

## Overview of NTCIR-8

NTCIR



Thanks for Teruko Mitamura, Tetsuya Sakai, Fred Gey, Yohei Seki, Daisuke Ishikawa, Atsushi Fujii, Hidetsugu Nanba, Terumasa Ehara for preparing slides

**Noriko Kando**

National Institute of Informatics, Japan

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

kando@nii.ac.jp

NTCIR



**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

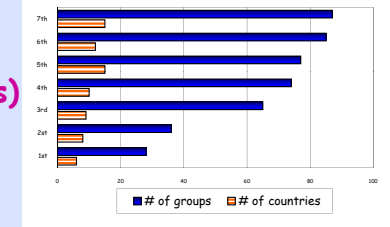
**Data sets (Test collections or TCs)**

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

**Tasks (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-7 participants**  
82 groups from 15 countries

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

NTCIR	1	2	3	4	5	6	7	8	
	'99	'01	'02	'04	'05	'07	'08	'09-	
User Generated Contents						■	■		■ Community QA ■ Opinion Analysis
Module-Based							■	■	■ Cross-Lingual QA + IR
IR for Focused Domain			■	■	■	■	■	■	■ Geo Temporal ■ Patent
Question Answering				■	■		■	■	■ Complex/ Any Types ■ Dialog ■ Cross-Lingual ■ Factoid, List
Summarization / Consolidation			■	■	■	■	■	■	■ Text Mining / Classification ■ Trend Info Visualization ■ Text Summarization
Web		■	■	■					■ Web
Crosslingual Retrieval	■	■	■	■	■	■	■	■	■ Statistical MT ■ Cross-Lingual IR ■ Non-English Search
Text Retrieval	■	■	■	■	■	■	■	■	■ Ad Hoc IR, IR for QA

## NTCIR-8 Tasks (2008.07–2009.06)

The 3<sup>rd</sup> Intl W on Evaluating Information Access (EVIA) refereed

- |   |  |
|---|--|
| <b>1. Advanced CL Info Access</b><br>- QA(CsC+JE->CsC+J)     Any Types of Questions   |  |
| GeoTime (E, J)     Geo Temporal Information   |  |
| <b>2. User Generated Contents (CGM)</b><br>- Opinion Analysis (Multilingual News) (Blog)<br>- [Pilot] Community QA (Yahoo! Chiebukuro)  |   |
| <b>3. Focused Domain: Patent</b><br>-Patent Translation; English -> Japanese<br>Statistical MT, The World-Largest training data (J-E sentence alignment), Summer School, Extrinsic eval by CLIR<br>-Patent Mining papers -> IPC<br>-Evaluation of SMT |   |

## NTCIR-7 & -8 Program Committee



Mark Sanderson, Doug Oard, Atsushi Fujii, Tatsunori Mori, Fred Gey, Noriko Kando (and Ellen Voorhees, Sung Hyun Myaeng, Hsin-Hsi Chen, Tetsuya Sakai)

## NTCIR-8 Coordination

- NTCIR-8 is coordinated by NTCIR Project at NII, Japan. The following organizations contribute to the organization of NTCIR-8 as Task Organizers
  - Academia Sinica
  - Carnegie Mellon Univ
  - Chinese Academy of Science
  - Hiroshima City University
  - Hitachi, Co Ltd.
  - Hokkai Gakuen University
  - IBM
  - Microsoft Research Asia
  - National Institute of Information and Communication Technology
  - National Institute of Informatics
  - National Taiwan Univ
  - National Taiwan Ocean Univ
  - Oki Electronic Co.
  - Tokyo Institute of Technology
  - Tokyo Univ
  - Toyohashi Univ of Technology and Science
  - Univ of California Berkeley
  - Univ of Tsukuba
  - Yamanashi Eiwa College
  - Yokohama National University

<p>[CCLQA] Carnegie Mellon Univ Dalian Univ of Technology National Taiwan Ocean Univ Shenyang Institute of Aeronautical Engineering Univ of Tokushima Wuhan Univ</p> <p>[IR4QA] Carnegie Mellon Univ Chaoyang Univ of Technology Dalian Univ of Technology Dublin City Univ Inner Mongolia Univ Queensland Univ of Technology Shenyang Inst of Aeronautical Engineering Trinity College Dublin Univ California, Berkeley Wuhan Univ Wuhan Univ (Computer School) Wuhan Univ of Science and Technology</p>	<p>[GeoTime] Dublin City Univ Hokkaido Univ INESC-ID, Portugal International Inst of Technology, Hyderabad Kioo Univ Nataional Inst of Materials Science Osaka Kyoiku Univ Univ California, Berkeley Univ of Iowa Univ of Lisbon Yokohama City Univ</p> <p>[MOAT] Beijing Uni of Posts and Telecommunications Chaoyang Univ of Technology Chinese Univ of HK+ Tsinghua Univ City Univ of Hong Kong (2 groups) Hong Kong Polytechnic Univ KAIST National Taiwan Univ NEC Laboratories China Peking Univ Pohang Univ of Sci and Tech SICS Toyohashi Univ of Technology Univ of Alicante Univ of Neuchatel Yuan Ze Univ</p>	<p>[Patent Mining] Hiroshima City Univ Hitachi, Ltd. IBM Japan, Ltd. Institute of Scientific and Technical Information of China KAIST National Univ of Singapore NEC Shanghai Jiao Tong Univ Shenyang Institute of Aeronautical Engineering Toyohashi Univ of Technology Univ of Applied Sciences - UNIGE</p> <p>[Patent Translation] Dublin City University, CNGL Hiroshima City University Kyoto University NiCT Pohang Univ of Sci and Tech tottori university Toyohashi University of Technology Yamanashi Eiwa College</p> <p>[Community QA] Microsoft Research Asia National Institute of Informatics Shirayuri College</p>
---	--	---

### NTCIR-8 Active Participants

## 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

## Tasks of NTCIR-8

## ACLIA

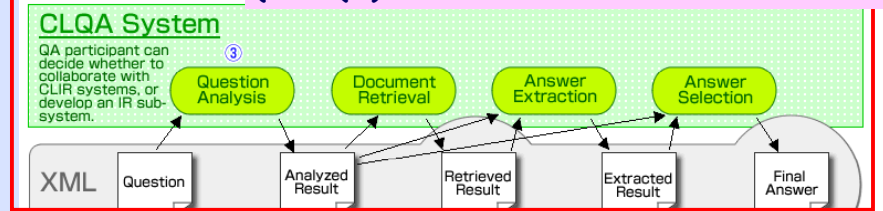
### Advanced Cross-Lingual Information Access

CCLQA: **C**omplex **C**ross-**L**ingual **Q**uestion **A**nswering

IR4QA: **I**nformation **R**etrieval **f**or **Q**uestion **A**nswering

Teruko Mitamura, Hideki Shima, Tetsuya Sakai,  
Noriko Kando, Tatsunori Mori, Koichi Takeda,  
Chin-Yew Lin, Ruihua Song  
Chuan-Jie Lin, Cheng-Wei Lee  
<http://aclia.lti.cs.cmu.edu/ntcir8>

## Complex Cross-lingual Question Answering (CCLQA) Task



**CLIR System**

Small teams that do not possess an entire QA system can contribute

- ① in source language.
- ② In collaboration with CLQA, CLIR system can also take translated keyterms and answer type analysis.
- ③ Translation often happens in here.

Different teams can exchange and create a "dream-team" QA system

IR Evaluation

QA Evaluation

IR and QA communities can collaborate

<http://aclia.lti.cs.cmu.edu/ntcir8/>

## Evaluation Topics - any types of questions -

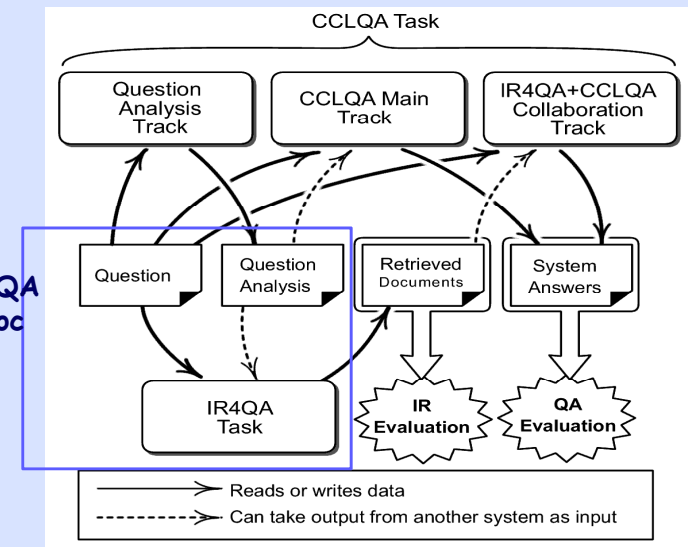
Type	#	Example Question	Related Past NTCIR Task
DEFINITION	10	What is the Human Genome Project?	ACLIA
BIOGRAPHY	10	Who is Howard Dean?	ACLIA
RELATIONSHIP	20	What is the relationship between Saddam Hussein and Jacques Chirac?	ACLIA
EVENT	20	What are the major conflicts between India and China on border issues?	ACLIA
WHY	20	Why doesn't U.S. ratify the Kyoto Protocol?	QAC-4
PERSON	5	Who is the Finland's first woman president?	QAC 1-3, CLQA 1,2
ORGANIZATION	5	What is the name of the company that produced the first Fairtrade coffee?	QAC 1-3, CLQA 1,2
LOCATION	5	What is the name of the river that separates North Korea from China?	QAC 1-3, CLQA 1,2
DATE	5	When did Queen Victoria die?	QAC 1-3, CLQA 1,2

13

NTCIR-8, June 16, 2010

## ACLIA Data Flow

IR for QA  
ad hoc  
IR



14

NTCIR-8, June 16, 2010

## SEPIA (Evaluation Toolkit) Standard Evaluation Package for Information Access

- Topic development tool
  - Adding Question, Information need, answer type
  - Extracting answer nuggets
  - Voting on vital nuggets
- IR/QA system response submission/exchange tool
- Human Evaluation tool
  - Relevance Judgment for IR
  - Nugget Matching for QA
- XML import/export
- Shared workspace for multiple user collaboration
  - Online system, IR system embedded.
- Admin tool for reporting task progress statistics, user activity logs, user contribution statistics, etc.
  - Will release as an open source

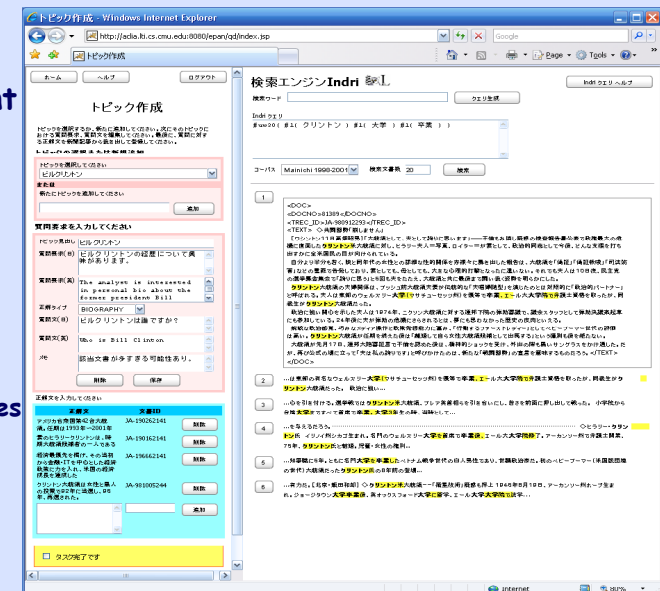
more info: visit poster on SEPIA today

15

NTCIR-8, June 16, 2010

## Topic Development Interface

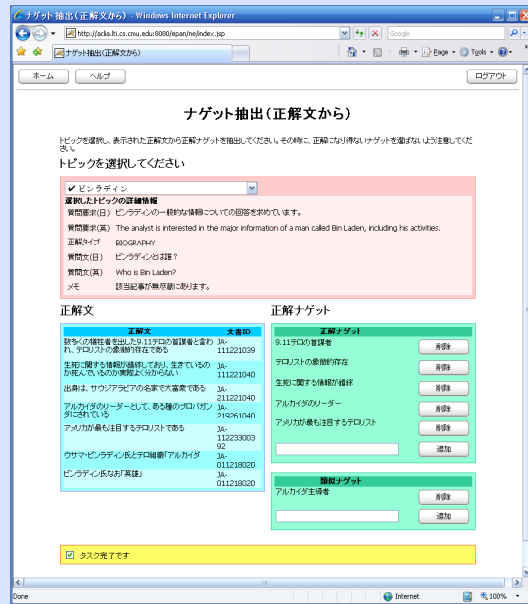
Search answers  
Extract sentences



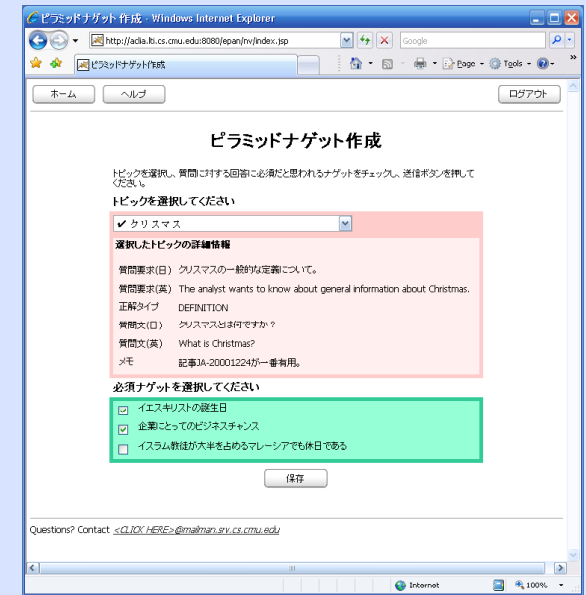
16

NTCIR-8, June 16, 2010

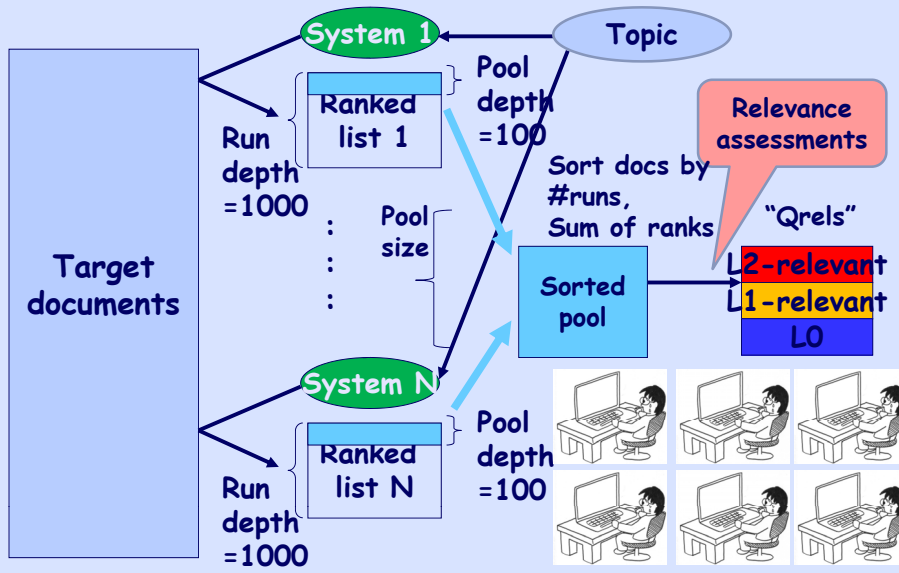
# Nugget Extraction Tool



# Nugget Voting Tool



# IR4QA Pooling/Relevance assessments



# A version of Normalised Discounted Cumulative Gain [Jarvelin&Kekalainen SIGIR00, TOIS02]

Discounted cumulative gain for system output

$$nDCG = \frac{\sum_{r=1}^l g(r) / \log(r + 1)}{\sum_{r=1}^l g^*(r) / \log(r + 1)}$$

Discounted cumulative gain for ideal output

The most popular graded-relevance IR metric



## Q-measure [Sakai AIRS04/05, SIGIR06/07, IPM07, EVIA07/08, IRJ09...]

$$Q\text{-measure} = \frac{1}{R} \sum_r I(r) \frac{C(r) + \beta \text{cg}(r)}{r + \beta \text{cg}^*(r)}$$

Average Precision

Cumulative gain for system output

Cumulative gain for ideal output

- Used for NTCIR CLIR, IR4QA, GeoTime and Community QA
- Used by other researchers e.g. Kazai/Lalmas TOIS06, Al-Maskari/Sanderson LWA06, Zhang/Park/Moffat IRJ09 etc.

## IR evaluation tool

[http://research.nii.ac.jp/ntcir/tools/ir4qa\\_eval-en.html](http://research.nii.ac.jp/ntcir/tools/ir4qa_eval-en.html)

NTCIR Project  
Tools  
ir4qa\_eval

[JAPANESE] [NTCIR]

Works for ad hoc IR,  
Diversity IR [Sakai et al EVIA2010],  
community QA [Sakai et al NTCIR-8 CQA]

ir4qa\_eval

These information retrieval (IR) evaluation scripts were developed for the NTCIR-7 ACLIA IR4QA subtask. They can be used for other IR tasks at NTCIR, TREC, etc. The scripts can compute average precision, Q-measure, nDCG and some other evaluation metrics. For more details, please read the README file included in the tar file.

Download

[http://research.nii.ac.jp/ntcir/tools/cqa\\_eval.tar.gz](http://research.nii.ac.jp/ntcir/tools/cqa_eval.tar.gz)

(A patch for the NTCIR-8 community pilot task. ir4qa\_eval2 also needs to be installed)

[http://research.nii.ac.jp/ntcir/tools/ir4qa\\_eval2.tar.gz](http://research.nii.ac.jp/ntcir/tools/ir4qa_eval2.tar.gz) (latest version)

[http://research.nii.ac.jp/ntcir/tools/ir4qa\\_eval.tar.gz](http://research.nii.ac.jp/ntcir/tools/ir4qa_eval.tar.gz)

## F-score definition

Let  
 $r$  sum of weights over matched nuggets  
 $R$  sum of weights over all nuggets  
 $a_{HUMAN}$  # of nuggets matched in SRs by human  
 $L$  total character-length of SRs  
 $C$  character allowance per match  
 $allowance$   $a_{HUMAN} \times C$

Then

$$recall = \frac{r}{R}$$

$$precision = \begin{cases} 1 & \text{if } L < allowance \\ \frac{allowance}{L} & \text{otherwise} \end{cases}$$

$$F(\beta) = \frac{(\beta^2 + 1) \times precision \times recall}{\beta^2 \times precision + recall}$$

CS:  $C=18$

CT:  $C=27$

JA:  $C=24$

Based on the micro-average character length of the nuggets in the formal run dataset

If the total length exceeds the allowance, the score is penalized.

## F3 Score

- $F(\beta=3)$  score weights recall 3x more than precision
- Example: 5 gold standard nuggets {1.0, 0.4, 0.2, 0.5, 0.7}  
A total characters of system responses = 200

$$recall = \frac{0.4 + 0.7}{1.0 + 0.4 + 0.2 + 0.5 + 0.7} = 0.39$$

$$precision = \frac{2 \times 24}{200} = 0.24$$

$$F(\beta=3) = \frac{10 \times 0.24 \times 0.39}{9 \times 0.24 + 0.39} = 0.37$$

- The evaluation score for the example question is  $F3 = 0.37$

**Table2. Corpora used in ACLIA2.**

Language	Corpus Name	Time Span	# document
CS	Xinhua	2002-2005	308,845
CT	UDN	2002-2005	1,663,517
JA	Mainichi	2002-2005	377,941

Topics from CCLQA with at least 5 relevant docs:  
73 for CS, 94 for JA and 87 for CT

## CT/JA-T IR4QA run rankings

	Mean Q		Mean nDCG
KDEG-CT-CT-01-T	0.4977**	KDEG-CT-CT-01-T	0.7056**
QUTIS-EN-CT-04-T	0.3354**	QUTIS-EN-CT-04-T	0.5412**
CYUT-EN-CT-02-T	0.1945	CYUT-EN-CT-02-T	0.3847

	Mean Q		Mean nDCG
LTI-JA-JA-01-T	0.4001	LTI-JA-JA-01-T	0.5977
BRKLY-JA-JA-02-T	0.3832**	BRKLY-JA-JA-02-T	0.5694**
CYUT-EN-JA-02-T	0.1788	CYUT-EN-JA-02-T	0.3638

## CCLQA Human Evaluation Preliminary Results: JA-JA

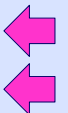
JA-JA Runs	ALL
LTI-JA-JA-01-T	0.1069
LTI-JA-JA-02-T	0.1443
LTI-JA-JA-03-T	0.1438

## JA-JA automatic evaluation

JA-JA Runs	ALL
LTI-JA-JA-01-T	0.2024
LTI-JA-JA-02-T	0.2259
LTI-JA-JA-03-T	0.2252

## Effect of Combination: IR4QA+CCLQA JA-JA Collaboration Track: F3 score based on automatic evaluation

		CCLQA
		LTI
IR 4 Q A	BRKLY-JA-JA-01-DN	0.2934
	BRKLY-JA-JA-02-T	0.2686
	BRKLY-JA-JA-03-DN	0.2074
	BRKLY-JA-JA-04-DN	<b>0.3000</b>
	BRKLY-JA-JA-05-T	0.2746



## NTCIR-GEOTIME GEOTEMPORAL INFORMATION RETRIEVAL

(New Track in NTCIR Workshop 8)

Fredric Gey and Ray Larson and Noriko Kando  
Relevance judgments system by Jorge Machado and Hideki Shima  
Evaluation : Tetsuya Sakai

Judgments: U Iowa, U Lisbon, U California Barkelay, NII  
Search with a specific focus on Geography + To distinguish  
from past GIR evaluations, we introduced a temporal  
component

Asian language geographic search has not previously been  
evaluated, even though about 50 percent of the NTCIR-6  
Cross-Language topics had a geographic component (usually a  
restriction to a particular country).

## NTCIR-GeoTime LANGUAGES and COLLECTIONS

Japanese and English

- Japanese: Mainichi News 2002-2005 (same as ACLIA-IR4QA)
- English: New York Times (NYT) 2002-2005 (used for MOAT)
  - Part of LDC's English Gigawords
  - cost \$50US for DVD
- Problems with missing news articles within NYT in 2003

## NTCIR-GeoTime TOPIC DEVELOPMENT

- 25 topics developed in English, then translated to Japanese
  - developed as questions with answers from Wikipedia
  - 5 topics of the event identification (when and where)
    - When and where did Katharine Hepburn die? or variation
    - How old was Max Schmeling (boxing champion) when he died and where did he die?
  - Other, somewhat more difficult topics:
    - How long after the Sumatra earthquake did its tsunami hit Sri Lanka?
- COMMUNITY-BASED DEVELOPMENT (a first at NTCIR)
  - Participating groups suggested some topics
  - English relevance assessment system developed by Jorge Machado of Portugal
  - English relevance assessment provided by three groups:
    - Univ of Iowa, Univ of Portugal, Lisbon, Univ of California, Berkeley
  - Japanese provided by NII

## NTCIR-GeoTime PARTICIPANTS

- Japanese Runs submitted by eight groups (two anonymous pooling supporters)

Team Name	Organization
Anonymous	Anonymous
BRKLY	University of California, Berkeley, USA
FORST	Yokohama National University, JAPAN
HU-KB	Hokkaido University, JAPAN
KOLIS	Keio University, JAPAN
ANON2	Anonymous submission, group 2
M	National Institute of Materials Science, JAPAN
OKSAT	Osaka Kyoiku University, JAPAN



## NTCIR-GeoTime PARTICIPANTS

- English Runs submitted by six groups

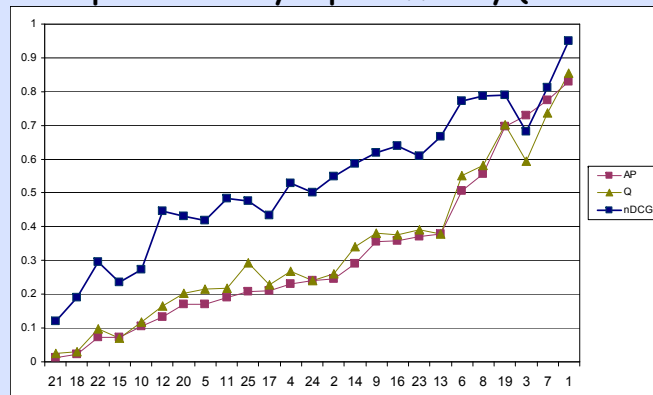
Team Name	Organization
BRKLY	University of California, Berkeley, USA
DCU	Dublin City University, IRELAND
IITH	International Institute of Technology, Hyderabad, INDIA
INESC	National Institute of Electroniques and Computer Systems, Lisbon, PORTUGAL
UIOWA	University of Iowa, USA
XLDB	University of Lisbon, PORTUGAL

## NTCIR-GeoTime Approached

- BRKLY: baseline approach, probabilistic + psued relevance feedback
- DCU, IITH, XLDB (U Lisbon) : geographic enhancements
- KOLIS (Keio U) : counting the number of geographic and temporal expression in top-ranked docs in initial search, then re-rank
- FORST (Yokohama Nat U): utilize factoid QA technique to question decomposition
- HU-KB (Hokkaido U), U Iowa: Hybrid approach combining probabilistic model and weighted boolean query formulation

## NTCIR-GeoTime: ENGLISH TOPIC DIFFICULTY by Average Precision

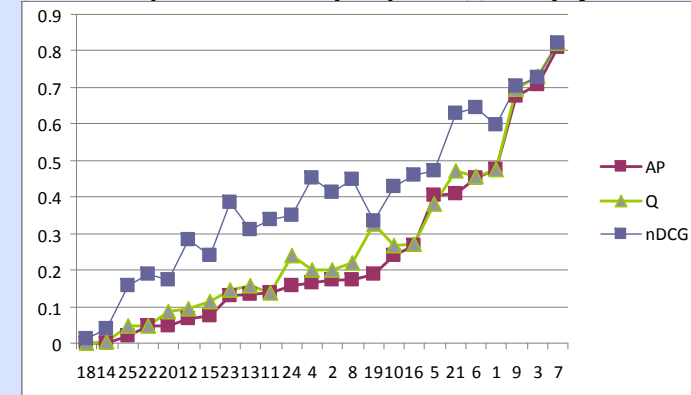
Per-topic AP, Q and nDCG averaged over 25 English runs for 25 topics sorted by topic difficulty (AP ascending)



Most Difficult English topic (21): *When and where were the 2010 Winter Olympics host city location announced?*

## NTCIR-GeoTime: JAPANESE TOPIC DIFFICULTY by Average Precision

Per-topic AP, Q and nDCG averaged over 34 Japanese runs for 24 topics sorted by topic difficulty (AP ascending)



Most Difficult Japanese topic 18: *What date was a country was invaded by the United States in 2002?*

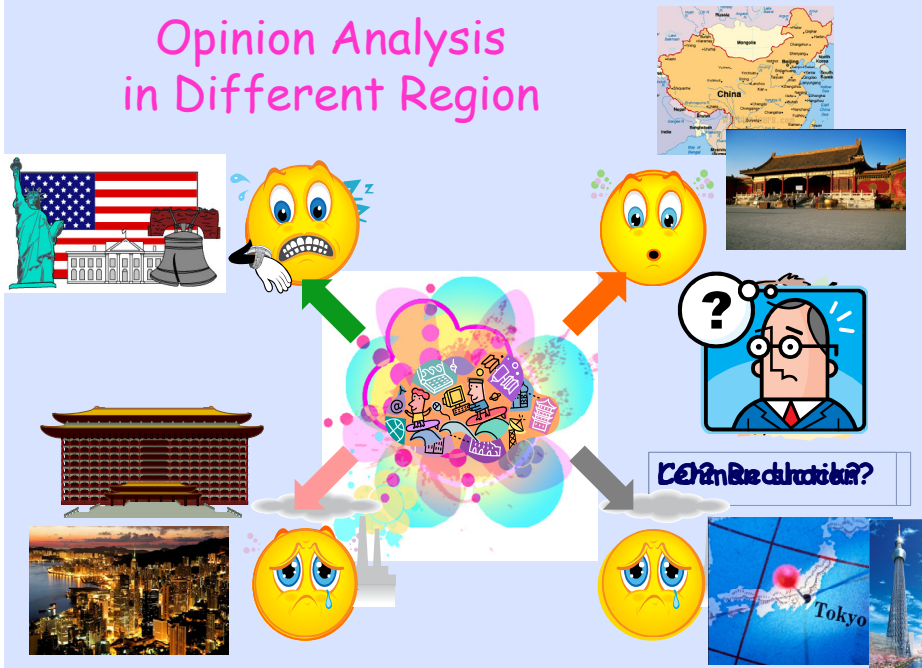
## NTCIR-GeoTime CHALLENGES

- Geographic reference resolution is difficult enough, but
- More difficult to process temporal expression ("last Wednesday") references
- Can indefinite answers be accepted ("a few hours")?
- Need Japanese Gazetteers
- Need NE annotated corpus for further refinement


## Multilingual Opinion Analysis Task (MOAT)

Yohei Seki (Toyohashi University of Technology),  
Lun-Wei Ku (National Taiwan University),  
Le Sun (Institute of Software, Academy of Sciences),  
Hsin-Hsi Chen (National Taiwan University),  
Noriko Kando (NII), and David Kirk Evans (Amazon Japan)

## Opinion Analysis in Different Region



## MOAT progress in NTCIR 6, 7, 8

- What's new in NTCIR-8 MOAT? 
- New subtask: *cross-lingual opinion Q&A*
- Focused application: *opinion Q&A*
- Update corpora: *NYT, UDN published in 2002-2005*

NTCIR	Language	subtask	Unit	Application	Copora	Period
6	E, J, TC	opinionated	sentence	IR	Mainichi, Yomiuri	1998-2001
		relevance			CIRB	
		polarity			Xinhua English	
		holder			Honkong Standard	
7	+SC	+target	opinion clause	Q&A	+Xinhua Chinese	-
8	-	+CLQA	-	Opinion Q&A	+NYT, UDN	2002-2005

## Task Design in NTCIR-8 MOAT

- Five conventional subtasks  
+ one new subtask

Subtask	Value	Annotation Unit
Opinionated Sentence	YES, NO	Sentence
Relevant Sentence	YES, NO	
Opinionated Polarities	POS, NEG, NEU	Opinion Clause
Opinion Holders	String, multiple	
Opinion Targets	String, multiple	
Cross-lingual Opinion Q&A	YES, NO	Sentence

## Research Question in NTCIR8 MOAT

- Application: estimate the opinion from different culture.
  - New subtask: *cross-lingual opinion Q&A*
- Clarify the effective approach across languages.
  1. Rich lexicon resource
  2. Accomplished feature selection
  3. Hot machine learning technique

### Changes in Task setting

- Common Framework for Annotation across Language
- Reduced # of Annotators, but increased the agreements
- English docs Xinhua English -> NYT



## Corpus: Sources and Topics

Language	Sources	Span	Topics (Opinion Questions)		
			Sum	Sample	Test
T-Chinese	United Daily News	2002-2005	21	1	20
Japanese	Mainichi		21	1	20
English	New York Times		21	1	20
S-Chinese	Xinhua English		20	1	19

## Opinion Question List

ID	Opinion Question
N01	What negative prospects were discussed about the Euro when it was introduced in January of 2002?
(N03)	(What reasons were discussed about Bomb Terror in Bali Island in October, 2002?)
N04	What reasons have been given for the Space Shuttle Columbia accident in February, 2002?
N05	What negative comments were discussed about Bush's decision to start Iraq war in March, 2003?
N06	What negative prospects and opinions were discussed about SARS which started spreading in March, 2003?
N07	What reasons are given for the blackout around North America in August, 2003?
N08	What reasons and background information was discussed about the terrorist train bombing that happened in Madrid in March, 2004?
N11	Why did supporters want to elect George W. Bush in the November 2004 American Presidential Election?
N13	What positive comments were discussed to help the victims from earthquake and tsunami in Sumatera, Indonesia in December, 2004?
N14	What objections are given for the US opposition to the Kyoto Protocol that was enacted in February 2005?
N16	What reasons have been given for the anti-Japanese demonstrations that took place in April, 2005 in Peking and Shanghai in China?
N17	In July 2005 there were terrorist bombings in London. What reasons and background were given, and what controversies were discussed?
N18	What actions by President George Bush were criticized in response to Hurricane Katrina's August 2005 landing?
N20	What negative opinions and discussion happened about the Bird Flu that started spreading in October, 2005?
N24	Identify opinions that indicate that Arnold Schwarzenegger is a bad choice to be elected the new governor of California in the October 2003 election.
N26	Find positive opinions about the reaction of Nuclear and Industrial Safety Agency officials to the Mihama nuclear powerplant accident in August 2004.
N27	What were the advantages and disadvantages of the direct flight between Taiwan and Mainland China commercially?
N32	What are good and bad approaches to losing weight?
N36	What are complaints about XIX Olympic Winter Games that were held in and around Salt Lake City, Utah, United States in 2002?
N39	What are the comments about China's first manned space flight which happened successfully in October 2003?
N41	What negative comments were discussed when in April 2004 CBS made public pictures showing cruel U.S. military abuse of Iraqi prisoners of war?

# Online Annotation tool in MOAT

# Participants

- 16 teams submitted 56 runs.
- Half the teams participated in more than two language related tasks.

TeamID	Affiliation	# of Submission Runs				
		EN	SC	TC	JA	CL
BUPT	Beijing University of Posts and Telecommunications		2			
CTL	City University of Hong Kong		1	1		
CityUHK	City University of Hong Kong			3		
cyut	Chaoyang University of Technology			3		
IISR	Yuan Ze University				3	
KAIST	Korea Advanced Institute of Science and Technology	2				
KLELAB	Pohang University of Science and Technology	3		3		
NECLE	NEC Laboratories China	2	2			
NTU	National Taiwan University	2		2		2
OPAL	University of Alicante	3				3
PKUTM	Peking University		3			
PolyU	The Hong Kong Polytechnic University	3	1			
SIGS	Swedish Institute of Computer Science	1				
TUT	Toyoashi University of Technology					3
UNINE	University of Neuchatel	2		1	1	
WIA	The Chinese University of Hong Kong		2	2		
# of teams		8	6	7	3	2
# of runs		26	11	15	7	5

# English Results

Group	RunID	Opinated			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F
UNINE	1	29.44	62.84	40.10	83.68	32.74	47.07	50.29	29.58	37.25
NECLC	bsf	26.50	58.74	36.52						
NECLC	bs1	21.79	78.84	34.14						
NECLC	bs0	25.85	58.11	35.78						
UNINE	2	19.32	81.79	31.26	84.39	36.01	50.48	48.35	37.80	42.43
KLELAB	3	19.68	68.00	30.53						
KAIST	2	19.00	65.26	29.43						
KLELAB	2	17.90	82.00	29.39						
KAIST	1	18.88	64.84	29.24						
NTU	2	17.02	93.68	28.81	77.75	94.02	85.11	49.22	46.23	47.68
NTU	1	16.80	95.47	28.57	77.95	96.06	86.06	52.17	49.94	51.03
KLELAB	1	16.82	95.37	28.60						
OPAL	1	17.99	45.16	25.73	82.05	47.83	60.43	38.13	12.82	19.19
OPAL	2	19.44	44.00	26.97	82.61	5.16	9.71	50.93	12.26	19.76
OPAL	3	19.44	44.00	26.97	76.32	3.94	7.49			
PolyU	1	24.58	21.47	22.92						
SIGS	1	13.87	31.37	19.24						
PolyU	2	19.51	13.37	15.87						

# CLOQA Results

Group	RunID	lang	evaluation type	Answer Extraction		
				P	R	F
NTU	2	TC	Agree	7.80	38.46	9.38
NTU	1	EN	Agree	5.63	45.81	7.65
NTU	2	EN	Agree	4.87	39.99	7.35
OPAL	1	TC	Agree	3.54	56.23	6.34
OPAL	3	TC	Agree	3.42	72.13	6.32
NTU	1	TC	Agree	5.54	39.07	5.99
OPAL	2	TC	Agree	3.35	42.75	5.78
OPAL	3	TC	Non-Agree	15.02	77.68	23.55
NTU	2	TC	Non-Agree	24.59	41.11	23.41
OPAL	1	TC	Non-Agree	14.62	60.47	21.36
OPAL	2	TC	Non-Agree	14.64	49.73	19.57
NTU	1	TC	Non-Agree	20.09	41.19	19.26
NTU	2	EN	Non-Agree	8.46	37.58	11.35
NTU	1	EN	Non-Agree	8.61	41.55	10.86

## Effective approaches

- The following teams attained good results with accomplished feature filtering, hot machine learning, and rich lexicon resources.

TeamID	Lang	Feature Filtering	Machine Learning	Lexicon Resource
UNINE	EN	Z score	logistic regression	SentiWordNet
PKUTM	SC	Iterative classifier	SVM (better than NB, ME, DT)	In House, NTU, and Jun LI's lexicon
CityUHK	TC	Supervised Lexicon	Ensemble	NTUSD, LCPW, LCNW, CPWP, SKPI

## Community QA Pilot Task

Daisuke Ishikawa †, Tetsuya Sakai ‡ and Noriko Kando †  
† National Institute of Informatics  
‡ Microsoft Research Asia

with special thanks to  
Yohei Seki and Kazuko Kuriyama

2010/6/17

NTCIR-8 Workshop Meeting, 2010

50

## Background

- Increase in activity surrounding social media research targeting community-type Q&A sites;
  - Yahoo! Answer
  - Yahoo! Chiebukuro (Japanese equivalent of Yahoo! Answer)  
(<http://chiebukuro.yahoo.co.jp/>)
  - Oshiete! Goo (<http://oshiete.goo.ne.jp/>)
- Identify high-quality content on these sites (Agichtein 2008, etc).

2010/6/17

51

NTCIR-8 Workshop  
Meeting 2010

## Yahoo! Chiebukuro Data

- **Best Answer:**
  - (1) Most convincing and most satisfying answer as the "Best Answer." by the asker
  - (2) The Best Answer can be selected by vote by other users.
- Only the Best Answers selected by the questioner (1) are recorded in the "Yahoo! Chiebukuro" data version 1.0.

### Details of Yahoo Chiebukuro data version 1.0:

- Data range: 2004/4/1 ~ 2005/10/31
- Questions resolved: 3116009 items(about 916 MB)
- Best answer: 3116008 items(about 935 MB)
- Other answer: 10361777 items(about 2.3 GB)
- Category: 285 categories

2010/6/17

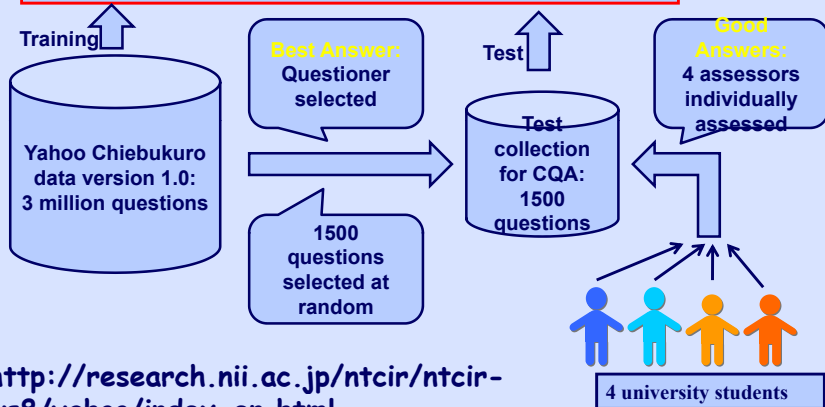
52

NTCIR-8 Workshop  
Meeting 2010



## Community QA Pilot Task

Rank all posted answers by answer quality (as estimated by system) for every question.

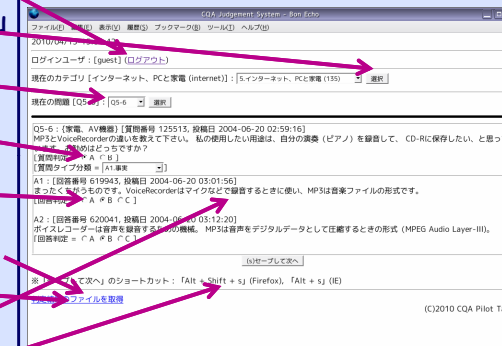


<http://research.nii.ac.jp/ntcir/ntcir-ws8/yahoo/index-en.html>

## Good Answer (GA) Assessment

CQA Online Judgment System Version 0.1  
(modification of MOAT judgment system)

- Login management menu
- Category selection menu
- Question number selection menu
- Question assessment function
- Answer assessment function
- Log preservation function
- Download function of preservation logs
- Display function
- Short cut function



## Good Answer (GA) Assessment Criteria

- Question criterion:
  - Grade A: question is a question.
  - Grade B: question is not actually a question
- Answer criterion:
  - Grade A: satisfactory answer to the question.
  - Grade B: partially relevant answer, or a partially irrelevant answer.
  - Grade C: answer unrelated to the question.

## Participating runs

Table 1: Participating teams.

team name	organisation	#runs
ASURA	National Institute of Informatics	2
BASELINE	Organisers	3
LILY	Shirayuri College	3
MSRA+MSR	Microsoft Research Asia and Redmond	5

- BASELINE-1: rank answers at random
- BASELINE-2: rank answers by length
- BASELINE-3: rank answers by timestamp

## Approaches

- MSRA+MSR (oral presentation)  
SVM-rank / analogical reasoning; relevance, authority, informativeness, discourse/modality
- ASURA (oral presentation)  
SVM; A-1 uses 5 answer features;  
A-2 uses 13 question and answer features
- LILY  
SVM; 10 features including readability but for some reason underperformed random baseline

## BA-Hit@1

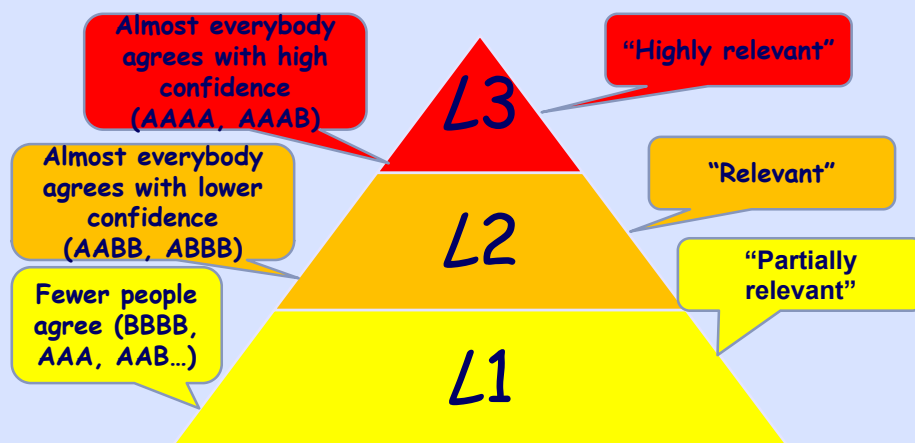
- 1 if top-ranked answer is BA; otherwise 0.

However, asker-selected BAs may be

- Biased (reflects one person's opinion)
- Nonexhaustive (there may be other good answers)

## The Pyramid

[Nenkova et al. 07, Lin/Demner-Fushman 06]



Retain different people's different views in the gold standard

## “Good Answers” (GA) -based metrics

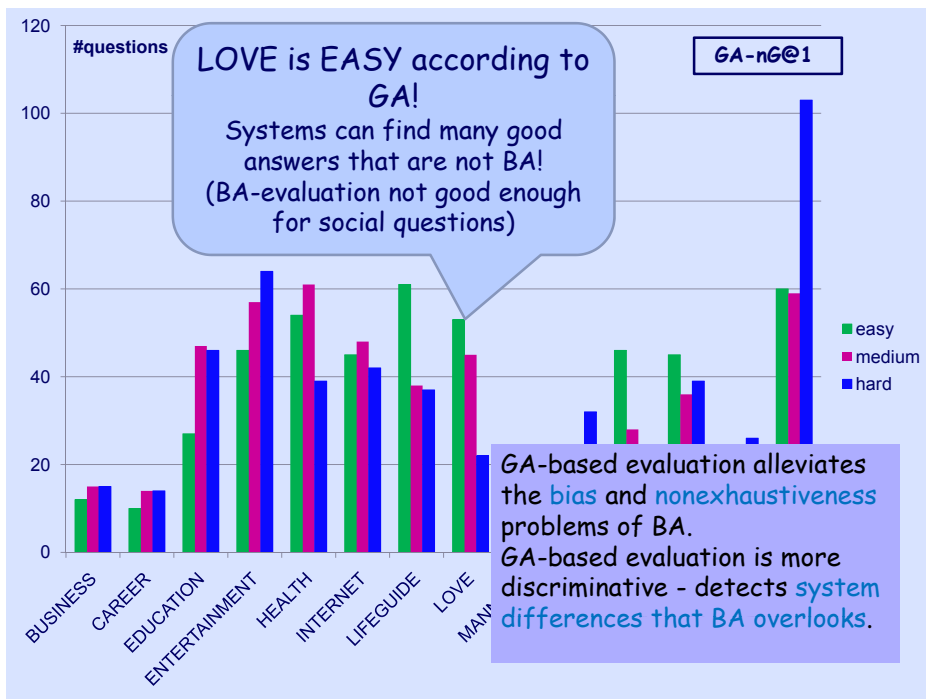
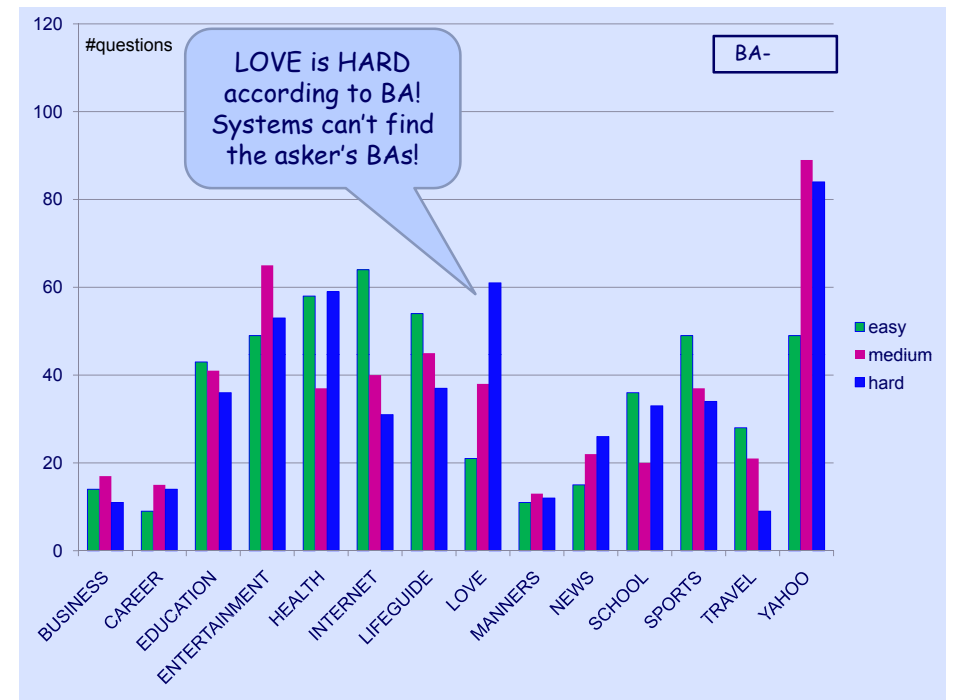
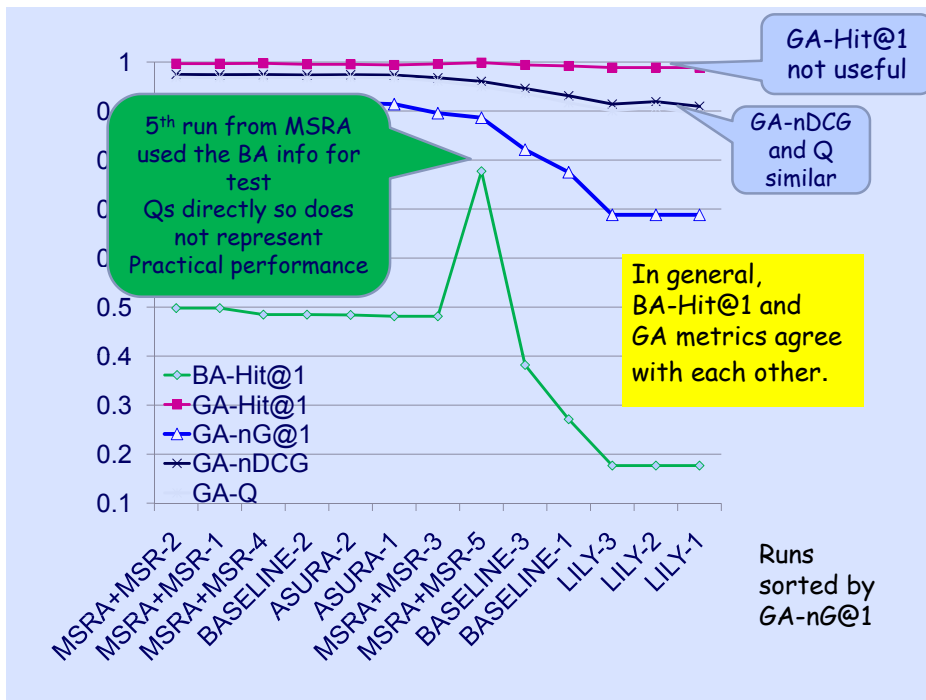
- GA-nG@1 (or nDCG@1)
- GA-nDCG (evaluates the entire answer list)

[Jarvelin/Kekalainen SIGIR00, TOIS02]

- GA-Q (evaluates the entire answer list)

[Sakai AIRS04/05, NTCIR-4, SIGIR06/07, IPM07, EVIA07/08...]

Binary-relevance GA-Hit@1 is not useful:  
almost all answers are at least somewhat relevant



## Patent Mining Task

Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama, and Taiichi Hashimoto

### Background

For a researcher in a field of high industrial relevance, retrieving research papers and patents has become an important aspect of assessing the scope of the field.

### Problem

The terms used in patents are often more abstract or creative than those used in research papers, to widen the scope of the claims.

### Purpose

To develop fundamental techniques for retrieving, classifying, and analyzing both research papers and patents

**Goal: Automatic Creation of technical trend maps from a set of research papers and patents.**

	Effect 1	Effect 2	Effect 3
Technology 1	[AAA 1993] [US Pat. XX-XXX]		[BBB 2002]
Technology 2		[DDD 2001]	
Technology 3	[CCC 2000]	[US Pat. YY-YYY]	[US Pat. ZZ-ZZZ] [US Pat. WW-W]

Research papers and patents are classified in terms of elemental technologies and their effects.

65

## Subtasks in NTCIR-8 PATMN

(Step 1) For a given field, collect research papers and patents written in various languages.

(Step 2) Extract "elemental technologies" and their "effects" from the documents and Classify the documents.

### (Step 1) Subtask 1: Research Paper Classification

Classification of research papers into the International Patent Classification (IPC) system.

### (Step 2) Subtask 2: Technical Trend Map Creation

Extraction of elemental technologies and their effects from research papers and patents.

66

## Evaluation

Subtask 1 (Research Paper Classification)

Metrics: Mean Average Precision (MAP)

- k-NN based approach is superior to machine learning approach.
- Re-ranking of IPC codes is effective.

Subtask 2: Technical Trend Map Creation

Metrics: Recall, Precision, and F-measure

- Top systems employed CRF, and the following features are effective.
  - Dependency structure
  - Document structure
  - Domain adaptation

67

## Patent Translation Task

Atsushi Fujii (Tokyo Tech, Japan)

Masao Utiyama (NICT, Japan)

Mikio Yamamoto (Univ. of Tsukuba, Japan)

Takehito Utsuro (Univ. of Tsukuba, Japan)

Terumasa Ehara (Yamanashi Eiwa Col.)

Hiroshi Echizen-ya (Hokkai-Gakuen Univ.)

Sayori Shimohata (Oki Co., Ltd.)

68

## History of Patent IR at NTCIR

- NTCIR-3 (2001-2002)
    - Technology survey
      - Applied conventional IR problems to patent data
  - NTCIR-4 (2003-2004)
    - Invalidation search
      - Addressed patent-specific IR problems
  - NTCIR-5 (2004-2005)
    - Enlarged invalidity search
  - NTCIR-6 (2006-2007)
    - Added English patents
- \* JPO = Japan Patent Office  
\* USPTO = US Patent & Trademark Office

2 years of JPO patent applications

5 years of JPO

Both document sets were published in 1993-2002

10 years of JPO patent applications

10 years of USPTO patents granted

69

## Patent MT at NTCIR-7 (2007-2008)

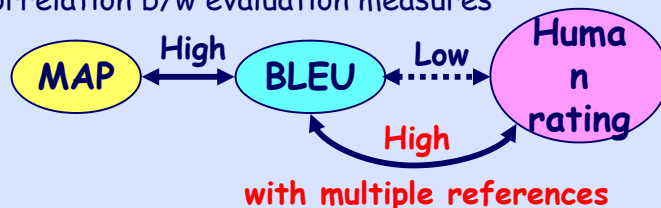
- Patent MT is realistic
  - Parallel corpus
  - Decoders for Statistical MT
- Participants can use any types of MT:
  - Statistical MT (SMT)
  - Rule-based MT (RBMT)
  - Example-based MT (EBMT)
- Utility of patent MT
  - Cross-lingual patent retrieval
    - Good for CLIR
    - Bi-directional CLIR (transl. Q and docs)
    - Need for users' relevance judgments
  - Filing patent applications in foreign countries
    - In Europe, CLIA to CJK Patents are critical and increasing the NEEDS

Important from science, engineering, & industry points of view

70

## Findings at NTCIR-7

- Which method was effective?
  - BLEU: Phrase-based SMT
  - Human rating: Rule-based MT
- Correlation b/w evaluation measures



- MT for regular sentences was effective for translating patent claims

71

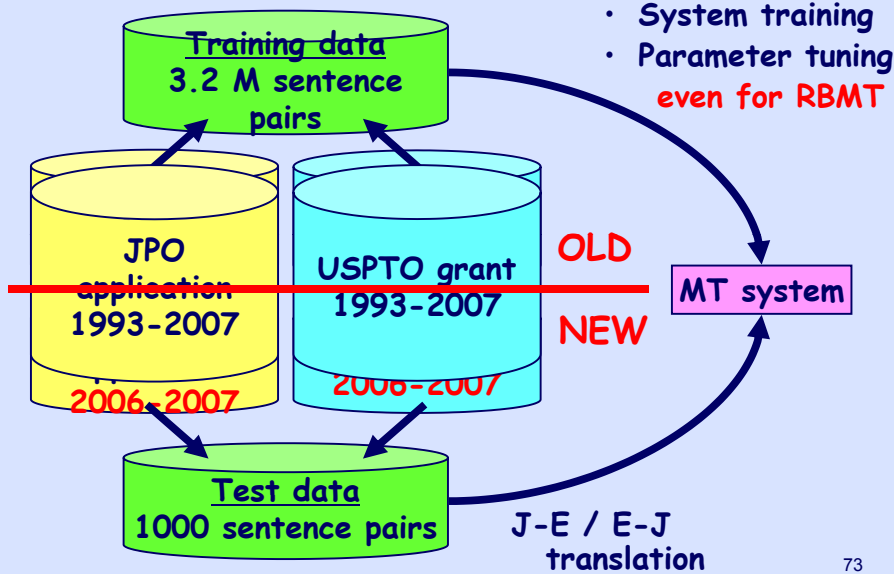
## Patent MT at NTCIR-8

- Larger document sets
  - Japanese/English patent documents (published in 1993-2002, )
- Subtasks
  - Machine translation
  - Cross-lingual IR **Cancelled (no participation)**
    - MT results by other participants can be used
  - Evaluation
    - Developing automatic evaluation methods highly correlated with human rating

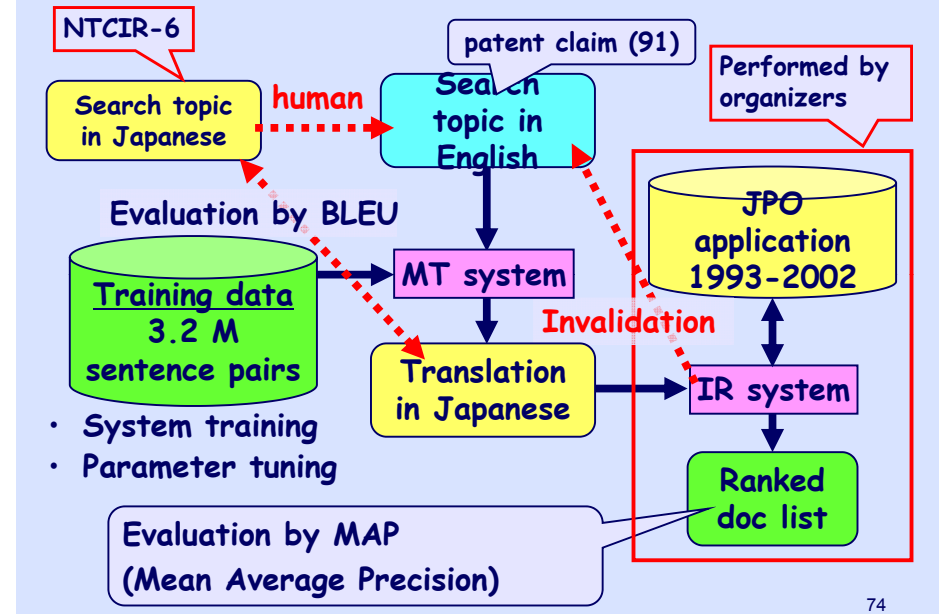
72



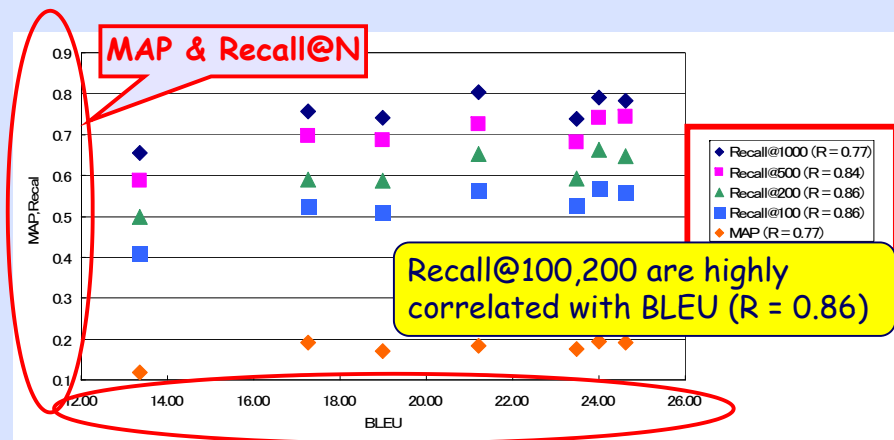
## Intrinsic evaluation



## Extrinsic evaluation



## Extrinsic E-J: BLEU & IR measures



## Overview of the Patent Translation Task Evaluation Subtask (AE subtask)

Terumasa EHARA  
Yamanashi Eiwa College

## Aim of the subtask

- To improve automatic evaluation methods for machine translation accuracy
- To overcome the difference of automatic evaluation and human evaluation [Fujii, 2007]

[Fujii, 2007]:

## Data 1

- The results of the NTCIR-7 patent translation task (only JE direction)
- Training data: NTCIR-7 patent translation task, dry run data
  - Source and reference data: 100 sentences
  - MT output data: 100 (sent.) × 11 (systems)
  - Human evaluation results (adequacy and fluency):  
100 (sent.) × 11 (systems) × 3 (human raters)

## Data 2

- Test data: NTCIR-7 patent translation task, formal run data
  - Source and reference data: 100 sentences
  - MT output data: 100 (sent.) × 12 (systems)
  - Human evaluation result (adequacy and fluency):  
100 (sent.) × 12 (systems) × 3 (raters)

## Result

- AE subtask has only one participant




participant	Correlation coefficients to the adequacy data				Correlation coefficients to the fluency data			
	Pearson		Spearman		Pearson		Spearman	
	Avg	All	Avg	All	Avg	All	Avg	All
HCU-1	0.2992	0.2463	0.2712	0.2234	0.2608	0.2285	0.2486	0.2126

Avg : average value for the correlation coefficients for the 12 test systems

All : correlation coefficient for all data of the 12 test systems

## NTCIR-8 Tasks (2008.07–2009.06)

The 3<sup>rd</sup> Intl W on Evaluating Information Access (EVIA) refereed

- |   |  |
|---|--|
| <b>1. Advanced CL Info Access</b><br>- QA(CsC+JE->CsC+J)    Any Types of Questions  |  |
| GeoTime (E, J)    Geo Temporal Information  |  |
| <b>2. User Generated Contents (CGM)</b><br>- Opinion Analysis (Multilingual News) (Blog)<br>- [Pilot] Community QA (Yahoo! Chiebukuro)  |   |
| <b>3. Focused Domain: Patent</b><br>-Patent Translation; English -> Japanese<br>Statistical MT, The World-Largest training data (J-E sentence alignment), Summer School, Extrinsic eval by CLIR<br>-Patent Mining papers -> IPC<br>-Evaluation of SMT |   |

## NTCIR-9 Test Collections

- Will be available for research purpose to non-participants for free of charge from either of
  - NII NTCIR, or
  - NII Information Research Data Repository (NII-IDR)

NII IDR is a center to collect and release the research data sets for research purpose.

All the NTCIR Test Collections will be transferred to NII-IDR in the near future. Some have already started to release from NII-IDR, ex. NTCIR Web Collection, Yahoo! Chiebukuro

## NTCIR-9

- Call for Formal Task Proposal, soon
- Tasks will be decided by August
- Call for Task Participation
- Call for Paper for EVIA 2011
- NTCIR-9 Meeting & EVIA 2011: Dec. 2011

### Provisional Organization Plan

- NTCIR-9 is co-organized by NICT and NII
- General Co-Chairs; Eiichiro Sumita (NICT), Tsuneaki Kato (U Tokyo), Noriko Kando (NII)
- Evaluation Chairs:
- EVIA Chairs: TBA

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/>

Will be moved to: <http://ntcir.nii.ac.jp/>