Forum for Information Retrieval Evaluation

Information Retrieval Laboratory, Indian Statistical Institute, Kolkata

FIRE

A TREC-like platform for Indian languages, providing test collections and a common forum for comparing models and techniques. FIRE is funded primarily by the Technology Development in Indian Languages (TDIL) group, housed within the Department of Information Technology, Government of India.

Goals

- Construct reusable large-scale test collections in 23 Indian languages.
- Provide a common evaluation infrastructure for comparing the performance of different IR systems.
- Explore new IR tasks.

Participation

Year	*Groups*	*Runs*
2008	9	64
2010	11	129
2011	7	73

Tasks

Task name	*2008*	*2010*	*2011*	*2012*
Adhoc (Mono-lingual and cross-lingual)	○	○	○	○
Retrieval and Classification from Mailing Lists and Forums		○		
Adhoc Wikipedia Entity Retrieval from News Documents		○		
Cross-language Indian Text Reuse			○	
Retrieval from Indic Script OCR'd Text			○	○
SMS-based FAQ Retrieval			○	○
Cross-language Indian News Story Search				○
Morpheme Extraction Task				○
Collaborative Information Retreival				○

Notes

- Interest in Adhoc task is on the wane.
- Mandate for creating more Adhoc test collections remain valid. 17 more languages to cover.
- Bias inherent in the query formulation process. Left to one institute, topics chosen are biased towards Bengali and Hindi Corpora.
- Efforts to increase participation.

Adhoc Task (Mono-lingual and Cross-lingual)

Figure 1: 2008

Figure 2: 2010

Figure 3: 2011

On the left are mono-lingual Adhoc runs by language. On the right are the cross-lingual runs.

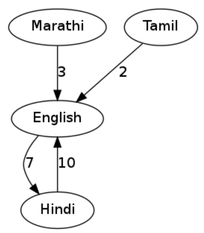

Figure 4: 2008

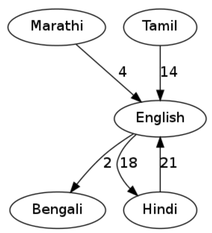

Figure 5: 2010

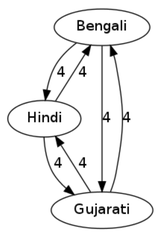

Figure 6: 2011

Topics

- 225 topics formulated in English so far.
- Assessors formulated topics by manually searching the corpus.
- Topics refined based on the number of relevant documents retrieved, balance of easy, medium and hard queries.
- Manually translated into 6 Indian languages.

Timeline

- 8 months - Data release (August)
- 1 month - Run submission (September)
- 2 months - Qrels and results (November)
- 15 days - Working notes submission (December)

Corpus

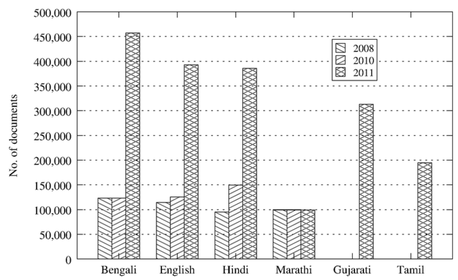

Figure 7: Corpus size

- All content converted to UTF-8
- Documents marked up in TREC format

Language	Bengali	English	Hindi	Marathi	Gujarati	Tamil
Size in GB	3.5	1.8	1.9	0.7	2.7	1.0

Qrels

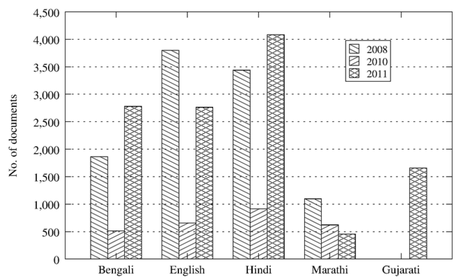

Figure 8: No. of relevant documents

- Aim was to find as many relevant documents as possible, limited to seeing 100 documents per topic.
- Assessors used boolean filters, relevance feedback, supervised query expansion.
- Pool depth varies, but usually around 50.

References

[1] P. Majumder, M. Mitra, D. Pal, A. Bandyopadhyay, S. Maiti, S. Pal, D. Modak, S. Sanyal, The FIRE 2008 Evaluation Exercise, ACM Trans. Asian Lang. Inf. Process. **9**, 2010.

[2] `http://www.isical.ac.in/~fire/2011/slides/fire.2011.majumder.prasenjit.pdf`

[3] `http://www.isical.ac.in/~fire/data.html`

Acknowledgement

- Mandar Mitra, Indian Statistical Institute, Kolkata
- Prasenjit Majumder, Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar
- TREC, CLEF, NTCIR.

Author information: Sauparna Palchowdhury, on behalf of the IR Lab, ISI, Kolkata. (sauparna.palchowdhury@gmail.com)