The BETTER Cross-Language Information Retrieval Datasets

Ian Soboroff
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
ian.soboroff@nist.gov

ABSTRACT

The IARPA BETTER (Better Extraction from Text Through Enhanced Retrieval) program held three evaluations of information retrieval (IR) and information extraction (IE). For both tasks, the only training data available was in English, but systems had to perform cross-language retrieval and extraction from Arabic, Farsi, Chinese, Russian, and Korean. Pooled assessment and information extraction annotation were used to create reusable IR test collections. These datasets are freely available to researchers working in cross-language retrieval, information extraction, or the conjunction of IR and IE. This paper describes the datasets, how they were constructed, and how they might be used by researchers.

CCS CONCEPTS

• **Information systems** → **Test collections**; **Multilingual and cross-lingual retrieval**; • **Computing methodologies** → **Information extraction**.

KEYWORDS

information retrieval, test collection, information extraction

ACM Reference Format:
Ian Soboroff. 2023. The BETTER Cross-Language Information Retrieval Datasets. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3539618.3591910

1 INTRODUCTION

Cross-language retrieval is a rich subject with a long history in research and practice. While CLIR experiments have taken many forms since the 1990s, the common element is that the information needs are expressed in a different language than the documents being searched. The Text REtrieval Conference (TREC) cross-language tracks pioneered test collections in European languages, Chinese, and Arabic. The NII-NACSIS Test Collection for IR Systems (NTCIR, now the NII Testbeds and Community for Information access Research) and the Cross-Language Evaluation Forum (CLEF, now the Conference and Labs of the Evaluation Forum) were spun out in the early 2000s as regional evaluation venues to support broader test collection development in non-English languages. The final TREC cross-language tracks in 2001 and 2002 built the first IR test collections for English-Arabic CLIR.[8] Recently, the IARPA Machine Translation for English Retrieval of Information in Any Language (MATERIAL) program [9] focused on CLIR, and the new TREC NeuCLIR track is building new test collections for cross-language and multilingual search.[5]

Approaches to CLIR include machine translation of the documents or queries, pivoting through a common vector encoding, and using multilingual large neural language models. There have been a number of excellent surveys of CLIR, including a very recent one by Galuščáková et al.[3]

Information extraction (IE) was defined as an evaluation task in the Message Understanding Conference (MUC) and the Automatic Content Extraction (ACE) evaluations. IE includes named-entity recognition, entity linking, relation finding, event extraction, identification of beliefs and complex sentiments, and more. NIST's Text Analysis Conference (TAC) alongside the DARPA Deep Exploration and Filtering of Text (DEFT), Active Interpretation of Disparate Alternatives (AIDA), and Knowledge-directed AI Reasoning Over Schemas (KAIROS) programs provided numerous refinements to the task. From the mid-1990s multilinguality was also an intense focus for information extraction research.

2 THE IARPA BETTER PROGRAM

The IARPA BETTER (Better Extraction from Text Through Enhanced Retrieval) program started in 2019, with the overarching goal of tying IR and IE closer together in a virtuous cycle, where extraction could benefit retrieval and retrieval could benefit extraction. The BETTER IR test collections are best described in the context of the program, and this section presents essential background on BETTER. Aside from this section, we do not describe the information extraction datasets in detail, instead focusing on the IR test collections.

BETTER program evaluations included three traditional IE evaluations over "abstract", "basic", and "granular" event schemas, and a CLIR evaluation including a human-in-the-loop component. There was an evaluation done in each of the three phases of the program, and so are referred to in this paper as the Phase 1, 2, or 3 datasets. NIST was responsible for the IR evaluation including development, assessment, and annotation of the collection.

There are three levels of information extraction annotated in BETTER. The first is "Abstract".[6] The Abstract schema describes events with agents, patients, an event anchor and a "quadclass". The quadclass is an event typology from the field of political science. The first part of the quadclass indicates whether the event is *material* (that is, a physical activity like an attack or a transaction) or *verbal* (for example, a communication, meeting, or statement). The other dimension of the quadclass indicates whether the event is *harmful*

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591910

or *helpful* towards the patient(s) of the event.[1] Abstract extraction is evaluated at the sentence level.

The second level, "Basic", is a lightweight event frame with an event type, agents, patients, an event anchor, a slot for "referred events" and a "state of affairs" flag. A referred event is a pointer to another Basic event, for example if a government issued a statement (a communication event) condemning a bombing (an attack event). An event was deemed a "state of affairs" if it represented a continuing situation rather than something happening just at that point in time. Basic extraction is evaluated on full documents, but an event and its arguments are constrained to occur within a single sentence. The event types are of the common ACE variety, including communications, legal actions, crimes, etc.

The third level, "Granular", comprises templates made up of Basic events and string fillers for non-event slots. Granular templates represent higher level events such as protests, disease outbreaks, cyber-attacks, instances of government corruption, and refugee migrations that were the focus of the BETTER program. Granular templates are constrained to a single document, and only a few slots might be filled. BETTER Granular templates are reminiscent of MUC scenarios.

The information extraction datasets were built by MITRE and ARLIS,[1]. The program IE evaluations were of the common variety where systems predict the annotations and are measured on how closely they match the gold-standard annotations. In the evaluation setting, participants did not have training data in the target language, but were expected to pivot from English training data.

3 IR IN BETTER

The information retrieval aspect in BETTER provided a user-level task. IR can be considered to fill the co-reference gap in Basic events and Granular templates, connecting related events across multiple documents. However, document relevance is not defined as the presence of an event or template filler, but rather as how useful the information in the document is to the user in completing their task. The user is taken to be an information analyst working on a report about a large phenomenon or series of events, and in order to compose that report they have a number of requests they will make against the search system. The user has examples of what they want and would prefer to just hand the examples to the system instead of composing queries.

BETTER information retrieval has a number of unusual aspects. Individual search requests (analogous to TREC topics) were grouped together into analytic tasks that defined a larger context for the requests. Rather than having access to parts of a topic statement, search was query-by-example: the tasks and requests had example relevant passages in English. The provided passages were annotated for Basic events, so systems could make use of information extraction examples during retrieval. These extractions were used as diversification aspects in the metric. There was an interactive "human-in-the-loop" component for user feedback to the system.

The document collections in each phase are divided into an English training corpus and an evaluation corpus. In the BETTER program, systems could be developed using the training corpus, and example passages in the tasks and requests come from that corpus. Systems did not see the evaluation corpus until the actual evaluation, which was conducted in a containerized environment at MITRE.

Collection	English to...	Tasks	Reqs	Docs	Rel
Phase 1	Arabic	8	54	864,971	918
Phase 2	Farsi	10	53	856,167	1,161
Phase 3	Russian, Chinese, Korean	10	48	1,000,000	1,346

Figure 1: A high level view of the three BETTER IR test collections.

The document collections are taken from CommonCrawl.[2] The phase 1 documents were collected by MITRE, and for phases 2 and 3 by ARLIS. Care was taken to identify subsets of the CommonCrawl collection that were likely to contain topical information.

An example of an analytic task from the Phase 3 collection is "Understand the breadth of Chinese investment and control of shipping port facilities in other countries." (See Listing 1) This is the *task statement*, and it is accompanied by a narrative paragraph similar to a TREC narrative, and sections listing "gray-area" facets that are in or out of scope of the task. There are two relevant passages from two different documents in the English training data included as examples of the kind of information the user considers to be within the scope of the task. The passages are annotated following the Basic scheme, and the annotations are included in what is shown to systems.

Each task has three or more analytic requests. These are analogous to individual topics in a TREC collection, and are the observational unit for averaging measures. An example of an analytic request from the above task is "Identify specific projects and their locations, either underway or in the planning stages or complete." This is the *request text* and is the only description given for a request. Each request also has two relevant passages from two documents as examples, and again the examples are annotated in the Basic scheme. The number of tasks and requests were such that the total number of requests was around 50.

In the BETTER program, search was *by example*: systems received the example passages for the task and for each request, but **not** the topic statement. In the *automatic* evaluation condition systems were required to search for relevant documents without any human intervention and only using the examples. In the contrastive human-in-the-loop (*HITL*) condition, a user was shown the topic statements and examples for the task and for the first two requests per task. Based on this information, the user could spend a limited amount of time interacting with a system and the English training corpus (not the documents in the target language) with the goal of refining search and/or extraction to the specific analytic tasks. Because the evaluation was containerized in a limited-access environment, actual interactive use was not possible, and so HITL-tuned systems were containerized and sent in to be evaluated. The program required that the HITL users not be the system engineers

[1] https://www.arlis.umd.edu/

[2] https://commoncrawl.org/ One could argue that this data was not truly "hidden" since the data is open and systems could have incorporated CommonCrawl data in their models. BETTER systems were restricted from doing this. See below in Section 8 for a forward-looking perspective.

```
  {
    "task-num": "IR-T1",
    "task-title": "Port facilities",
    "task-stmt": "Understand the breadth of Chinese investment and
control of shipping port facilities in other countries.",
    "task-narr": "As part of the Belt and Road Initiative, China has
invested heavily in ports in a number of countries.  Because of the
enormous presence of shipping in the world economy, coupled with
transportation in international waters as opposed to links controlled
by countries, the Belt and Road Initiative has taken a special
interest in container ports.  In some cases, China has taken full
control of the port, either through being the majority investing
partner, or by taking control upon the country defaulting on the loan.
The goal of this task is to identify ports that have come under
Chinese control, what entities are involved, and what is the degree
of control exerted.",
    "task-in-scope": "This task is focused on sea ports, and
primarily on container shipping.  ",
    "task-not-in-scope": "Airports or landlocked port logistics
facilities are not in scope.",
    "task-docs":
      "296ed0fa-7af1-4f7e-8813-bb540efb5cf7": {
        "doc-id": "296ed0fa-7af1-4f7e-8813-bb540efb5cf7",
        "entry-id": "296ed0fa-7af1-4f7e-8813-bb540efb5cf7",
        "annotation-sets": {
          "basic-events": {
            "events": {
              "event-1": {
                "agents": [ "ss-2" ],
                "anchors": "ss-4",
                "event-type": "Construct-Project",
                "eventid": "event-1",
                "money": [],
                "patients": [ "ss-1" ],
                "ref-events": [],
                "state-of-affairs": false
              },
            ...
        },
        "segment-text": "The growing web of trade routes, including
the Silk Road Economic Belt and the Maritime Silk Road Initiative,
now extends into at least 76 countries, mostly developing nations in
Asia, Africa, and Latin America, plus a handful of countries on the
eastern edge of Europe.\n\nChina's plans to build or rebuild dozens
of seaports have sounded alarm bells in Washington and New Delhi: how
many of those docks will end up hosting Chinese warships?\n\n",
        "segment-type": "highlight"
    },
    ...
    "requests": [
      {
        "req-num": "IR-T1-r1",
        "req-text": "Identify specific projects and their locations,
either underway or in the planning stages or complete.",
        "req-docs": {
          "78432d30-d106-4694-bdb6-b4b772d337bb": {
            "doc-id": "78432d30-d106-4694-bdb6-b4b772d337bb",
            "entry-id": "78432d30-d106-4694-bdb6-b4b772d337bb",
            "annotation-sets": {
            ...
          }
          ...
}
```


Listing 1: The first analytic task and its first request in the Phase 3 collection. The task shows part of the first task-level example: the example passage and one annotation. The requests follow the task examples.

or have strong IR or NLP backgrounds, in order to keep HITL approaches aimed at BETTER's target user base, analysts who are experts in their subject domain but not in computational linguistics, information retrieval, or computer science. Typically HITL users were students at that research institution but from outside the BETTER project.

Since the obvious interactive thing to do given current IR technology is to ask the user to compose a query and then do relevance feedback, there was a standard *auto-HITL* condition where the system had access to the tasks and first two requests as in HITL, but had to process those completely automatically.

Aside from these experimental condition requirements, BETTER systems could use any approach to retrieval and extraction. The drive of the program was that systems should make use of extraction to improve retrieval. BETTER systems ran the gamut from traditional probabilistic systems to employing large language models.

Relevance was marked on a graded scale:

irrelevant (0) The document is not at all relevant.

topical (1) The document is topically relevant to the request. That is, it's in the ballpark of the analytic task, but doesn't help answer the specific request.

specific info (2) The document contains specific information that contributes to an understanding of the analytic request.

direct answer (3) The document contains a direct answer to the analytic request.

decisional (4) A document is "decisional" if it will drive a direct decision on the situation giving rise to the analytical task. It is a "home run" document, it is a primary citation in the report, it is the "smoking gun," pick the metaphor of your choice.

Note that the scale is grounded in a conception of task completion or task success. That is to say, the relevance grade is directly tied to how useful the document is to the user situated within the context of a notional report-writing task. This is quite different from a trinary irrelevant/relevant/highly-relevant scale where the distinction between relevant and highly relevant is undefined, or a perfect/excellent/good/fair/bad (PEGFB) scale where the levels may not be strongly defined or placed in the context of the user's task. By aligning relevance levels to the task we hoped to limit assessor disagreement due to the judgment scale. Note also that relevance level 1 is equivalent to TREC relevance, given that instead of a topic BETTER has a specific information request. Only relevance level 2 and above counted as "relevant" towards scoring metrics in the BETTER evaluation.

Each phase of the BETTER program had (a) different target language(s) and analytic focus area(s) for search needs. In phase 1, the target language was Arabic, with a focus on government corruption and protests regarding it. For phase 2, the target language was Farsi, and the foci were natural disasters, disease outbreaks, and refugee migration. For phase 3, there were three target languages, Chinese, Russian, and Korean, and two target foci, cyberattacks and energy, transportation, and infrastructure projects occurring broadly within the Chinese Belt and Road initiative.

4 TOPIC DEVELOPMENT PROCESS

Topic development[3] is the process of identifying user search needs that are good fits for the test collection: they should be realistic in the context of the task domain, they should not be too difficult considering our understanding of the state of the art, and they should not seem to have an overabundance of relevant documents in the target collection. For BETTER, we adapted the process used for many of the TREC adhoc collections.

For the phase 1 and 2 collections, the analytic tasks and requests were developed by assessors who were bilingual in English and the target language; these assessors would also perform relevance assessments. We provided the assessors with a web-based search tool using Patapsco[4] as a back-end. Patapsco is based on Pyserini,[5] a toolkit for repeatable IR experiments itself based on the open-source Lucene search library,[6] and adds high quality tokenizers for foreign languages. The web-based tool offers two tabs to separately search the English training corpus and the target language evaluation corpus, and the search tool supported stemming of search terms and quoted phrases. Note that these are both monolingual search tools provided to bilingual assessors, rather than a cross-language search tool. Additionally, the tool supports marking relevant documents in searches and composing the analytic task and request text sections.

The assessors were instructed to first explore the English training corpus to find topics or events within the analytic focus area(s). After arriving at a potential analytic task, they searched the evaluation corpus to determine if there were too many relevant documents about the analytic task. This was done by executing a search and counting relevant in the top 25 results; if there was less than one or more than 20 relevant documents,[7] then the task needed to be revised. After arriving at an acceptable analytic task, the assessor would compose the textual sections (task statement, etc.) for the task. They would also select two relevant passages from two relevant English training corpus documents that would serve as the task-level examples.

After the analytic task was defined, the assessors would compose 4-10 analytic requests as components of the analytic task. This development was done the same way as for the task: light exploration in the English training corpus, and a counting search in the foreign language evaluation corpus. If this succeeded they would compose (or finalize, as often they would compose prior to searching) the request text and identify two relevant English passages for the request.

In phase 3, due to unforeseen circumstances we did not have bilingual assessors under contract at the time they were needed. The JHU HLT Center of Excellence processed the evaluation collection with their state-of-the-art neural machine translation system, translating Russian, Chinese, and Korean into English, and we developed the topics using the machine translation to work with the foreign language evaluation corpus. This probably resulted in somewhat lower diversity of tasks in the phase 3 collection due to a single person contributing topics.

There is also a difference between phase 1 and the later phases. In phase 1, the analytic requests were appropriate and reasonable requests to make in the process of composing a report on the analytic task, but they didn't have any particular connection to the Basic or Granular extraction schemas. As a result, relevant documents rarely had extractions that coincided with the task. Since the goal of the program was to explore an intertwined retrieval and extraction setup, the phase 2 and 3 collection requests were changed to specifically target an appropriate granular template for that task, and as such are more event- and entity-focused than those in phase 1. An example phase 1 request which we would not have used in later phases is, "Have the protests in Ethiopia led to democratic changes?" This request looks like a typical TREC topic but is quite different from the example regarding infrastructure projects described in Section 3.

Despite the question-like structure of many of the analytic requests and their just-the-facts flavor, the BETTER IR task is not a question-answering task. The user task is to find documents to support their report on the analytic task, and an answer is not sufficient for a report citation.

Following topic development, the English example passages were annotated in the Basic schema by annotators at ARLIS. The JSON topic format includes the annotations with references to character offsets in the passage.

5 POOLING AND ASSESSMENT

There were four "performer" teams in the BETTER program, and each could submit one container each for the automatic, auto-HITL, and HITL conditions. Submissions were staggered: after the automatic and auto-HITL submission deadline, the HITL version of the analytic tasks were released and teams could start their users interacting with their system. Teams had two weeks within which to conduct the interactions and prepare the HITL container. The submitted containers were run within a closed AWS-based evaluation environment, and limited to five days of runtime.

In phase 1, performer systems had to output both a top-1000 ranking for each request and Basic extractions on the top 100 documents. This proved to be too much extraction for systems to accomplish within the five-day evaluation execution window, and so for phases 2 and 3 we switched to running extraction on an annotated document set drawn from the retrieval pools as a separate run. For purposes of pooling, we had twelve runs for each phase: three conditions from four teams. We planned to assess the documents in the pool and select a subset of relevant documents for annotation.

The primary goal was to compare the performance of each team across the conditions, and so in the first phase we assessed depth-10 pools from all twelve runs. For the phase 2 and 3 collections, the pools were drawn to depth 50 in hopes of having a better estimate of recall.

The assessment procedure had the assessors review the pool documents in random order. If they decided that a document was relevant at the level of "specific info" or higher, they were to highlight the passage containing that information, or the answer (if it was a "direct answer" or "decisional"). They were instructed to

[3] Alongside the BETTER program terminology of "analytic tasks" and "requests", in this paper we also use the more traditional term "topic" to refer to an analytic task with its set of requests.
[4] https://github.com/hltcoe/patapsco
[5] https://github.com/castorini/pyserini
[6] https://lucene.apache.org/
[7] During this stage relevance was not formalized into the scale above, but left as "contains an answer to the request."

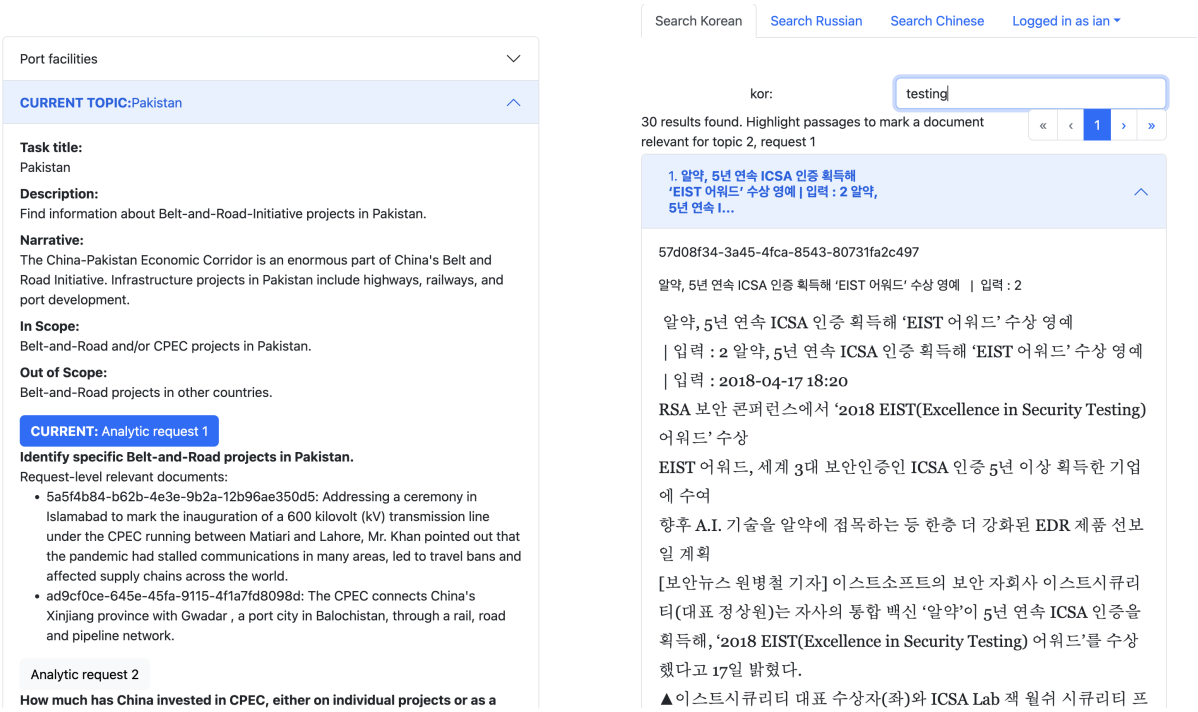

Figure 2: A screenshot of the search tool provided to assessors. This is the version that they used to conduct manual searches for relevant documents for the phase 3 collection.

Collection	Tasks	Reqs	0	1	2	3	4
Phase 1	8	54	2,474	511	458	334	126
Phase 1b	4	25	1,115	198	253	72	105
Phase 2	10	53	7,165	912	968	614	29
Phase 3	10	48	26,757	1,382	845	424	75
Phase 3b	9	42	11,013	813	397	265	7

Figure 3: Tasks, requests, and number of judged documents at each relevance level. The 1b and 3b collections are the secondary relevance judgments.

keep passages short, with a maximum of three sentences. Only a single passage was permitted and it was required to be a contiguous segment of text, so for some documents on the high end of the relevance scale the selected passage is just one useful part of a very useful document.

We were able to collect assessments from a second assessor in phase 1, covering 25 requests from four analytic tasks. The inter-assessor agreement as measured by Krippendorff's α between the primary and secondary judgments is 0.51, a typical level of agreement for IR relevance judgments. For phase 3, we have secondary assessments for nine out of ten tasks, mostly for Chinese; for request-document pairs where we have two judgments, the $\alpha = 0.49$.

In phase 3, since we had developed the topics in English using MT, we had the assessors do manual searches in their language to try to find as many relevant documents as they could. We gave them up to an hour per request but often they reported finishing in 20 or 30 minutes. Figure 2 shows the interface used for searching. The assessor would highlight a passage that contained an answer to the request, if they found one, so these judgments are equivalent to "specific information" (2) or higher. These documents were included in the pools and reviewed by the assessor for that analytic task. The purpose of this was to improve the recall base of the test collection by providing relevant documents from an unpooled source.

In phase 1 and 2, a sample of relevant documents were selected for Basic annotation, and in phase 3, we chose to only annotate the highlighted passages in order to cover more documents. These annotations serve two purposes. The primary driver is to support scoring with the combined IR-IE metric described in Section 6. Additionally, performer's Basic extraction systems were run on these documents to see if extraction performance was different in relevant documents versus in an arbitrary document sample.

In phase 3, the primary retrieval was a multilingual list of documents from (potentially) all three languages. The teams also returned a monolingual ranking for each language, and so for a given task-request-language pool, there were two rankings that could be pooled from each team and condition. Additionally, there was an extra condition where teams were to run their retrieval process with no information extraction processing, to make the "plain IR" baseline clear. In total, each team returned up to four rankings for each request in each of six conditions, (automatic, HITL, auto-HITL) x (IE, no-IE).

Collection	nDCG	MAP	P@10
Phase 1	0.1	0.04	0.05
Phase 2	0.06	0.06	0.07
Phase 3	0.1	0.07	0.06

Figure 4: Minimum meaningful score differences in each collection with a 5% error rate, as computed by Sakai's discriminative power method.

Collection	nDCG		MAP		P@10	
	min	max	min	max	min	max
Phase 1	0	0.047	0	0.061	0	0.140
Phase 2	-0.091	-0.038	-0.031	-0.010	0	0.020
Phase 3	-0.005	0.005	-0.008	0.003	0	0.006

Figure 5: Minimum and maximum absolute score differences observed by holding a group's unique contributions from the relevance judgments.

6 METRICS

The primary metric for these collections is normalized discounted cumulative gain (nDCG) [4], with the gain values set as follows:

Relevance category	Gain
irrelevant (0)	0
topical (1)	0
specific-info (2)	4
direct-answer (3)	8
decisional (4)	20

These gain values were selected arbitrarily to give disproportionate weight to "decisional" documents. Average precision is also a reasonable metric for the phase 2 and 3 collections, where the deeper pools provide better assurance of completeness. For precision-based metrics, "specific information" (2) should be the minimum relevance level.[8]

Figure 4 shows the minimum meaningful score differences for MAP, P@10, and nDCG. These are computed using NIST's implementation of Sakai's discriminative power [7] and taking the value at the full topic set and a 5% error rate. For phase 3, this difference is with respext to the multilingual ranking. Score differences of less than these figures are not meaningful according to this procedure. These scores are computed from the evaluations of the BETTER performers, and the range in scores for those systems is quite close, so these minimum differences may be overestimates.

Since the BETTER program emphasized the combination of IR and IE, we developed a special metric integrating the two, based on the diversity metric α-nDCG.[2] In this metric, if a document addresses a known subtopic, gain is accumulated for that subtopic, and gain for that subtopic is discounted for subsequent documents that cover the same subtopic. BETTER subtopics were defined as follows: recall that during topic development, the example passages for tasks and requests were annotated according to the Basic schema. Each Basic event was considered a "subtopic". The extraction subtopics were called "critical extractions" in the sense that they contained information that the searcher expected to find.

In giving gain for retrieving a relevant document, the metric takes into account the presence of matching annotations in the highlight passage of the relevant document. Hence, retrieving documents which capture more of the Basic events specified in the topic cover more of the subtopic range, whereas retrieving more of the same events earns decreasing gain. We did not have the resources to manually align the annotations between the task/request examples and the relevant sections, so matches were made on the basis of event type, the type(s) of any referred events, and whether the events are states-of-affairs or not. The original definition of α-nDCG assumes that each document only addresses one subtopic, whereas our implementation allows credit for multiple matching extractions in a single document. As this metric is quite new and purpose-built for the BETTER program, we do not analyze it further in this paper, but the implementation of the measure is included with the datasets.

[8]The trec_eval incantation is `trec_eval -M1000 -q -l 2 -mall_trec -mndcg.0=0,1=0,2=4,3=8,4=20`. For programmatic reasons, the BETTER evaluation used a measurement depth of 100 (`-M100`).

7 REUSABILITY

Are the BETTER IR collections reusable? A common technique to measure reusability is to hold a group's unique contributions to the pool out of the relevance judgments, and then re-score the group's runs with these diminished relevance judgment sets.[10]

The results of holding a group out are shown in Figure 5 as the minimum and maximum score difference for each metric when a group's unique contributions the pool are held out. We would consider these differences quite small, except for P@10 in the phase 1 collection and nDCG in the phase 2 collection. The largest differences are for the phase 1 collection, which is reasonable as it was only pooled to depth 10. Since the BETTER program only had a few performing organization, they had more potential to find unique relevant documents. We recommend that when comparing systems using these test collections, if there are many unjudged documents retrieved above the pool depth of the collection, one should consider differences between systems that fall in these ranges to be insignificant.

This technique is not foolproof, since in the real reuse scenario we may be measuring a run that is quite different than those that were pooled. In the experiment, we have taken runs with complete coverage above the pool depth and pretended that they have no coverage there, except for what was found by other systems. In this case we recommend reviewing the unjudged documents retrieved in the top 10 ranks, but beware of confirmation bias: we tend to believe that documents our system retrieves are relevant.

8 RECOMMENDATIONS ON USE

The BETTER IR collections are new datasets for cross-language IR in Arabic, Farsi, Russian, Chinese and Korean. The Arabic collection may suffer from some incompleteness as procedures were refined for the later collections. We have presented guidelines for common metrics given the systems that participated in BETTER. They can reasonably be used to run standard CLIR experiments for these languages. Although there is no obvious candidate for a provided short-query condition, the `req-text` field is similar to the TREC description field.

Moreover, the Russian, Chinese and Korean collections have a common set of topics (in the BETTER parlance, analytic tasks and requests) and judgments across all three languages, making the phase 3 collection fully multilingual.

The structuring of requests into analytic tasks may be of interest for research on query context or sessions.

The tasks and requests include example relevant passages in English, and so might be of interest in feedback experiments. Since the examples do not come from the target collection, no special care (such as using a residual-collection evaluation approach) is needed when using the passages for feedback.

BETTER explored several unusual paths for IR experiments. The most unusual is ranked search by example. This was a challenging condition for BETTER performers since the passages contain many distracting terms and are presented to the system without context. While query-by-example has been studied in IR and particularly in the TREC routing and filtering tasks [8], many of those experiments are quite old now and may bear re-examination in light of modern user tasks.

The human-in-the-loop condition as implemented in the program evaluation enforced a strong separation of the users from the target data. Researchers might take a look at closing that gap. The evaluation task took advantage of having related requests in a task to offer "training" and "test" requests on a common topic, which may be useful for interactive experiments.

Lastly, the goal of the program of incorporating retrieval and extraction in a tight cycle still has a lot to be explored. Having event annotations on query examples and in retrieval results offers some unique possibilities for feedback term selection in a pure retrieval context. The tight turnaround window in the BETTER program evaluation kept performers from doing significant amounts of extraction inference during retrieval, and so that is an under-explored topic waiting for research.

The open nature of the CommonCrawl source of the document collection presents an issue that the content is likely to have been incorporated into the large language models used in most modern retrieval and extraction systems. So in a sense using a language model postdating the collection is equivalent to using a later snapshot of Wikipedia as training data — the answers are already in there. This is an unresolved research issue larger than this paper can address.

9 AVAILABILITY

All the BETTER program datasets including the retrieval collections are available at https://ir.nist.gov/better/. All collections are free of cost, being developed on content from the CommonCrawl dataset. The search tool used for developing topics can be found at https://github.com/isoboroff/bench, and the assessment tool at https://github.com/isoboroff/assess-react.

The retrieval collections are each arranged into a structure with directories for "corpora", "annotations", "tasks" and "scripts". The corpora are in JSON-lines format. The tasks directory contains the analytic tasks and requests. Relevance judgments and annotations can be found in the annotations directory. The scripts directory includes evaluation scripts, including the implementation of the BETTER α-nDCG metric (in `eval-better-ir.py`).

10 ACKNOWLEDGMENTS

This work was supported by the IARPA BETTER program. We gratefully acknowledge the many contributions from John Beieler, Carl Rubino, Tim McKinnon, and the BETTER performer teams. Additionally, we are also grateful for the assistance of Allison Powell, Leland Vakarian, and Marc Vilain at MITRE and Aric Bills and Emily Lord at ARLIS, and others from both institutions who have moved on during the course of the program. Lastly, we thank Dawn Lawrie, Jim Mayfield, Cash Costello and Paul McNamee of the JHU HLT Center of Excellence for help with Patapsco and machine translation.

11 DISCLAIMER

Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

[1] John Beieler. 2016. Generating Politically-Relevant Event Data. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, Austin, Texas, 37–42. https://doi.org/10.18653/v1/W16-5605

[2] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 659–666. https://doi.org/10.1145/1390334.1390446

[3] Petra Galuščáková, Douglas W. Oard, and Suraj Nair. 2021. Cross-language Information Retrieval. https://doi.org/10.48550/ARXIV.2111.05988

[4] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (oct 2002), 422–446. https://doi.org/10.1145/582415.582418

[5] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldani, and Eugene Yang. 2022. Overview of the TREC 2022 NeuCLIR Track. In *Proceedings of the 31st Text REtrieval Conference (TREC 2022)*, Ian Soboroff (Ed.).

[6] Timothy Mckinnon and Carl Rubino. 2022. The IARPA BETTER Program Abstract Task Four New Semantically Annotated Corpora from IARPA's BETTER Program. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 3595–3600. https://aclanthology.org/2022.lrec-1.384

[7] Tetsuya Sakai. 2006. Evaluating Evaluation Metrics Based on the Bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, Washington, USA) *(SIGIR '06)*. Association for Computing Machinery, New York, NY, USA, 525–532. https://doi.org/10.1145/1148170.1148261

[8] Ellen M. Voorhees and Donna K. Harman (Eds.). 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.

[9] Ilya Zavorin, Aric Bills, Cassian Corey, Michelle Morrison, Audrey Tong, and Richard Tong. 2020. Corpora for Cross-Language Information Retrieval in Six Less-Resourced Languages. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*. European Language Resources Association, Marseille, France, 7–13. https://aclanthology.org/2020.clssts-1.2

[10] Justin Zobel. 1998. How Reliable Are the Results of Large-Scale Information Retrieval Experiments?. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 307–314. https://doi.org/10.1145/290941.291014