FIRE: A Community building exercise

Prasenjit Majumder
DAIICT, India
On behalf of FIRE

FIRE: Goals

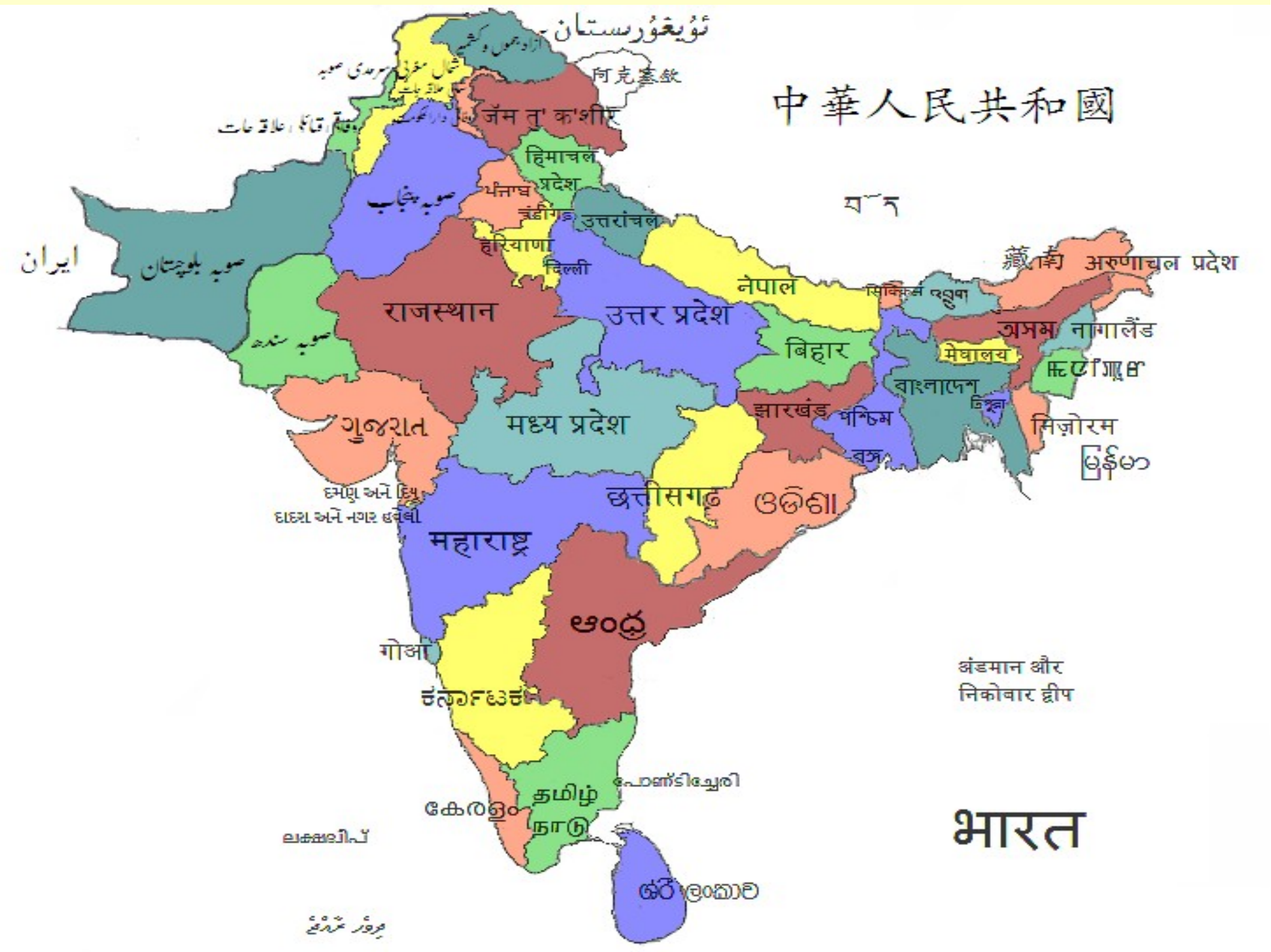
- To encourage research in South Asian language Information Access technologies by providing reusable large-scale test collections for ILIR experiments
- To provide a common evaluation infrastructure for comparing the performance of different IR systems
- To explore new Information Retrieval / Access 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

In short

Community Building

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%)


中華人民共和國
भारत
ایران
ئۇيغۇرستان
阿克塞哈
صوبہ بلوچستان
صوبہ پنجاب
صوبہ سندھ
جمّ تٗ ک'شیٖر
ہماچل پردیش
پنجاب
چنڈیگڑھ
اتّراکھنڈ
ہریانہ
دہلی
راجستھان
اتر پردیش
بہار
نیپال
सिक्किम
अरुणाचल प्रदेश
असम
नागालैंड
मेघालय
বাংলাদেশ
ত্রিপুরা
মিজোরম
ঝাড়খণ্ড
পশ্চিম বঙ্গ
ગુજરાત
મધ્ય પ્રદેશ
છત્તીસગઢ
ଓଡ଼ିଶା
દમણ અને દિવ
દાદરા અને નગર હવેલી
महाराष्ट्र
ಆಂಧ್ರ
ಗೋವಾ
ಕರ್ನಾಟಕ
ಪೊಣ್ಡಿಚ್ಚೇರಿ
தமிழ் நாடு
കേരളം
ലക്ഷദ്വീപ്
ශ්‍රී ලංකාව
अंडमान और निकोबार द्वीप

+ve

- Funding
- New groups
- Companies

Community

Consortia:

1. CLIA (12 Institutes)
2. MT (10 Institutes)
3. OCR (10 Institutes)

Conferences:

COLING 2012

CCLING

Challenges

- “idea” of Evaluation
- NLP and IR
- No national level conference (should FIRE fill the gap?)
- No journal/periodical in Indian Languages.

Suggested Measures

1. To held periodical workshops.
2. To organize college level competitions.
3. To attract non-active participants.
4. identify the area of interest.

   Adhoc, Adhoc with MET

Ad-hoc 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 dropped
	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

- All content converted to UTF-8
- Minimal markup

Ad-hoc Datasets

- 225 topics (numbers 1-225) 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

Ad-hoc participation

Year	#Teams	#Runs
2008	9	64
2010	11	129
2011	7	73
2012	15	~200

Tasks

CLEF2007	Translate queries
FIRE2008	Adhoc (Hindi, Bangla, Marathi)
FIRE2010	Adhoc (Hindi, Bangla, Marathi)
FIRE2011	Adhoc (Hindi, Bangla, Gujarati, Marathi, Tamil) CL!TR, RISOT, SMS-based FAQ Retrieval
FIRE2012	Adhoc (Hindi, Bangla, Gujarati, Marathi, Tamil, Odia) CL!NSS (new avatar of CL!TR) MET RISOT SMS-based FAQ Retrieval CIR

Future

- New proposals:
  - Machine Translation
  - Spoken Documents
- Shared task
- Looking for Collaborative tracks

Thanks!