Evaluation in the CLIA Project

Mandar Mitra
Indian Statistical Institute
Kolkata
Email: mandar@isical.ac.in

Prasenjit Majumder
DAIICT
Gandhinagar
Email: p majumder@daiict.ac.in

Abstract:- ***An overview of the Evaluation sub-system of the CLIA project is given in this paper. We start with a review of standard practices in Information Retrieval (IR) evaluation. The process followed in Phase I of the project is described along with the achievements. The evaluation strategy to be followed in Phase II is outlined next. We conclude by discussing some of the challenging issues in IR Evaluation.***

Keywords- *CLIA project; Cranfield paradigm; evaluation; FIRE; metrics;*

I. INTRODUCTION

The CLIA (Cross-Lingual Information Access) project was initiated by the Government of India with the aim of building a cross-lingual information access portal on the Web where a user will be able to enter a query in his/her preferred language and get back information from diverse sources. One of the distinguishing features of the portal is that while the information returned to the user may originally be in a different language, it will be presented to the user in her preferred language.

The aim of the Evaluation sub-system is to create a framework for quantitatively evaluating various components of the CLIA system, in particular, the ranking of search results returned by the CLIA system. This ranking determines how users perceive the system. The evaluation of this ranking can therefore be regarded as indicative of the overall, end-to-end performance of the system.

In general, an evaluation framework is important for the following reasons. First, it can be used to tune various parameters of the CLIA system in order to improve performance. More generally, it can be used to compare different models, algorithms, and techniques in a fair and uniform way. This allows us to measure progress. The availability of large-scale, standardised test data also provides the impetus necessary to drive new research in an empirical discipline like IR. Thus, the broader motivation behind the Evaluation subsystem is to create a framework that is not limited in its applicability to the CLIA system. Our goal is to create benchmark datasets that will continue to be useful to the IR / Language Processing community as a whole — both in India and abroad — even after the completion of the CLIA project.

II. EVALUATION METHODOLOGY

In accordance with standard practices followed by the IR research community all over the world (and in particular, by the best-known evaluation initiatives such as TREC1, CLEF2, and NTCIR3), we have adopted the Cranfield paradigm [1] in our approach. In this paradigm, a system indexes a test collection of documents and returns results for a set of test queries. The system is given a quantitative score (using standard evaluation metrics) based on the quality of the search results. The system gets a high score if useful / relevant documents are returned early. These various components that

1http://trec.nist.gov
2http://www.clef-campaign.org
3http://research.nii.ac.jp/ntcir/index-en.tml

comprise a benchmark dataset are described in more detail below.

Corpus: A corpus is a collection of documents through which the system must search for useful information in response to a user-query. A corpus is often created by crawling pages from the Web. Ideally, the documents in the benchmark corpus should all use a standard character encoding (e.g. UTF-8). However, since the actual data available on the Web does not always conform to this standard, in order to be realistic, the benchmark data should also contain pages that use proprietary, non-standard encodings in addition to UTF-8 pages. Further, the corpus should be large, consisting preferably of at least 100,000 pages.

Topics: This is a set of sample queries that are used for testing the search system. The set should consist of at least 50, and preferably 100 or more queries. The queries should be modeled on real-life information needs.

Relevance judgements: Relevance judgements provide information about which documents in the corpus are useful (or relevant) for which queries. Exhaustive relevance judgements can be created by examining each document for each query and determining its usefulness. However, this method is not scalable, as it may involve judging many hundreds of thousands of query-document pairs. The other alternative is to use pooling [2] to create a smaller set of query-document pairs that are manually examined in order to create the relevance judgements. Pooling has been used at all major evaluation fora, and has also been adopted to create the relevance judgements for some of the CLIA evaluation data.

Pooling works as follows: each of several diverse IR models / methods / algorithms are used to retrieve N documents for a given query. The union of these result lists is constructed.

This forms the so-called pool, and only this pool is manually judged. Any document that is not in the pool is assumed to be non-relevant, on the assumption that if a document were relevant, it would have been retrieved by at least one of the different IR systems contributing to the pool. Thus, the (large) portion of the corpus that is not in the pool does not need to be explicitly judged, saving time and effort. Note that, for the assumption underlying the pooling approach to be accurate, diverse effective systems need to contribute to the pool.

Common Metrics

Most evaluation metrics are based on the ideas of recall and precision, given by the following expressions:

Informally, recall measures how comprehensive the results returned by a search engine are, while precision measures how focused (or accurate) they are.

$$\text{recall} = \frac{\text{\# rel. docs. retrieved by a system}}{\text{\# rel. docs. present in the corpus}}$$

$$\text{precision} = \frac{\text{\# rel. docs. retrieved by a system}}{\text{total \# docs. retrieved by the system}}$$

The metrics that are commonly used for performance evaluation of search systems include precision at rank k (P@k), and average precision (AP). These metrics are formally defined below, using the following notation.

$$\begin{aligned} N &= \text{no. of documents retrieved in response to query } Q \\ R &= \text{no. of documents that are relevant for query } Q \\ D_i &= \text{document retrieved at rank } i \ (i = 1, \ldots, N) \\ I_i &= 1 \text{ if } D_i \text{ is relevant to } Q \\ &\quad\ 0 \text{ otherwise} \end{aligned}$$

In other words, P@k simply measures the proportion of relevant documents in the list of k top-ranked documents. The definition of AP

$$P@k = \sum_{i=1}^{k} I_i/k$$

$$AP = \frac{1}{R} \sum_{I_i=1} (\sum_{j=1}^{i} I_i)/i$$

is somewhat more involved. It is calculated as follows. If Di, the i-th document retrieved by a system (i = 1, . . . ,N), is relevant (i.e. Ii = 1), the precision is calculated at rank i. These precision values are added up and the total is divided by R, the total number of relevant documents for that particular query. Standard practice involves calculating these measures for each query in a benchmark query set, and reporting the overall (arithmetic) mean or average.

Since for most web-searches, user satisfaction is determined more by the precision or accuracy of results, we focus on P@k, with k - {5, 10, 20}. This decision was taken based on the observation that users rarely look at more than 2 pages of results returned by a search engine (with each page typically containing 10 results). Also, given that the "corpus" for the CLIA system is in some sense the entire Web, there is no easy way to measure recall. Thus, we avoid measures such as AP.

FIRE 2008 data				
Language	Bengali	Hindi	Marathi	English
# queries	75	75	75	75
# docs. judged	11,966	23,587	8,155	18,656
FIRE 2010 data				
Language	Bengali	Hindi	Marathi	English
# queries	50	50	50	50
# docs. judged	8,655	22,572	20,761	15,135

TABLE I
RELEVANCE JUDGEMENT DATA CREATED IN PHASE I (GENERAL DOMAIN)

III. EVALUATION IN PHASE I

As mentioned in the Introduction, the goal of the Evaluation sub-system was two-fold: (i) to tune the CLIA system and measure the quality of results returned by it; and (ii) to create benchmark data that can be used by IR researchers for testing general purpose IR systems. Accordingly, the benchmark data created in Phase I of the CLIA project falls into two categories:

1) Data drawn mostly from the News genre for testing general purpose IR systems. This data was created following well-established practices that are already in use at TREC, CLEF and NTCIR, and has been used for the FIRE4 (Forum for Information Retrieval Evaluation) evaluation campaigns [3].

2) Tourism domain data for specifically testing the CLIA system. The process of creating relevance judgement (RJ) data for the tourism domain is described below.

A substantial amount of benchmark data has been created. Some details are provided in the Tables I and II.

Tourism domain data

A number of tourism-related queries were formulated by each language group within the CLIA consortium. Some of these queries were about tourist destinations that are widelyknown and of national interest (e.g. the Taj Mahal). From these queries, a set of common queries were selected for testing by all language groups. Other queries that were about local attractions that may not be as well-known in other parts of the country (e.g. the Bandel church close to Kolkata) were used by individual language groups. In all, about 100 queries were used for testing the system in each language.

For each topic, the top 20 results fetched by the CLIA system as well as Google were manually examined. The quality of results returned by Google serves as a state-of-theart baseline against which the CLIA system can be compared. The CLIA system retrieves pages in

4http://www.isical.ac.in/93fire

Doc. language →	Bengali		Hindi		English	
Query language ↓	CLIA	Google	CLIA	Google	CLIA	Google
Bengali (100 queries)	223 (27 rel.)	1,718 (325 rel.)	1,592 (93 rel.)	N/A	1, 671 (487 rel.)	N/A
Hindi (100 queries)	N/A	N/A	749 (126 rel.)	1,705 (408 rel.)	1,609 (362 rel.)	N/A

TABLE II
RELEVANCE JUDGEMENT DATA CREATED IN PHASE I (TOURISM DOMAIN)

the original language, Hindi and English; for Google, only monolingual results were judged. Due to inadequate coverage of the CLIA system crawl, fewer than 20 (sometimes as few as zero) results were retrieved for some queries.

IV. LOOKING AHEAD: PHASE II

The overall goals of the Evaluation sub-system in Phase II of the CLIA remain the same as in Phase I. Since benchmark data for the general domain (i.e. FIRE data) was by and large successfully created in Phase I following well-established procedures, the same procedures are being followed in Phase II as well. The creation of tourism domain data, primarily intended for the evaluation of the CLIA system, is significantly different in some ways from the creation of FIRE data. The FIRE corpora were created from a one-time crawl of Web sources; these corpora are therefore static. In contrast, the tourism-domain corpus is formed from the pages crawled by the CLIA system. Since the CLIA system is a live search engine, its crawl needs to be regularly updated. The corpus is therefore constantly changing: new pages are added, existing pages may disappear, or their content may change. Thus, the relevance of a page with respect to a particular query may also also change. More generally, the values of the evaluation metrics may change over time for the same system and for the same set of queries, simply because the underlying document base has changed.

Keeping the dynamic nature of the evaluation test-bed in mind, the following instructions will be followed in Phase II to create RJ data.

1) Formulate an information need. Assign a serial number to the information need. Write down what you are looking for, in plain English (e.g. “I want to find information about houseboats in Kashmir, how to book them, what are the charges, etc.”). The Lonely Planet website and Yahoo Answers are good sources of realistic travel-related information needs.

2) Submit a suitable query to (i) the CLIA system (ii) Google. Use the same string in both cases (e.g. Kashmir houseboat reservation).

3) Note down in order in a PLAIN TEXT file the top 20 URLs from each set of results returned by CLIA (English, Hindi, language of query (e.g. Bangla)), and the top 20 URLs from Google (mono-lingual and cross-lingual whenever available). Each set of results should be in a file named using the following convention: hquery languagei.hdocument languagei. henginei (e.g. bn.hi.google, en.ta.clia).

4) For each URL, look at the corresponding page, and determine if it addresses your information need. It is not enough for the page to simply contain the keywords that you entered (e.g. if a page says “Kashmiri

houseboats are very famous, but this page talks about houseboats in Honolulu", then the page is not relevant). If it does, mark it Y; if it does not, mark it N; if it is partially relevant, i.e. it is about the topic of the query, but does not provide the precise information sought, mark it P. If it does not contain useful information, but useful links, mark it L. If the page is otherwise problematic (e.g. it is not in the language that it claims to be, or the URL is invalid), mark it X. Each line in the judgement file should look like: hquery numberi hURLi hY/N/P/L/Xi

5) Document each step, and repeat with different information needs / queries. Please also note down the date when the query was tried.

Some important issues that need to be borne in mind while evaluating the CLIA engine are briefly discussed below.

- Since Google has recently started providing cross-lingual search results in addition to the traditional monolingual results, it is important to compare the results of the CLIA system with the results returned by Google (see step 3 above). This will serve to quantify the addition of value by the CLIA project to services that are already freely available.
- Many of the tourism-domain queries that were used in Phase I were fairly simple, one or two-word queries. The information need underlying such queries is usually of the form "I am looking for any and all useful information about X" (where X may be "Taj Mahal", "Golden Temple", etc.). A study of prominent tourism-related portals such as Lonely Planet (or Yahoo Answers) shows that information needs are often far more complex in practice. In order to do a fair and realistic evaluation of the CLIA system, the set of test queries should include an increasing number of such complex queries in Phase II.
- Creating RJ data is effort-intensive and time-consuming. It is important to carefully document all work done, so that no work is wasted. Rather than imposing the additional burden of documentation on the assessors, it would be a good idea to design the interface so that it automates the documentation process, to the extent possible.

  With this end in mind, the search interface — a publicly available web-portal — will be modified as follows.

  – Assessors from the language groups will be able to create accounts using which they will access the interface. General users will not (at least for the time being) be given accounts.

  – Once assessors login to their accounts, the interface will keep track of their actions.

  – Specifically, authenticated users will be permitted (indeed, required) to enter a detailed description of their information need in addition to providing the usual set of search keywords.

  – A judgement button will be provided next to each result, in addition to the usual title and snippet.

  – For each user session, the information need, the keyword query, the list of results returned (by the CLIA engine, as well as by external engines like Google), and the judgement for each result will be logged by the interface along with an associated timestamp.

– Since detailed logs will be maintained for each user, it will be possible for the system to determine whether a huser, query, pagei triple is a duplicate of an existing triple. Such duplicates may be specially flagged so that the assessor need to redo an assessment that was completed earlier. Of course, if the content of a page corresponding to a particular URL changes, this triple will also change. Thus, assessors will be required to reassess pages that have been modified since the previous assessment.

V. CHALLENGES

There are a number of issues that need to be resolved when formulating an evaluation strategy. In this section, we discuss some of these issues.

- Creating a corpus of adequate size: while Indian language content on the Web is growing, not all languages are equally well represented. Obtaining adequate data for some language + domain combinations has been difficult.
- Creating a sufficiently large set of search topics that can be translated into multiple languages and then used for testing the performance of the system for these languages has been difficult. For several queries, relevant pages exist in the language in which the query was originally formulated, but there are no / few relevant pages in other languages.
- For the quantitative evaluation measures to be stable, the test topics must be neither too difficult, nor too easy. Formulating topics that will have this property across multiple languages has proved to be difficult.

In Phase II of the CLIA project, we will try to address some of these issues.

REFERENCES

[1] C. W. Cleverdon, “The significance of the Cranfield tests on index languages,” in Proc. ACM SIGIR. ACM Press, 1991, pp. 3–12.

[2] K. S. Jones and K. van Rijsbergen, “Report on the need for and provision of an ”ideal” information retrieval test collection,” British Library Research and Development Report, Computer Laboratory, University of Cambridge, Tech. Rep. 5266, 1975.

[3] P. Majumder, M. Mitra, D. Pal, A. Bandyopadhyay, S. Maiti, S. Pal, D. Modak, and S. Sanyal, “The FIRE 2008 evaluation exercise,” ACM TALIP, vol. 9, no. 3, 2010.