Forum for Information Retrieval Evaluation

Prasenjit Majumder
DAIICT, Gandhinagar, India
Dipasree Pal
ISI, Kolkata, India
(currently at U. Tampere, Finland)

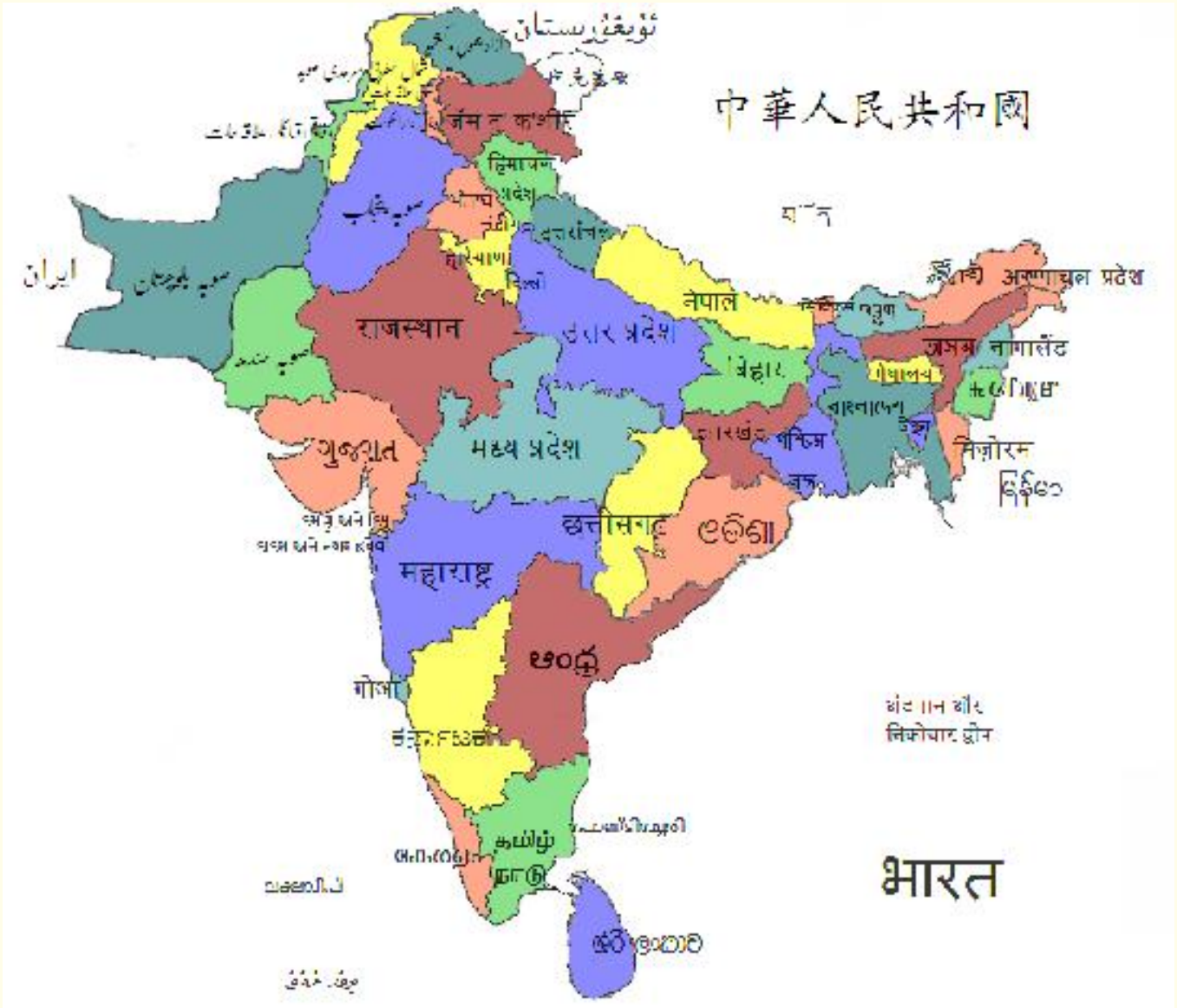
On behalf of the
FIRE Team

September 21, 2010
www.isical.ac.in/~fire

Overview

- Background
- History
- Tasks
- Participants
- Techniques
- Problems

Lingual Diversity



Background

Lingual diversity of Indian sub-continent

- Countries: India, Pakistan, Bangladesh, Sri Lanka, Nepal, Bhutan.
- Population: about 1,300 million.
- Official languages: about 25.
- Many spoken languages, with different scripts, different dialects.
- Hindi and Bengali rank among the top ten most-spoken languages of the world.
- Web content in IL substantial (growth rate 700%)

History

Year	Tasks
CLEF2007	Translate queries
FIRE2008	Adhoc
FIRE2010	Adhoc, MLaF, WikEND
FIRE2011	Adhoc, MLaF, WikEND, SMS*, ...

*SMS Task — proposed by IBM

Adhoc task

- Tasks: monolingual and cross-lingual retrieval
- Documents: news articles from Sept '04 to Sept '07

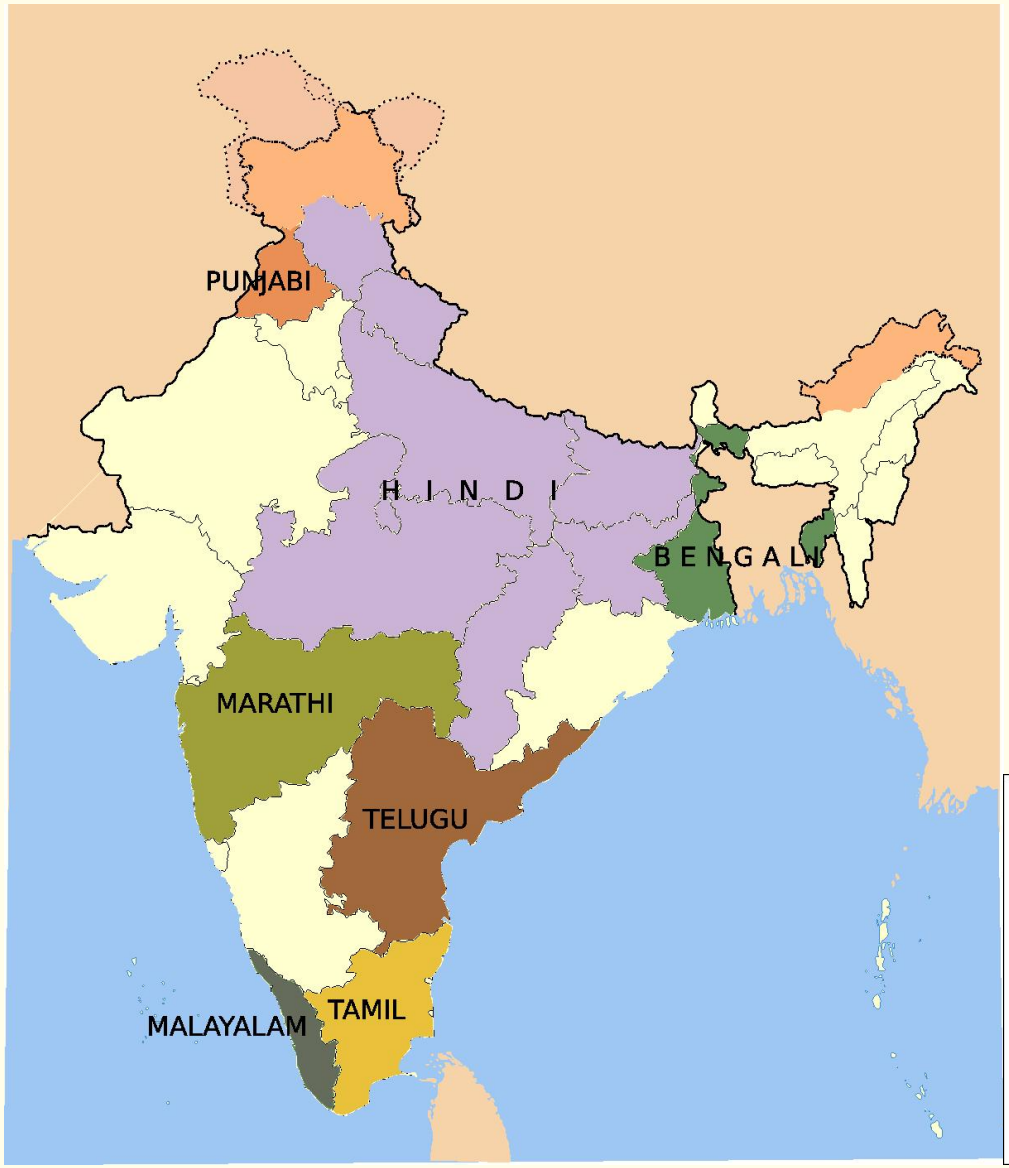
Language	Corpus Source	Corpus size	# docs
Bengali	Ananda Bazar patrika	768M	123,047
Hindi	Jagran	753M	95,215
Hindi	Amar Ujala	356M	54,269
English	The Telegraph	333M	125,638
Marathi	Maharastra Times	511M	99,926

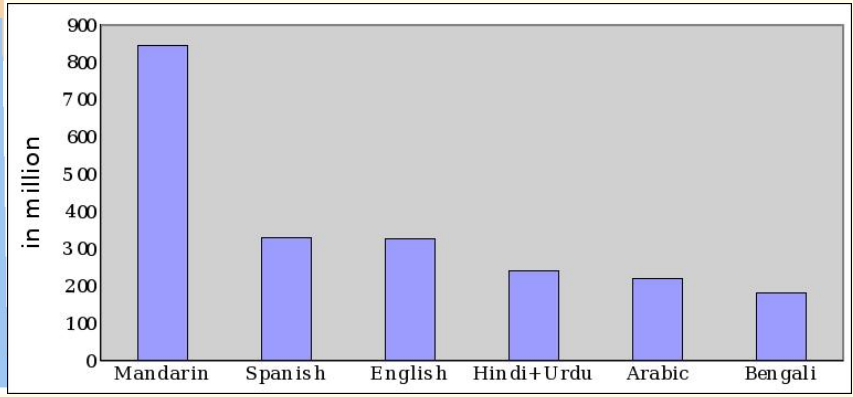
- Queries: Bengali, Hindi, Marathi, English + some others languages

FIRE	Training	Test	Languages
2008	1 - 25	26 - 75	Tamil, Telugu, Malayalam, Punjabi
2010	26 - 75	76 - 125	Gujarati

+ Transliterated queries (Bengali, Hindi queries in Roman script)

FIRE 2008 adhoc task





Mailing List and Forum Task

Message threads of technical problems.
Two sub-tasks:

- Ad-hoc retrieval (25 queries)
- Classification of messages (304 msg id)
  (ASK_QUESTION, DITTO, ASK_CLARIFICATION, FURTHER_DETAILS, SUGGEST_SOLUTION, SOLUTION_FEEDBACK_NEG, SOLUTION_FEEDBACK_POS)

Documents: Messages before June 2009

Corpus Source	Corpus size	# docs
Ubuntu-users archive	769M	187293
Tugindia archives	20M	4833
Tech forums	133M	16996
Tech support forums	34M	3010

WikEND

- "Query": page from Yahoo News
- To find: Wikipedia pages related to entities mentioned on the query page

President Raul Castro said on Saturday he would not change Cuba's communist .....Cuban National Assembly, Castro acknowledged the United States under President Barack Obama was less "aggressive" .....Secretary of State Hillary Clinton for saying repeatedly that Washington expected Havana to make ..... succeeded his brother Fidel Castro as president last year. .....are members of the Communist Party . "We are ready to talk about everything, but ... a news ticker on the U.S. Interests Section in Havana that Cuba viewed ..... and political prisoners . ..... battered by the global financial crisis ..... (Editing by Peter Cooney)

Relevant Wiki Entities:
Raul Castro- http://en.wikipedia.org/wiki/Raul_Castro
Barack Obama- http://en.wikipedia.org/wiki/Barack_Obama
Hillary Clinton- http://en.wikipedia.org/wiki/Hillary_Rodham_Clinton
Washington- http://en.wikipedia.org/wiki/Washington
Havana- http://en.wikipedia.org/wiki/Havana
Fidel Castro- http://en.wikipedia.org/wiki/Fidel_Castro
Communist Party- http://en.wikipedia.org/wiki/Communist_party
Political prisoners- http://en.wikipedia.org/wiki/Political_prisoner
Global financial crisis-
http://en.wikipedia.org/wiki/Global_financial_crisis_of_2008

WikEND

Dataset

- Source: Wikipedia dump (as on 9th Sept, '09)
- Size: 5.2 GB (bz2) and 24 GB (uncompressed)
- Queries: 25 Yahoo News articles
- Submissions: 100 results per topic

Participants

Run submission details.

Institute	Country	2008 #groups 9 #runs 64	2010 #groups 11 #runs 129
AU-KBC	India	2	2
IIT Bombay (1)	India	1	30
IIT Bombay (2)	India	10	2
Microsoft Research	India	5	32
U. Neuchatel	Switzerland	14	18
ISI Kolkata	India	5	
Johns Hopkins U.	USA	12	
IIIT Hyderabad	India	12	
U. Maryland	USA	3	
Dublin City U.	Ireland		17
IBM	India		2
Jadavpur U.	India		2
MANIT	India		9
U. North Texas	USA		8
U. Tampere	Finland		6

Techniques tried (Adhoc)

- Tokenization

  Lemmatizer, Morphological analyser, Stemming, N-gram tokenization, Zonal indexing, Term conflation.

- CLIR

  Bilingual dictionary, Transliteration (only nouns), probabilistic transliteration using parallel corpora, Google transliteration.

- Other

  Language model, QE (wordnet), Data fusion.

Publications

Special Issue of the ACM Transactions on Asian Language Information Processing (TALIP) on Information Retrieval for Indian Languages

- Volume 9 , Issue 3 (September 2010)
- Volume 9 , Issue 4 (December 2010)

Guest editors:

Donna Harman,
Noriko Kando,
Prasenjit Majumder,
Mandar Mitra,
Carol Peters.

Problems/Challenges/Prospects

- Wider participation
- Diversity / quality of pool
- New tasks, languages
- FIRE winter school 2010
  - Stimulating interest of researchers
  - Basic idea of handling IL issues in IR
  - Task proposals

Information Retrieval Society of India

http://www.irsi.res.in

THANK YOU!

- DIT, Govt. of India
- Workshop sponsors: Google, HP, MSR, Yahoo, IBM, SNLTR
- Anandabazar Patrika, Jagran, Amar Ujala, Maharastra Times, The Telegraph
- Donna Harman, Ellen Voorhees, Carol Peters, Noriko Kando
- Doug Oard, Stephen Robertson, Mark Sanderson, Amit Singhal
- FIRE steering committee
- Fellow members of the CLIA consortium
- And many more ...