Overview of the FIRE 2011 RISOT Task

Utpal Garain,[1*] Jiaul Paik,[1*] Tamaltaru Pal,[1]
Prasenjit Majumder,[2] David Doermann,[3] and Douglas W. Oard[3]

[1] Indian Statistical Institute, Kolkata, India
{utpal|tamal}@isical.ac.in
jia.paik@gmail.com

[2] DAIICT, Gandhinagar, India
Prasenjit.majumdar@gmail.com

[3] University of Maryland, College Park, MD USA
{doermann|oard}@umd.edu

Abstract

RISOT was a pilot task in FIRE 2011 which focused on the retrieval of automatically recognized text from machine printed sources. The collection used for search was a subset of the FIRE 2008 and 2010 Bengali test collections that contained 92 topics and 62,825 documents. Two teams participated, submitting a total of 11 monolingual runs.

1. Introduction

The first Retrieval of Indic Script OCR'd Text (RISOT) task was one of seven tasks at the Third Forum for Information Retrieval Evaluation (FIRE), which was held in Mumbai, India in December, 2011. The focus of the task was on evaluation of Information Retrieval (IR) effectiveness for errorful text generated from machine printed documents in an Indic script using Optical Character Recognition (OCR). Substantial effort has been invested in developing OCR for Indic scripts[1], but RISOT is the first effort to formally characterize the utility of such systems as part of an information retrieval application. The track has three primary goals: (1) supporting experimentation of retrieval from printed documents, (2) evaluating IR effectiveness for retrieval based on Indic script OCR, and (3) providing a venue through which IR and OCR researchers can work together on a challenge that requires perspectives drawn from both communities. RISOT was included in FIRE 2011 as a pilot task to begin the development of test collections and to provide an opportunity for multidisciplinary research teams to come together and collaborate.

This paper presents an overview of activities in this first year of the RISOT task. Section 2 briefly reviews prior work in evaluation of IR from printed documents, Section 3 describes the test collection and evaluation method, Section 4 introduces the participating teams and presents aggregate results, and Section 5 concludes the paper with a brief discussion of the future of the RISOT task.

2. Background

The design of the RISOT task was influenced by two previous TREC (Text Retrieval Conference) evaluations that had similar goals: the Confusion Track and the Legal Track. The TREC Confusion track was part of TREC-4 in 1995 [1] and TREC-5 in 1996 [2]. In the TREC-4 Confusion Track, random character insertions, deletions and substitutions were used to model degradations (with electronic text as the starting point). The collection to be searched included about 260,000 English electronic text documents from multiple sources, and distortion modeling was applied to either 10% or 20% of the characters. This use of character distortion models for collection development was useful as a way of quickly gaining some experience with the task, but such an evaluation design raises fidelity concerns, particularly when error models are also used in the retrieval process. The concern arises from the potential for unmodeled phenomena (e.g., correlated errors) yielding evaluation results that might not be representative of actual applications. For the TREC-5 Confusion Track, about 55,000 government announcements that had been printed, scanned, and then OCR'd (with a roughly 5% or a roughly 20% character error rate) were used instead. Electronic text for the same documents was available for comparison. Relevance judgment costs were minimized in the TREC-5 Confusion Track by using a known-item evaluation design in which each query was designed to retrieve a single item from the collection. All experiments in both years of the TREC Confusion Track were run in automatic mode (i.e., with no human

* The authors to whom correspondence should be directed.

[1] The Ministry of Information Technology, Govt. of India has been funding a nationwide consortium for developing robust OCR system for ten Indic script/languages: http://tdil.mit.gov.in/Research_Effort.aspx

intervention). Participants experimented with techniques that used error models in various ways and with techniques that sought to accommodate OCR errors by using relatively short overlapping character n-grams.

TREC returned to evaluating retrieval of printed documents in the Legal Track each year between 2006 and 2009 [3,4,5,6]. A collection of about 7 million scanned English business documents (e.g., memoranda, reports, and printed email) was searched, with the same collection being used in each of the four years. These documents were made available as part of the so-called "Tobacco Lawsuits" which took place in the USA between 1999 and 2004. Access to the printed and scanned documents provided a more natural range of variability than the documents of the TREC-5 Confusion Track, although no corresponding electronic text was available. A notable feature of the TREC Legal Track was the use of rich topic statements that were representative of those commonly used to request evidence in the legal process of "e-discovery." The TREC-2006 Legal Track included only automated experiments. This same collection (with new topics) was used in the TREC-2007 and TREC-2008 Legal Tracks with the involvement of real users being incorporated in one of two ways. The first was the use of relevance feedback experiments in which some pre-existing relevance judgments were provided with the query (2007-2009). The second was the use of fully interactive experiments in which users could work as a part of a human-machine system to obtain optimal results for a smaller number of topics (2007 and 2008).

Experiments with retrieval from printed documents were, of course, also conducted outside of community-based evaluation venues such as TREC. Most notably, early experiments with OCR-based retrieval on small collections were reported by Taghva and his colleagues at the University of Nevada, Las Vegas (UNLV) as early as 1993 [7]. Perhaps the best known among the early work was that of Singhal and his colleagues using larger collections (simulated with character corruption modes, as in TREC-5), which showed that linear document length normalization models were better suited to collections containing OCR errors than the quadratic (cosine normalization) models that were widely used at the time [8]. OCR-based retrieval is now widely used in many applications, most notably Google Books [9]. To date, however, none of this work has focused on Indic languages.

3. Test Collections

FIRE 2008 and 2010 were the first information retrieval community evaluation venues to create large-scale IR text collections for Indic languages. The RISOT 2011 data set is a subset of the existing FIRE Bengali test collection, which contains articles from a leading Bengali newspaper that were published between 2004 and 2006. The subset contains 62,825 documents, about half the FIRE Bengali collection. We refer to the electronic text from that collection as the "clean" text collection or simply the "TEXT" collection.

For RISOT, each document in the clean text collection was rendered as a document image using standard next rendering software at a resolution of 300 dots per inch (dpi). Some documents generated multiple pages, and on average 2.8 images were generated per document. The correspondence between a text page and its corresponding image(s) was maintained using a file naming convention. The resulting images are of high quality, free from the kinds of skew, distortion and spurious marks that might be found in scanned images of actual newspaper pages.

3.1. OCR Collection

A Bengali OCR system was used to convert these images into electronic text using a feature-based template matching approach [10]. Automatic evaluation [11] found the Unicode glyph accuracy to be about 92.5%. A single Bengali character might be represented using two or more Unicode glyphs, so glyph accuracy somewhat understates character accuracy. For example, if <ক><ু> were misrecognized as <ঙ>, two Unicode glyph errors would be counted. Similarly, if <স><ো><ন> were misrecognized as <ল>, three Unicode glyph errors would be counted. In each case, only one Bengali character substitution would actually have occurred.

The principal causes of OCR errors are segmentation errors (specifically, errors in the division of words into characters) and character misclassification. The Bengali alphabet contains about 250 characters, counting both basic characters and conjuncts. There are also some vowel modifiers which can attach to consonants, forming yet more new shapes. Our current OCR system treats all of these shapes as separate classes, resulting in about 700 shapes that character classification must distinguish. Thus, the character recognition problem in Bengali is nearly an order of magnitude more challenging than is the case for English.

When a single document generated multiple images, the OCR outputs for each of those images are reassembled to produce a single OCR'd document. There are therefore 62,825 OCR'd documents, and this collection is referred to simply as the "OCR" collection.

3.2 Topics

The 92 RISOT Bengali topics were taken from FIRE 2008 (topics 26-50) and FIRE 2010 (topics 51-125). Each topic consists of three parts – a Title (T), a Description (D) and a Narrative (N) – and a unique query number. The title represents what a searcher might initially type into a Web search engine, the title and description together (which we call TD) represents what the searcher might say to a human intermediary who has offered to help them with their search, and the title, description and narrative together (which we call TDN) represents what that intermediary might understand the information need to be after some discussion with the searcher. The machine's task is then to take a T or TD query and to return results that would be judged to be relevant on the basis of the full TDN topic description. The topic statements are available in several languages, but only Bengali queries were used in the 2011 RISOT pilot task. A sample topic is shown in Bengali and English below.

<top>
<num>26</num>
<title>সিঙ্গুরে জমি অধিগ্রহণ সমস্যা</title>
<desc>সিঙ্গুরে বামফ্রন্ট সরকারের জমি অধিগ্রহণ কর্মসূচি এবং ভূমি উচ্ছেদ প্রতিরোধ কমিটির বিক্ষোভ সংক্রান্ত নথি খুঁজে বার করো। </desc>
<narr>শিল্পোন্নয়নের জন্য সিঙ্গুরে কৃষি জমি অধিগ্রহণ, বামপন্থী ও বিরোধী দলের মধ্যে সঙ্ঘর্ষ, সাধারণ মানুষকে নিষ্ঠুর ভাবে হত্যা, সমাজের বিভিন্ন স্তরের মানুষের প্রতিবাদ ও সমালোচনা প্রাসঙ্গিক নথিতে থাকা উচিত। </narr>
</top>

<top>
<num>26</num>
<title>Singur land dispute</title>
<desc>The land acquisition policies of the Left Parties in Singur and the protest of Bhumi Ucched Protirodh Committee against this policy.</desc>
<narr>Relevant documents should contain information regarding the acquisition of agricultural land for industrial growth in Singur, the territorial battle between the Left Parties and the opposition parties, the brutal killing of the innocent people and the protests and the criticism by people from different sections of society.</narr>
</top>

3.3 Relevance Judgments

Relevance judgments had been created for these topics in 2008 or 2010 as part of the FIRE ad hoc task [12]. The existing FIRE relevance judgments have been limited to the documents in the RISOT 2011 collection and we reused those relevance judgments for the 2011 RISOT pilot task. Only a subset of the documents was judged (generally, those that were highly ranked by some participating system in the 2008 or 2010 ad hoc task); unjudged document were treated as not relevant.

3.4 Evaluation

RISOT 2011 participants were asked to evaluate their runs using the relevance judgments provided by the organizers and version 9.0 of the trec-eval package.[2] Participants were asked to report MAP and P@10 for both the TEXT and the OCR conditions, and to explain how they had formed their queries (e.g., as T, TD or TDN).

4. Results

Two teams participated in RISOT 2011, one from the Indian Statistical Institute, Kolkata, India (ISI) and one from the University of Maryland, College Park, USA (UMD). Both teams submitted TEXT and OCR runs with no special processing as baseline conditions. The ISI team also experimented with rule-based error correction and with query expansion. The UMD team also experimented with stemming and with statistical accommodation of likely errors. Table 1 shows the reported results for the 12 submitted runs.

As Table 1 illustrates, the best results (by P@10) were obtained using TD queries on clean text. Stemming yielded apparent improvements for each condition in which it was tried (TD TEXT, T TEXT, TD OCR) and these observed differences are statistically significant. Error modeling yielded apparent improvements for the OCR condition in all three cases in which it was tried (TD unstemmed, TD stemmed, T stemmed). Among these improvements error modeling on TD unstemmed and TD stemmed produced statistically significant

[2] http://trec.nist.gov/trec_eval/

improvements but improvement for T stemmed is observed to be statistically not significant. Notably, ISI and UMD used rather different error modeling techniques. The best results for the OCR condition achieved 88% of the P@10 (and 90% of the MAP) achieved by the same team's TEXT condition. These results suggest that practical search applications for printed Bengali documents could be constructed now. Moreover, in view of the fact that this year's relevance judgments could serve as training data for next year's RISOT task, continued research using more highly tuned approaches to error modeling and to stemming for Bengali OCR results might reasonably be expected to yield further improvements.

Run	Query	Docs	Processing	P@10	MAP
umdT2	TD	TEXT	Stemming	0.3554	0.4229
isiT1	TD	TEXT	None	0.3239	0.3540
umdE5	TD	OCR	Stemming + OCR single-error model	0.3008	0.3521
umdT1	T	TEXT	Stemming	0.2967	0.3487
isiE1	TD	OCR	OCR multiple-error model	0.2859	0.3193
umdE2	T	OCR	Stemming + OCR single-error model	0.2686	0.2967
umdE1	T	OCR	OCR single-error model	0.2583	0.2588
umdO4	TD	OCR	Stemming	0.2489	0.2915
isiO1	TD	OCR	None	0.2293	0.2318
umdO3	TD	OCR	None	0.2217	0.2293
umdO2	T	OCR	Stemming	0.2187	0.2349
umdO1	T	OCR	None	0.1901	0.1922

Table 1. RISOT 2011 results.

5. The Future

In subsequent years, we anticipate conducting an extended version of RISOT. Future evaluations may consider a number of changes:

- For the 2011 pilot task we asked participants to compute their own results using existing relevance judgments; in future years we expect to conduct blind evaluations using new relevance judgments.
- For this year's task we generated clean images. In future years, image degradation models could be applied before running the OCR. Alternatively, we could model the actual application with even higher fidelity by printing and then re-scanning at least a part of the collection. Indeed, even higher fidelity might be achieved by finding a subset of the documents that have actually been printed in the newspaper and scanning those newspaper clippings. With these approaches we could generate as many as four versions of an OCR collection.
- Some participants in future years might wish to contribute additional OCR results, or to perform retrieval tasks using image domain techniques. For such cases, the participants would need to be provided with an image collection along with the clean text collection.
- Documents in other Indic scripts such as Devanagari may also be added in future years.
- Additional evaluation measures such as Normalized Discounted Cumulative Gain (NDCG) or inferred Average Precision (infAP) may also be considered in future years.

The specific design of the task in future years will, of course, be discussed among the potential participants. We therefore encourage the broadest possible participation in the forthcoming RISOT task in order to provide a rich basis for those discussions.

References

1. D. Harman, "Overview of the Fourth Text Retrieval Conference," in *The Fourth Text Retrieval Conference*, Gaithersburg, MD, USA, pp. 1-24, 1995.
2. P. Kantor and E. Voorhees, "Report on the TREC-5 Confusion Track," in *The Fifth Text Retrieval Conference*, Gaithersburg, MD, USA, pp. 65-74, 1996.
3. J. Baron, D. Lewis and D. Oard, "The TREC-2006 Legal Track," in *The Fifteenth Text Retrieval Conference*, Gaithersburg, MD, USA, 2006.
4. S. Tomlinson, D. Oard, J. Baron and P. Thompson, "Overview of the TREC 2007 Legal Track," in *The Sixteenth Text Retrieval Conference*, Gaithersburg, MD, USA, 2007.
5. D. Oard, B. Hedin, S. Tomlinson and J. Baron, "Overview of the TREC 2008 Legal Track," in *The Seventeenth Text Retrieval Conference*, Gaithersburg, MD, USA, 2008.
6. B. Hedin, S. Tomlinson, J. Baron, and D. Oard, "Overview of the TREC 2009 Legal Track," in *The Eighteenth Text Retrieval Conference*, Gaithersburg, MD, USA, 2009.
7. K. Taghva, J. Borsack and A. Condit, "Results of Applying probabilistic IR to OCR Text," in *The Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, pp. 202-211, 1994.
8. A. Singhal, G. Salton and C. Buckley, "Length Normalization in Degraded Text Collections," in *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, USA, pp. 149-162, 1996.
9. L. Vincent, "Google Book Search: Document Understanding on a Massive Scale," in *Ninth International Conference on Document Analysis and Recognition*, Curitiba, Brazil, pp. 819-823, 2007.
10. U. Garain and B. Chaudhuri, "Compound character recognition by run number based metric distance," in Proceedings of the IS&T/SPIE 10th International Symposium on Electronic Imaging: Science & Technology, SPIE Vol. 3305, pp. 90-97, San Jose, CA, USA, 1998.
11. J. Sauvola, H. Kauniskangas, D. Doermann and M. Pietikainen. Techniques for automated testing of document analysis algorithms. In *Brazilian Symposium on Document Image Analysis*, Curitaba, Brazil, pp. 201–212, 1997.
12. P. Majumder, M. Mitra, D. Pal, A. Bandyopadhyay, S. Maiti, S. Pal, D. Modak and S. Sanyal, "The FIRE 2008 Evaluation Exercise," ACM Transactions on Asian Language Information Processing 9(3)10:1-10:24, 2010.