Text Collections for FIRE

Prasenjit Majumder, Mandar Mitra, Dipasree Pal, Ayan Bandyopadhyay,
Samaresh Maiti, Sukanya Mitra, Aparajita Sen, and Sukomal Pal
Indian Statistical Institute
Kolkata, India.
fire@isical.ac.in

ABSTRACT

The aim of the Forum for Information Retrieval Evaluation (FIRE) is to create a Cranfield-like evaluation framework in the spirit of TREC, CLEF and NTCIR, for Indian Language Information Retrieval. For the first year, six Indian languages have been selected: Bengali, Hindi, Marathi, Punjabi, Tamil, and Telugu. This poster describes the tasks as well as the document and topic collections that are to be used at the FIRE workshop.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Performance evaluation (efficiency and effectiveness)

General Terms

Experimentation, Languages, Measurement, Performance

1. INTRODUCTION

The success of TREC, CLEF, and NTCIR has established the importance of building reusable, large-scale standard test collections in information access research. The aim of the Forum for Information Retrieval Evaluation (FIRE) is to create a similar platform for Indian Language Information Retrieval (ILIR) in order to encourage research in ILIR by providing the data and a common forum for comparing models and techniques. This effort is a part of a nation-wide project (called the *Cross-Lingual Information Access* (CLIA) project) that is funded by the Indian government, and is being carried out by a consortium of ten academic and industrial institutions. The broad goal of this project is to develop resources for ILIR, along with a complete, cross-lingual information access system for English and six other Indian languages, viz. Bengali, Hindi, Marathi, Punjabi, Tamil, and Telugu. Of these languages, Hindi and Bengali rank among the top ten most-spoken languages of the world.

FIRE addresses the evaluation-related issues pertaining to ILIR. Its aim is to build test collections in the six chosen Indian languages. These are the first test collections to be constructed for IR experiments in these languages (except Hindi, which was addressed in the TIDES surprise language exercise [1]). As usual, participants will run their systems on the test collections. Results of the system evaluations will be discussed at a workshop to be held during 12-14 December, 2008 in Kolkata, India.

This poster briefly describes the test collections that will be used for the FIRE[1] workshop. In the next section, we describe the IR tasks proposed for the workshop. Section 3 covers the test collections to be used for the workshop. Section 4 lists some of the issues that need to be investigated in the immediate future.

2. TASKS

For the first year, only two tasks will be considered:

1. **Ad-hoc monolingual retrieval** for Bengali, Hindi, Bengali, Marathi, Punjabi, Tamil, and Telugu.

2. **Ad-hoc cross-lingual retrieval.** This task will be further subdivided into two parts:

    - queries in Bangla, Hindi, Marathi, Punjabi, Tamil, Telugu; documents in English and Hindi;
    - queries in English; documents in any of the six Indian languages.

In both cases, systems will be evaluated using the standard metrics implemented in `trec_eval`[2].

A Call for Participation for the workshop was sent out in January. A language-wise breakup of the number of groups that have officially registered so far is given in Table 1. Several other groups have also expressed an intention to participate in the various tasks.

Language	No of Participants
Hindi	7
Bangla	4
Telugu	3
Tamil	2
Marathi	1
Punjabi	1

Table 1: No. of registered participants

3. THE TEXT COLLECTIONS

Text collection construction is complete for some languages; for the other languages, corpus development is in progress.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

[1] http://www.isical.ac.in/~fire
[2] http://trec.nist.gov/trec_eval/

Language	No. Of Documents	Size in MB
Bangla	159142	1224
Hindi	100000	786
Marathi	200000	564
Punjabi	17906	235
Tamil	10000	168
Telugu	25000	250

Table 2: Corpus sizes

The corpus consists predominantly of news articles published in the six languages during the period from September 2004 to September 2007 in various on-line news sources. Besides news articles, the corpus also includes some documents from the health and tourism domains.

We are currently in the process of obtaining permission from the respective publishing houses to distribute the corpus to interested groups for research use.

3.1 Documents

The original articles are often written using non-standard, font-based encoding schemes. All such documents have been transcoded to use the UTF-8 encoding scheme, so that the corpus is uniformly in Unicode. In most cases, each document contains a single news article, and consists of a title, the author's / correspondent's name and the body of the article. Some statistics about the corpus in its present form is given in Table 2.

The Bengali corpus consists of 555,124 unique words. After light stemming using a statistical stemmer, the dictionary size reduces to 312,411, while after aggressive stemming using the same stemmer, the lexicon size reduces to 168,437. The maximum, minimum and mean document size in this corpus are 61891, 136, and 6278.40 bytes respectively. The corpus is also classified (based on the original categorization of the news articles) into major categories like Business, *Rajya* (state news), Travel, Editorials, *Bidesh* (international news), *Desh* (national news), Sports, Health, etc. Similar details for the other languages will be available soon.

3.2 Topics

A set of 95 topics has been created based on manual inspection of the news published in the six languages during the period September 2004 to September 2007. This news can be divided into three categories, viz. international, national, and regional. While there is a large overlap across languages in terms of international and national news content, regional news is often specific to each language. The topics were created keeping this in mind. The topics have been translated to all the six languages and English. A typical topic has a title, a description, and a narrative section. Some of these topics are expected to be discarded as either too easy or too difficult on the basis of preliminary experiments. Of the remainder, 50 topics will be used for the final evaluation. A training set of about 30 topics will also be distributed in early June, 2008.

3.3 Relevance Judgments

Since the number of participants in the final evaluation may not be as large as at the other major evaluation fora, some preliminary pooling will be done for each query. Three automatic retrieval methods based on (i) two variants of the Divergence From Randomness (DFR) [2] model, (ii) a language model, and (iii) the BM25 scheme [4] will be used to create the pools. The Terrier [3] system will be used for this purpose. The pool will be supplemented with the results of a manual retrieval run that uses the SMART [5] system. The 100 top-ranked documents from each run are expected to contribute to the pool.

Binary relevance judgments will be used. The judging process is in progress. The judges are senior students / recent graduates from technical disciplines as well as the humanities, who use a Web-based tool developed for the purpose.

Some preliminary statistics about the pool created so far for Bengali are given in Table 3. This pool was created by taking the top 20 documents from each of five runs for a set of 25 queries. The pool consists of 824 documents, indicating that there is a large overlap in the ranked lists. The pool will be supplemented by the results of a manual run.

Run ID	Unique contribution to the pool(in docs.)
BM25 + light stemming	27
BM25 + aggr. stemming	47
DFR + no stem	64
DFR + light stemming	11
DFR + aggr. stemming	41

Table 3: Pool statistics

4. CONCLUSION

The test collection is still in a formative stage. The three most important tasks that need to be addressed in the immediate future are: 1. running preliminary interactive experiments across the six languages with the query set to determine how many of them are easy / hard / of intermediate difficulty; 2. doing a more careful analysis of the pool to determing which strategies are likely to be the most effective contributors of relevant documents to the pool; and 3. enriching the pool with the results of a manual run.

The eventual test collection should serve as a benchmark that can be used to compare the performance of various techniques proposed for ILIR.

5. REFERENCES

[1] *ACM Transactions on Asian Language Information Processing (TALIP) 2:2,3.* ACM, 2003.

[2] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.

[3] I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in terrier. *Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*, 2007.

[4] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC. In *Text REtrieval Conference*, pages 21–30, 1992.

[5] G. Salton, editor. *The SMART Retrieval System—Experiments in Automatic Document Retrieval.* Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

Acknowledgements

Our sincere thanks to Donna Harman, Noriko Kando, Doug Oard, Carol Peters, Stephen Robertson, Mark Sanderson and Ellen Voorhees for their support. We would like to thank the publication houses who agreed to share their data.