Overview of FIRE 2012

Prasenjit Majumder
on behalf of the FIRE team

DAIICT

Gandhinagar

Overview

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

Background

- 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

Evaluation fora

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

The big ones

- **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)

FIRE: goals

- 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 third year.

Tasks

- Ad-hoc monolingual / cross-lingual retrieval
  - documents in Bengali, Gujarati, Hindi, Marathi, Tamil and English
  - queries in Bengali, Gujarati, Hindi, Marathi, Tamil, Telugu and English
- SMS-based FAQ Retrieval
- Cross-Language Indian Text Reuse (CL!TR)
- Personalised IR (PIR)
- Retrieval from Indic Script OCRed Text (RISOT)
- WSD for IR
- Adhoc Retrieval from Mailing Lists and Forums (MLAF) — scrapped

Timeline

Ad-hoc monolingual and cross-lingual document retrieval

Corpus Release	Aug 01 2011
Query Release	Aug 16 2011
Run Submission	Sep 01 2011
	Sep 15 2011
Qrel Release	Nov 15 2011
Working Note Due	Nov 28 2011

Datasets

Documents

Lang.	Source	# docs.	Size (GB)	Remarks
Bengali	Anandabazar Patrika (IN)	374,203	3.0	Expanded
	BDNews24 (BD)	83,167	0.5	New
Gujarati	Gujarat Samachar	313,163	2.7	New
Hindi	Amar Ujala	54,266	0.2	DJ droppe
	Navbharat times	331,599	1.7	New
Marathi	Maharashtra Times, Sakal	99,275	0.7	
Tamil	Dinamalar	194,483	1.0	New
English	Telegraph (IN)	303,291	1.4	Expanded
	BDNews24 (BD)	89,286	0.4	New
Odia		32871	0.2	new

- All content converted to UTF-8
- Minimal markup

Datasets

Topics

- 50 topics (numbers 126-175) in TREC format (title + desc + narr)
- Queries formulated parallely 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 manually into other languages

Relevance assessments

- Preliminary pooling using TERRIER
- Pool from submissions
  - pool depth = 200 (ben), 100 (hin), 100 (eng), 30 (guj), 30 (odia)
- 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	English	Gujarati	Odia
Minimum	436	594	474	36	49
Maximum	1258	1070	1073	89	1000
Total	44823	39827	36499	3127	45236

Relevance assessments

Number of relevant documents

	Bengali	Hindi	English	Gujarati	Odia
Minimum	7	5	2	1	1
Maximum	194	236	294	55	49
Mean	51.62	46.18	70.78	12.52	9
Median	39.5	39.5	53.5	9	3
Total	2581	2309	3539	576	189
FIRE 2011	2778	4084	2761	1659	-
FIRE 2010	510	915	653	-	-
FIRE 2008	1863	3436	3779	-	-

Queries with 5 or more rel. docs.

	Bengali	Hindi	English	Gujarati	Odia
# queries	49	50	49	34	8

Participants

Institute	Country	# runs submitted
DCU	Ireland	12
Gujarat University	India	15
IIT, Kharagpur	India	10
ISM, Dhanbad	India	8
ITER, Bhubaneswar	India	1
JU	India	1

Year	**# teams**	**# runs**
2008	9	64
2010	11	129
2011	7	73
2012	7	47

Submissions

Query language	Docs retrieved	# runs
Bengali	Bengali	14 (4 unofficial)
Hindi	Hindi	0 (4 unofficial)
Marathi	Marathi	18
English	English	2
Gujarati	Gujarati	0 (7 unofficial)
Bengali	Hindi	0 (4 unofficial)
Bengali	Gujarati	0 (4 unofficial)
Gujarati	Bengali	0 (4 unofficial)
Gujarati	Hindi	0 (4 unofficial)
Hindi	Bengali	0 (4 unofficial)
Hindi	Gujarati	0 (4 unofficial)

Results

Bengali (15 runs, 3 groups)

Baseline (Model: In_expC2, Stemmer: Yass)

Title Only: 0.2142

Title Desc: 0.2887

Best Mono (T) : 0.0505

Best Cross (T) (E-B) : 0.0111

Best (TD) from FIRE 2011: 0.3798

Best (TD) from FIRE 2010: 0.4862

Best (TD) from FIRE 2008: 0.4719

Hindi (8 runs, 2 groups)

Baseline (Model: In_expC2, Stemmer: Yass)

Title Only: 0.2485

Title Desc: 0.3391

Best Mono (T) : 0.0747

Best Cross (T) (B-H): 0.3206

Best (TD) from FIRE 2010: 0.4459

Best (TD) from FIRE 2008: 0.3487

Gujarati (15 runs, 1 group)

Best Mono (T) : 0.3378
Best Mono (TD) : 0.4015

Odia (1 run)

Best Mono (TD) : 0.1513

English (8 runs, 2 groups)

Baseline (Model: In_expC2, Stemmer: Yass)

Title Only: 0.3428

Title Desc: 0.4218

Best Mono (TDN) : 0.2264

Best Cross (T) (B-E) : 0.1325

Best (T) from FIRE 2010: 0.4846

Best (TD) from FIRE 2008: 0.5572

Bengali

Marathi

Mono-lingual retrieval (18 runs)

TD runs

RunID	Group	MAP
qListDFR_IneC2-c1d5-NNN.trec_2	UniNE	0.2350
qListOkapi-b0d75k1d2-NPN.trec_2	UniNE	0.2318
fcg-80	ISI and UTA	0.2223
qListDFR_PB2-c1d5-NNN.trec	UniNE	0.2222
qListDFR_PB2-c1d5-NNN.trec_3	UniNE	0.2222

Best from FIRE 2010 0.5009
Best from FIRE 2008: 0.4483

Problems and prospects

- Wider participation
- New tasks, languages
- More after the Steering Committee meeting

There will be a next time.

Steering committee

James Allan
Ricardo Baeza-Yates
Hsin-Hsi Chen
Tat-Seng Chua
Christian Fluhr
Norbert Fuhr
Donna Harman
Gareth Jones
Noriko Kando
Krishna Kummamuru
Mun Kew Leong
Ee Peng Lim
Paul McNamee
Sung Hyon Myaeng
Hwee Tou Ng
Iadh Ounis
Carol Peters
Doug Oard
Prabhakar Raghavan
Stephen Robertson
Tetsuya Sakai
Mark Sanderson
Jacques Savoy
Fabrizio Sebastiani
Amit Singhal
Ian Soboroff
Tony Veale
Ellen Voorhees

Thank you!

- Members of our steering committee
- Anandabazar Patrika, Amar Ujala, etc.
- Assessors, participants, and speakers
- Sponsors: Google, Microsoft Research, SNLTR, and DIT, Govt. of India
- And many more . . .