



Overview of FIRE 2010

**Mandar Mitra
on behalf of the FIRE team**

**Indian Statistical Institute
Kolkata**

- Background
- Tasks
- Data
- Results
- Problems and prospects
- People

- People have been working on Indian language IR for several years
- Need standard benchmarks
 - to identify what works and what does not
 - to measure progress

- Data
 - document collection
 - query / topic collection
 - relevance judgments - information about which document is relevant to which query
- Platform for comparing results, techniques, models, etc.

■ TREC

- Organized by NIST every year since 1992
- Primary focus on English text

■ CLEF

- Started in 2000 (CLIR track at TREC-6 (1997))
- Focus on European languages

■ NTCIR

- Started in late 1997
- Held every 1.5 years at NII, Japan
- Focus on East Asian languages
(Chinese, Japanese, Korean)

- To encourage research in South Asian language Information Access technologies by providing reusable large-scale test collections for ILIR experiments
- To provide a common evaluation infrastructure for comparing the performance of different IR systems
- To explore new Information Retrieval / Access tasks that arise as our information needs evolve, and new needs emerge
- To investigate evaluation methods for Information Access techniques and methods for constructing a reusable large-scale data set for ILIR experiments.
- To build language resources for IR and related language processing tasks

This is our second year.

- Ad-hoc monolingual retrieval (**repeat**)
 - Bengali, Hindi Marathi and English
- Ad-hoc cross-lingual document retrieval (**repeat**)
 - documents in Bengali, Hindi, Marathi, and English
 - queries in Bengali, Hindi, Marathi, Tamil, Telugu , Gujarati and English
 - Roman transliterations of Bengali and Hindi topics
- Retrieval and classification from mailing lists and forums (**new**)
 - pilot task being offered by IBM India Research Lab
- Ad-hoc Wikipedia-entity retrieval from news documents (**new**)
 - pilot task being offered by Yahoo! Labs, Bangalore.

Ad-hoc monolingual and cross-lingual document retrieval

Training data release	Aug 15 2009	(FIRE 2008 data)
Test data release	Nov 01 2009	(Topics)
Adhoc run submission	Dec 11 2009	
Results release	Feb 01 2010	

Documents

- Bengali: Anandabazar Patrika (123,047 docs)
 - Hindi: Dainik Jagran (95,215 docs) +
Amar Ujala (54,266 docs)
 - Marathi: Maharashtra Times, Sakal (99,275 docs)
 - English: Telegraph (125,586 docs)
-
- All from the Sep 2004 - Sep 2007 period
 - All content converted to UTF-8
 - Minimal markup

Hindi – Dainik Jagran corpus

के लिए स्थान की परेशनायी हो जाएगी।

सूत्रों का कहना है कि टाटा मोटर्स प्रबंधन जमशदपुर प्लांट की क्षमता को बढ़ाने के लिए सरकार के स्तर पर लगातार पहल कर रहा है। टाटा मोटर्स के वैश्विक बाजार में कदम रखने के साथ ही एक श्रेणी के वाहनों का एक ही प्लांट में उत्पादन करने की महती योजना तैयार करने की प्रक्रिया शुरु हो गयी है। इसी क्रम में टाटा मोटर्स जमशदपुर के वर्तमान प्लांट में केवल भारी वाहनों के उत्पादन की योजना पर चर्चा शुरु हो गयी है। पुणे प्लांट में हल्के वाहन तैयार किए जाएंगे। प्रस्तावित प्लान के लागू होने की स्थिति में जमशदपुर प्लांट की उत्पादन क्षमता एक हजार वाहन प्रतिदिन हो जाएगी। वही प्रबंधन के समक्ष सबसे बड़ी परेशानी यह है कि वर्तमान में ही उत्पादित वाहनों को रखने के लिए पार्किंग का स्थान कम पड़ रहा है। हाल ही में टाटा कमिंस के सामने नया पार्किंग स्थल तैयार किया गया है। एक हजार वाहन उत्पादन की स्थिति में प्रबंधन की ओर से सरकारी स्तर पर वार्ता शुरु कर दी गयी है। भारी वाहन के साथ ही मुंबई मुख्यालय पिछले दिनों महाराष्ट्र व दक्षिण के एक प्रदेश में लांच किए गए टाटा के एक टन वजन के नए वाहनों को पूर्वोत्तर भारत में एक साथ लांच करने की योजना है। एक टन वाले उक्त वाहन का आकार विक्रम टेम्पो के समान होगा जो वाहन बाजार में टेम्पो श्रेणी के वाहनों को कड़ी चुनौती देगा। दक्षिण के राज्यों में उक्त दोनों ही वाहनों की मांग में काफी तेजी से वृद्धि

Topics

- 50 topics (numbers 76-125)
- Queries formulated parallelly in Bengali, Hindi by browsing the corpus
- Refined based on initial retrieval results
 - ensure minimum number of relevant documents per query
 - balance easy, medium and hard queries
- Translated into Marathi, Tamil, Telugu , Gujarati and English
- TREC format (title + desc + narr)

Relevance assessments

- Preliminary pooling using TERRIER
- Pool from submissions
 - pool depth = 60
- Interactive search
 - aim: find as many relevant documents as possible
 - tools: boolean filters, relevance feedback, supervised query expansion
 - limit: look at about 100 documents

Pool size across queries

	Bengali	Hindi	Marathi	English
Minimum	96	280	233	87
Maximum	300	704	616	553
Total	8,655	22,572	20,761 +	15,135

Relevance assessments

Number of relevant documents

	Bengali	Hindi	Marathi	English
Minimum	2	2	0 (11)	1
Maximum	29	74	72	47
Mean	10	18	12	13
Median	8	14	6	11
Total	510	915	621	653
FIRE 2008	1863	3436	1095	3779

Queries with 5 or more rel. docs.

	Bengali	Hindi	Marathi	English
# queries	40	45	26	42

Participants

Institute	Country	# runs submitted
AU-KBC	India	2
Dublin City U.	Ireland	17
IBM	India	2
IIT Bombay (1)	India	30
IIT Bombay (2)	India	3
Jadavpur U.	India	2
MANIT	India	9
Microsoft Research	India	32
U. Neuchatel	Switzerland	18
U. North Texas	USA	8
U. Tampere	Finland	6
11 (9 @ FIRE 2008)	TOTAL	129 (up from 64 @ FIRE 2008)

Submissions

Query language	Docs retrieved	# runs
Bengali	Bengali	16
Hindi	Hindi	19
Marathi	Marathi	20
English	English	15
English	Bengali	2
English	Hindi	18
Hindi	English	21
Marathi	English	4
Tamil	English	14

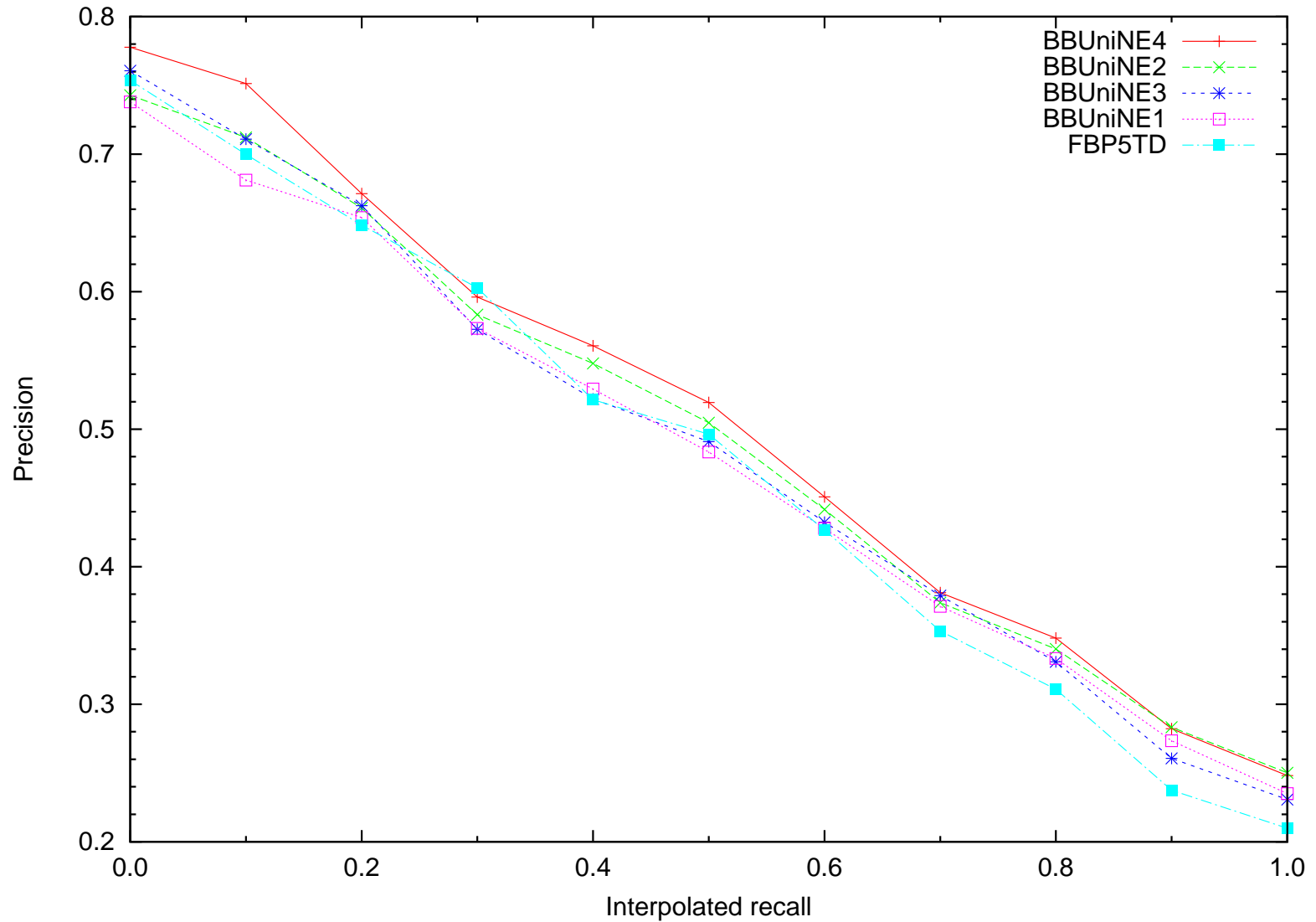
Results

Mono-lingual retrieval (16 runs)

TD runs		
RunID	Group	MAP
BBUniNE4	UniNE	0.4862
BBUniNE2	UniNE	0.4731
BBUniNE3	UniNE	0.4684
BBUniNE1	UniNE	0.4646
FBP5TD	DCU	0.4526

Best from FIRE 2008: 0.4719

Best EN → BN run: 0.3771 (about 78% of best mono)



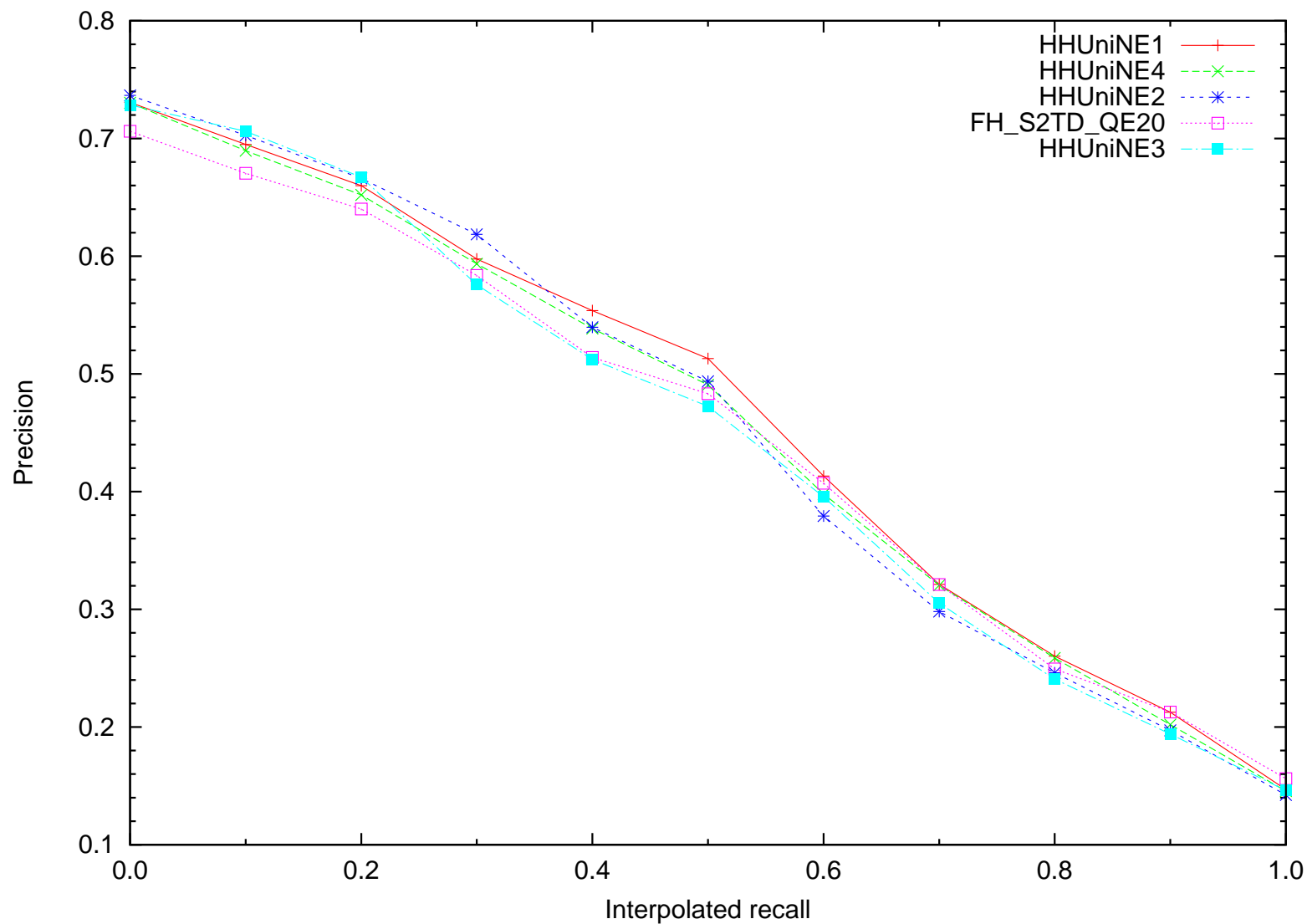
Mono-lingual retrieval (16 runs)

All runs			
RunID	Group	Fields	MAP
BBUniNE6	UniNE	TDN	0.5438
BBUniNE5	UniNE	TDN	0.5329
tdnLemurStale	UTA	TDN	0.5058
BBUniNE4	UniNE	TD	0.4862
BBUniNE2	UniNE	TD	0.4731

Mono-lingual retrieval (19 runs)

TD runs		
RunID	Group	MAP
HHUniNE1	UniNE	0.4459
HHUniNE4	UniNE	0.4373
HHUniNE2	UniNE	0.4334
FH_S2TD_QE20	DCU	0.4305
HHUniNE3	UniNE	0.4284

Best from FIRE 2008: 0.3487



Mono-lingual retrieval (19 runs)

All runs			
RunID	Group	Fields	MAP
HHUniNE6	UniNE	TDN	0.4864
HHUniNE5	UniNE	TDN	0.4769
HHUniNE1	UniNE	TD	0.4459
HHUniNE4	UniNE	TD	0.4373
HHUniNE2	UniNE	TD	0.4334

Cross-lingual retrieval (EN → HI, 18 runs)

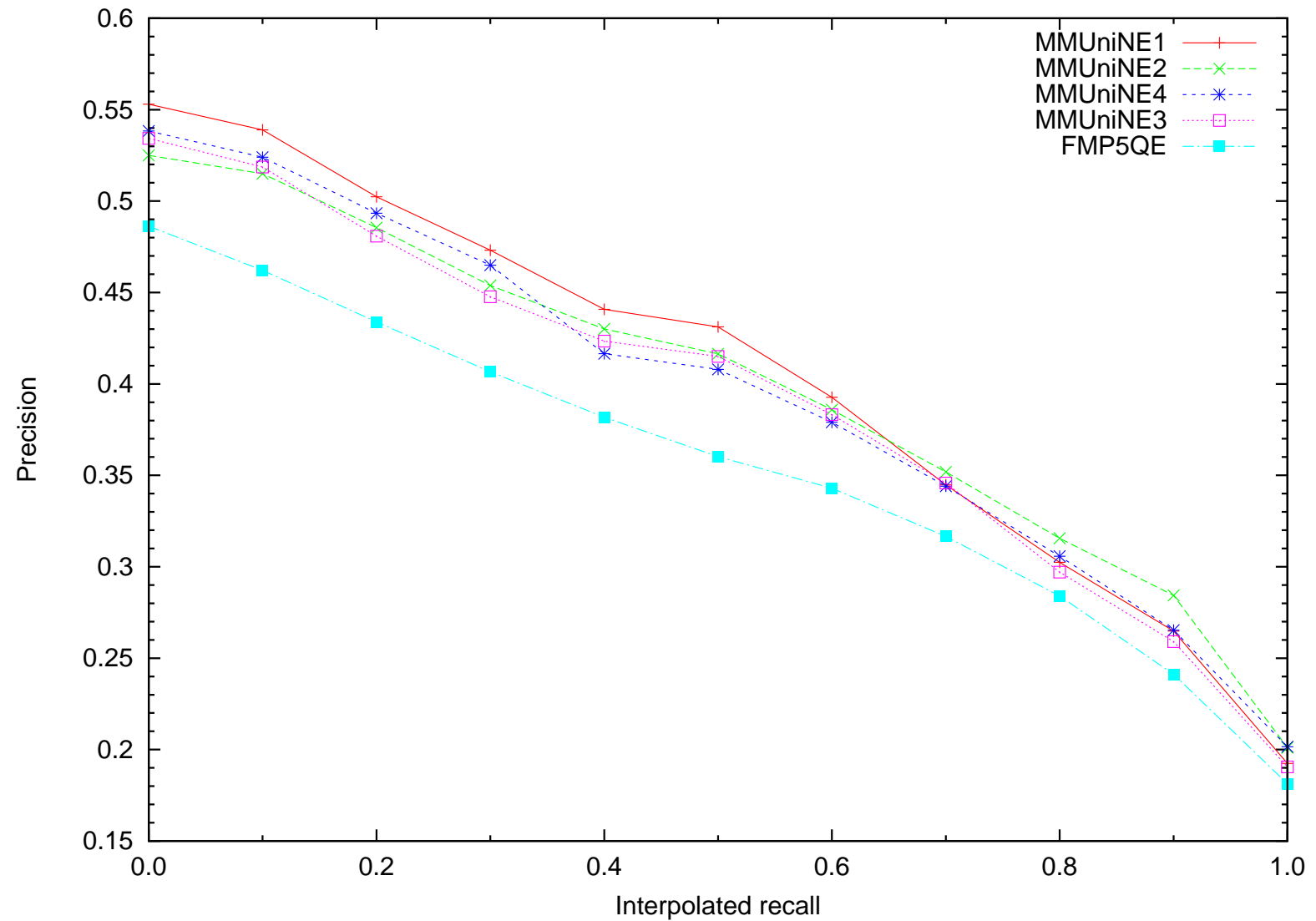
RunID	Group	Fields	MAP
FHan_P5TD_QE20	DCU	TD	0.3771
UNTclenhi	UNT	TD	0.3757
FHan_S2TD_QE20	DCU	TD	0.3747
FHgt_S2TD_QE20	DCU	TD	0.3684
FHgt_P5TD_QE20	DCU	TD	0.3647

Best mono-lingual run: 0.4459

Mono-lingual retrieval (20 runs)

TD runs		
RunID	Group	MAP
MMUniNE1	UniNE	0.5009
MMUniNE2	UniNE	0.4897
MMUniNE4	UniNE	0.4885
MMUniNE3	UniNE	0.4817
FMP5QE	DCU	0.4373

Best from FIRE 2008: 0.4483



Cross-lingual retrieval ($X \rightarrow$ EN, 18 runs)

X	RunID	Group	MAP
HI	2010FIREHE110	MSRI	0.4376
HI	2010FIREHE112	MSRI	0.4375
HI	2010FIREHE102	MSRI	0.4369
HI	2010FIREHE101	MSRI	0.4336
HI	2010FIREHE100	MSRI	0.4042
TA	2	AUKBC	0.3980
MR	MRENFEEDBACKT50	IITBCFILT	0.2771*

Best mono-lingual run: 0.4846* (FES1QE, DCU)

* – Title only

- IR models: vector space, BM25, **DFR**, LM (Hiemstra)
- **Stemming**
 - light – removal of inflectional suffixes attached to nouns and adjectives (verbs ignored)
 - aggressive – also removes some frequently used derivational suffixes
- Pseudo-relevance feedback
- Score fusion

Problems and prospects

- Quality of queries
- Diversity and quality of pool
- Wider participation
- New tasks, languages
- More after the Steering Committee meeting

There will be a next time.



Steering committee

James Allan

Hwee Tou Ng

Ricardo Baeza-Yates

Iadh Ounis

Hsin-Hsi Chen

Carol Peters

Tat-Seng Chua

Doug Oard

Christian Fluhr

Prabhakar Raghavan

Norbert Fuhr

Stephen Robertson

Donna Harman

Tetsuya Sakai

Gareth Jones

Mark Sanderson

Noriko Kando

Jacques Savoy

Krishna Kumnamuru

Fabrizio Sebastiani

Mun Kew Leong

Amit Singhal

Ee Peng Lim

Ian Soboroff

Paul McNamee

Tony Veale

Sung Hyon Myaeng

Ellen Voorhees

Ad-hoc retrieval:

Pushpak Bhattacharyya (pb@cse.iitb.ac.in)

Dipasree Pal (dipasree_t@isical.ac.in)

Retrieval from mailing lists and discussion forums:

Debapriyo Majumdar (debapriyo at in dot ibm dot com)

Ayan Bandyopadhyay (ayan_t@isical.ac.in)

Ad-hoc Wikipedia-entity retrieval from news documents:

Ashwin Tengli (ashwint@yahoo-inc.com)

Pabitra Mitra (pabitra@cse.iitkgp.ernet.in)



Thank you!

- Donna Harman, Ellen Voorhees, Carol Peters, Noriko Kando, Doug Oard, Mark Sanderson, and other members of our steering committee
- Anandabazar Patrika, Jagran, Amar Ujala
- Assessors, participants, and speakers
- Staff and students of DA-IICT
- Sponsors: Google, HP, Yahoo, SNLTR, and DIT, Govt. of India
- And many more . . .