtask : Formula Search Example

Retrieved formula



$$\chi(t,t_w) = \lim_{h_0 \to 0} \frac{m[h](t)}{h_0}. \quad (8)$$

Then we have that

$$\chi(t,t_w) = \beta \int_{t_w}^{t} dt' X\left[C\left(t,t'\right)\right] \frac{\partial C\left(t,t'\right)}{\partial t'}, \quad (9)$$

and by performing the change of variables $u = C\left(t,t'\right)$ we finally obtain the key equation

$$\chi(t,t_w) = \beta \int_{C(t,t_w)}^{1} du X(u), \quad (10)$$

where we have used the fact that $C(t,t) \equiv 1$ in Ising models.

n be easily extracted simply measuring the

integrated response to a small external field

$$_w) = S[C(t,t_w)], \quad (11)$$

$$= \int_{C}^{1} du X(u). \quad (12)$$

eed is encoded in the shape of the function $S(C)$.

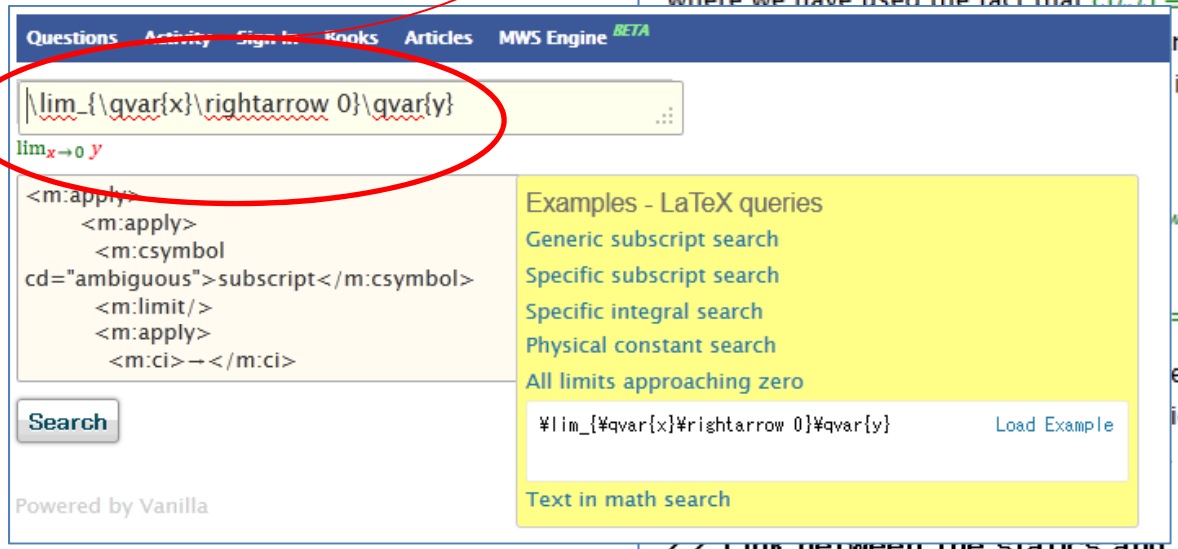
ion factor is equal to one and the relation becomes

$$C(t,t_w) \text{ or } S(C) = 1 - C. \quad (13)$$

2.2 Link between the statics and the dynamics

To get information on the thermodynamical properties of the model we should match the violation factor $X(C)$ to some static observable. This can be done using the following

Query formula



Math Understanding Subtask

- **Task definition**
  - Extract natural language definitions of mathematical expressions in a document (Basic Task) with their semantic interpretation (Challenge Task, TBA).

- **Dataset (scheduled)**
  - Development Data
    - 10 papers selected from ACL-Anthology Reference Corpus
    - 30 papers selected from ArXiv.org dataset which will be also used in Math Retrieval Task.
  - Data for Formal Run (submission period: five days)
    - 10 papers selected from ACL-Anthology Reference Corpus
    - 10 papers selected from ArXiv.org dataset which will be also used in Math Retrieval Task.

Math Understanding Subtask : Dataset Example

Original text (pdf)

The similarity between G_1 and G_2 is then defined as follows:

$$groupSim(G_1, G_2) = RI(G_1, G_2) \times RC(G_1, G_2)$$

where

$$RI(G_1, G_2) = \frac{2AI(G_1, G_2)}{II(G_1) + II(G_2)}$$

is the relative interconnectivity and

$$RC(G_1, G_2) = \frac{AC(G_1, G_2)}{\frac{|G_1|}{|G_1|+|G_2|} IC(G_1) + \frac{|G_2|}{|G_1|+|G_2|} IC(G_2)}$$

is the relative closeness.

MathML parallel markup

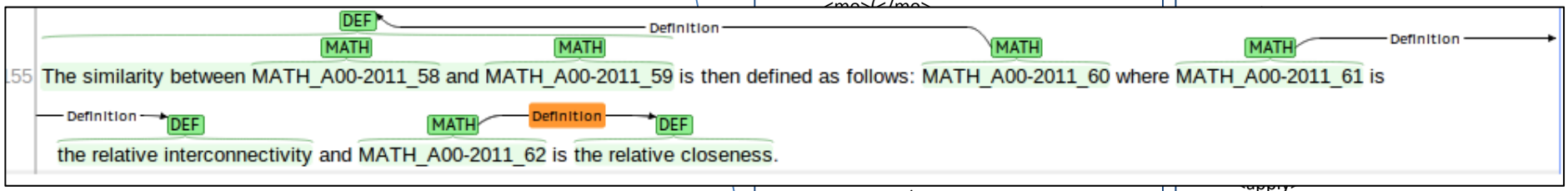
Example of Presentation MathML

```
<mrow>
  <mrow id="math_b6a1eb64.1">
   <mi mathvariant="italic">groupSim</mi>
   <mo>?</mo>
   <mrow>
    <mo>(</mo>
    <mrow>
     <msub>
      <mi>G</mi>
      <mn>1</mn>
     </msub>
     <mo>,</mo>
     <msub>
      <mi>G</mi>
      <mn>2</mn>
     </msub>
    </mrow>
    <mo>)</mo>
   </mrow>
  </mrow>
  <mo>=</mo>
  <mrow id="math_b6a1eb64.2">
   <mrow>
    <mi mathvariant="italic">RI</mi>
    <mo>?</mo>
    <mrow>
     <mi>G</mi>
     <mn>2</mn>
    </mrow>
```

Example of Content MathML

```
<annotation-xml encoding="MathML-Content"
  <apply>
   <eq />
   <apply xref="math_b6a1eb64.1">
    <ci>groupSim</ci>
    <apply>
     <selector />
     <ci>G</ci>
     <cn>1</cn>
    </apply>
    <apply>
     <selector />
     <ci>G</ci>
     <cn>2</cn>
    </apply>
   </apply>
   <apply xref="math_b6a1eb64.2">
    <times />
    <apply>
     <ci>RI</ci>
     <apply>
      <selector />
      <ci>G</ci>
      <cn>1</cn>
```

Definitions of mathematical objects in text




```
    <mi>G</mi>
    <mn>2</mn>
   </msub>
  </mrow>
```

```
    <ci>RC</ci>
    <apply>
     <selector />
     <ci>G</ci>
```

* Annotation tool (http://www.nactem.ac.uk/brat-annotation/)

Schedule

February-April, 2012	Task framework development
April, 2012	Call for participation
May, 2012	Dataset and example topics release
Early-mid October, 2012	Topic release for Open IR search type
Late October, 2012	Topic release for Formula, Full-text search types, and Math understanding subtask
Early November, 2012	Results submissions due for Formula, Full-text search types, and Math Understanding subtask
Mid November, 2012	Results submissions due for Open IR search type
February, 2013	Evaluation Results Released
March, 2013	Draft papers for NTCIR-10 Proceedings Due
May, 2013	Camera ready for NTCIR-10 Proceedings Due
June, 2013	NTCIR-10 Meeting

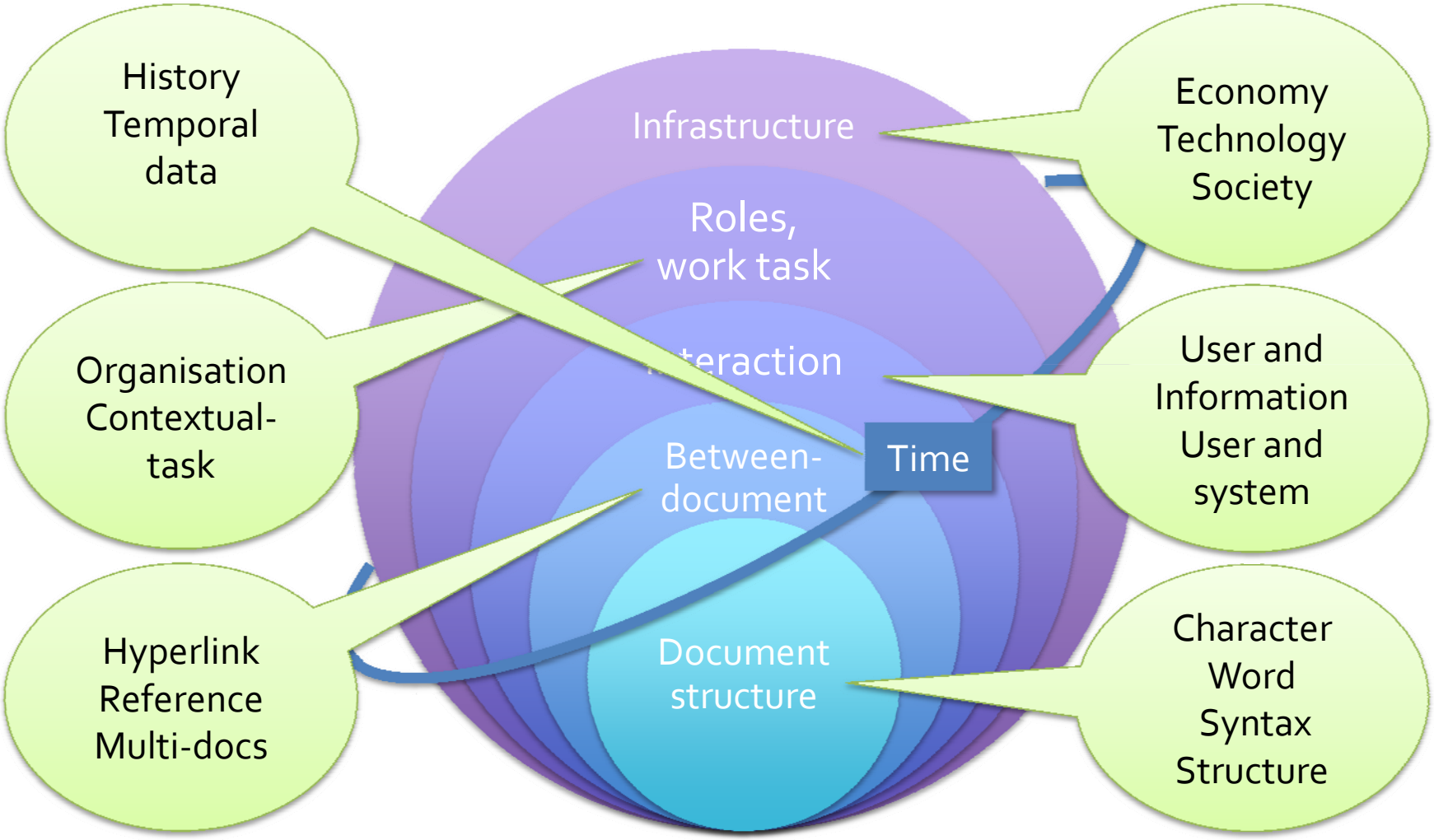
Task information

- **Contact**
  - ntcir10adm-math@nii.ac.jp
- **Task Web page**
  - http://ntcir-math.nii.ac.jp/
- **Task Organizers**
  - Akiko Aizawa (National Institute of Informatics, Japan)
  - Michael Kohlhase (Jacobs University Bremen)
  - Iadh Ounis (University of Glasgow)
- **Task Advisors**
  - Noriko Kando (National Institute of Informatics, Japan)
  - Fredric C. Gey (University of California, Berkeley)

NTCIR-10 Task Map

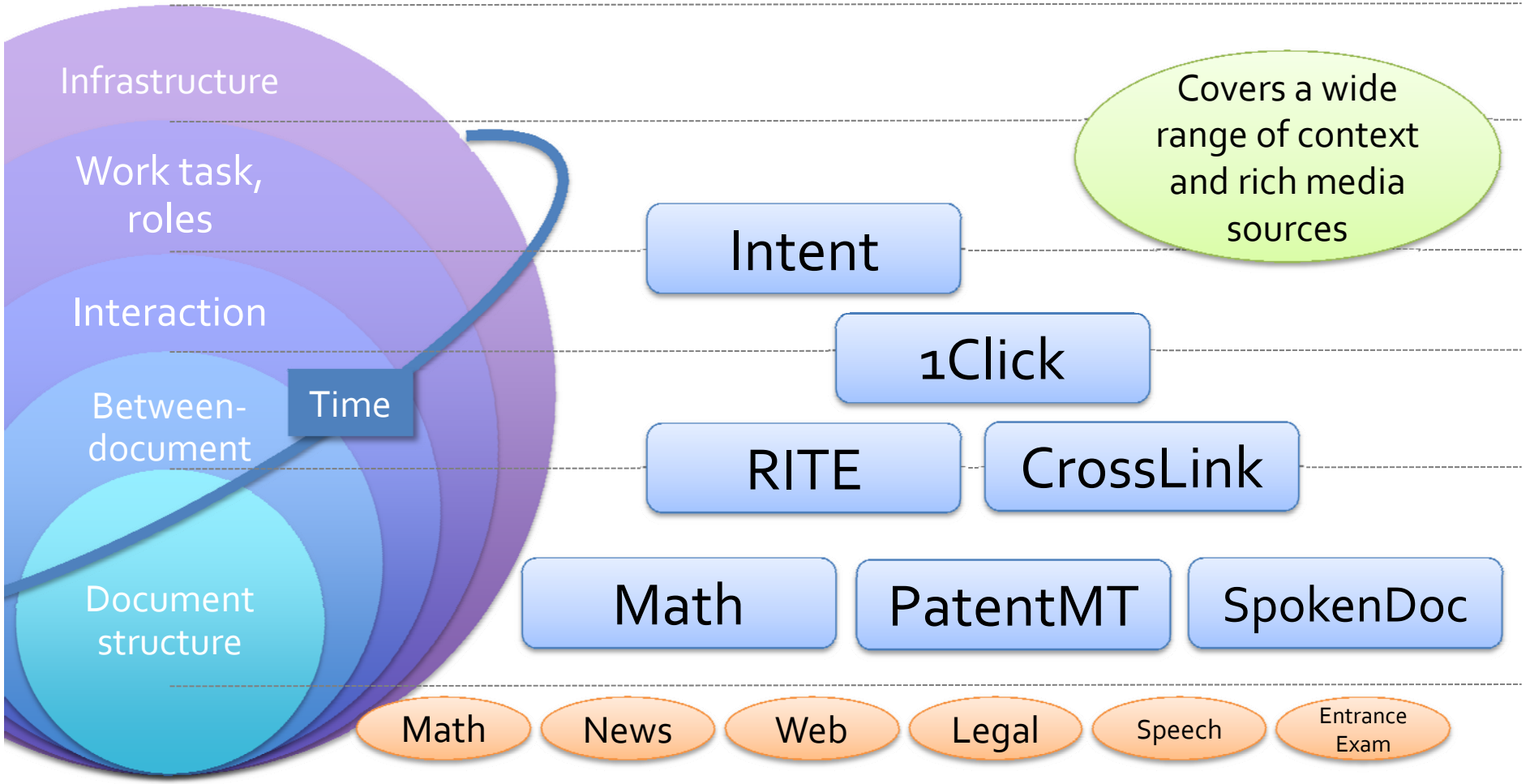
Summary

Context of Information Access

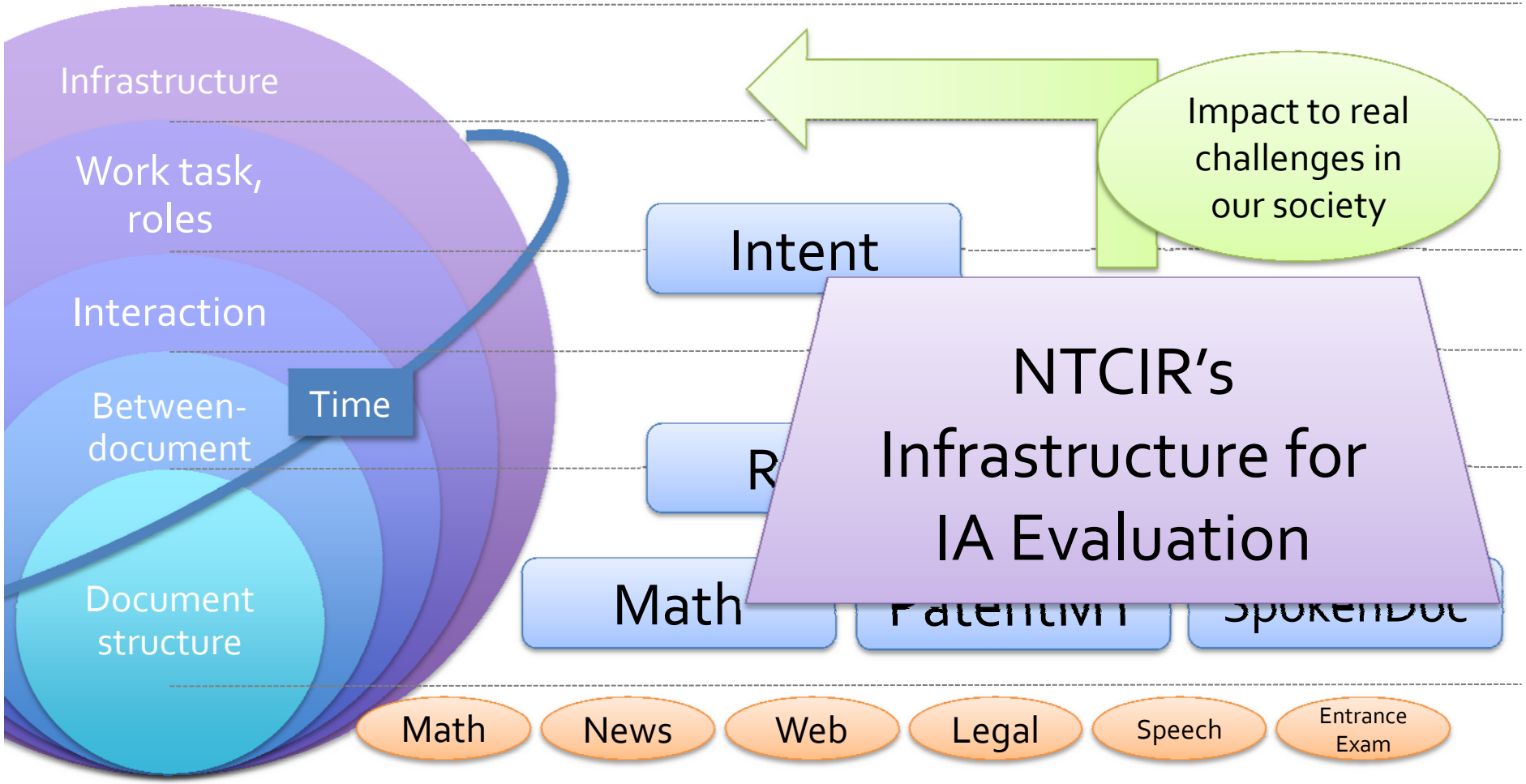


Adapted from Ingwersen & Järvelin (2005)

NTCIR-10 Tasks



NTCIR's Long-Term View



Why participate?

Case for students and industry

Why participate? (Students)

- Easy start-up
  - Much of experimental setup is provided
  - Performance measures are (often) defined
- Publications
  - Comparison with other participants can produce stronger arguments
  - Inspired by the international community for future work
- Diverse tasks
  - Range of Information access tasks to tackle

Why participate? (Industry)

- Establish your brand
  - To your end-users and competitors
  - Recruit smart people
- Fair benchmarking
  - Comparison with your own products can be biased
  - Critical self-assessments
- Faster development
  - Brush up your product or eliminating bugs in a short period of time
- Early access to resulted resources
  - Secondary resources developed by the task are yours, too

How to participate

Simple six steps

How to participate

1. Read the task description and CFP carefully
2. Contact a TO if you have questions
3. Decide a task to participate
4. Register as a participant at NTCIR website
5. Fill in User Agreement Forms
6. Keep an eye on a task's ML, website, etc. to follow the activity

Reminder for participants

Participant's agreement

- Submit a paper to NTCIR-10 Workshop Meeting
- Attend to the meeting and give a poster presentation

Important Dates

For your diary

Important Dates

08/03/2020	Kick-off event in Tokyo
30/06/2012	Task registration due
01/07/2012	Document set release
07 - 12/2012	Dry run
09/2012 - 02/2013	Formal run
01/02/2013	Evaluation results due
01/02/2013	Task overview partial release
01/03/2013	Participant paper submission due
01/05/2013	All camera-ready copy for the Proceedings due
18-23/06/2013	NTCIR-10 Workshop Meeting, NII, Tokyo, Japan

Contact TO for the exact schedule

Wrap-up

Memorial

- The tenth cycle of NTCIR has started
  - Even stronger structure
- Seven exciting tasks are running
  - Organised by 48 researchers worldwide
- Lots of opportunities for innovative work
  - Exchange great ideas with the community
- What's missing is **your participation**!

http://research.nii.ac.jp/ntcir/ntcir-10/

Thank you for your attention!

For further enquiries, contact the NTCIR office
ntc-secretariat@nii.ac.jp

Q & A