# The BOLT IR Test Collections of Multilingual Passage Retrieval from Discussion Forums

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

Kira Griffitt and Stephanie Strassel
Linguistic Data Consortium
Philadelphia, PA
kiragrif@ldc.upenn.edu, strassel@ldc.upenn.edu

## ABSTRACT

This paper describes a new test collection for passage retrieval from multilingual, informal text. The task being modeled is that of a monolingual English-speaking user who wishes to search discussion forum text in a foreign language. The system retrieves relevant short passages of text and presents them to the user, translated into English. The test collection contains more than 2 billion words of discussion thread text, 250 queries representing complex informational search needs, and manual relevance judgments of forum post passages, pooled from real systems. This information retrieval test collection is the first to combine multilingual search, passage retrieval, and informal online genre text.

## 1. INTRODUCTION

The DARPA Broad Operational Language Translation (BOLT) program was "aimed at enabling communications with non-English-speaking populations and identifying information in foreign-language resources."[2] As part of the program, performers were tasked to complete a multilingual passage retrieval task: in response to a textual query in English, retrieve relevant passages from English, Arabic and Chinese discussion forums, translated into English.

To evaluate this task, we constructed a test collection of 2 million discussion forum threads totalling nearly 70 million posts and 250 search topics with free-text information need statements. The information needs represent users looking for information about, relationships between, effects of, and opinions about current events and socially prominent subjects. Systems participating in the BOLT program retrieved short passages in response to each query and translated them into English. These passages were manually judged for relevance by bilingual speakers with access to the original documents in their original language. The first 100 topics have an average of 185 judged passages each, and the latter 150 topics have an average of 486 judged passages. The combination of multilinguality, passage retrieval, and informal online discussion text makes this test collection unique.



*SIGIR '16, July 17 - 21, 2016, Pisa, Italy*



ACM ISBN 978-1-4503-4069-4/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2911451.2914674

Passage retrieval has been extensively studied in information retrieval [4], information extraction [9], and question answering [7]. As shown in the TREC HARD track [1], passage retrieval tasks can be complicated to specify. They also present a challenge to reproducibility: unless systems retrieve the exact same passages as were judged, interpreting passage-level relevance judgments is not straightforward.

## 2. EVALUATION TASK

The BOLT IR task is to retrieve short relevant passages of text from informal text in response to a natural language English sentence representing a complex information need. We imagine an English-speaking intelligence analyst searching across multiple languages. The analyst is looking to survey diverse views and different relationships among people, both document authors and people mentioned in the documents. Because the documents in the collection are in multiple languages, passages in languages other than English must be translated to English for the end user.

We have found it helpful to describe this task as an imaginary interface for the analyst that presents retrieved snippets in context; the user can select a snippet to drill down to the full document, and switch back and forth between the translated text and the original (presumably with a native speaker or linguist assisting). Because the user views the retrieved passages in context, it is not critical that passages exactly contain just the relevant information, nor do they need to disambiguate or coreference things mentioned in the passage. Because the native language is one step lower in the drill-down, translation quality is inherent in the task and affects the interaction. This perspective was reflected in the BOLT IR evaluation which did not require passages to contain only relevant content, and which measured retrieval effectiveness separately from translation quality. Although we do not report details of the evaluation in this paper, these considerations are important for understanding the relevance assessments.

## 3. DOCUMENT COLLECTION

The document collection is a set of two million online discussion forum threads collected by the Linguistic Data Consortium (LDC) containing 2.3 billion words of text. The threads are available in the original HTML and also in a cleaned XML format which was used in all stages of building the collection. The threads come from a number of different forums on different subjects. The threads come from three different identified language sources: English, Arabic, and Mandarin Chinese. The Arabic subcollection is intended to

| Language | English | Mandarin | Arabic |
|---|---|---|---|
| number of tokens | 694,914,443 | 1,050,254,101 | 616,719,471 |
| number of threads | 457,970 | 789,077 | 773,861 |
| average messages per thread | 13.23 | 50.77 | 30.82 |
| average thread size in bytes | 14,651.2 | 5,769.3 | 11,380.5 |

| Topic set | P2 | P3 |
|---|---|---|
| number of topics | 100 | 150 |
| avg relevant passages per topic | 114.56 | 252.46 |

**Table 1: Statistics on the BOLT IR collection. In Mandarin, 1.5 tokens = 1 word, otherwise 1 token = 1 word.**

target the Egyptian Arabic dialect, but the posts often contain Modern Standard Arabic and/or other Arabic dialects.

From a total forum crawl of roughly 3 billion words, LDC selected a subset of roughly 700 million words from each language to form the final collection. Table 1 gives statistics on the size of the collection.

# 4. TOPICS

The user's information needs that give rise to the queries are called "topics". The topics were developed by LDC annotators, which included native speakers of Mandarin and Egyptian as well as English speakers. The topics were developed using a collection exploration process: the annotator thought of a topic and did preliminary searches to determine whether the topic had any relevant information in the collection. The annotator then formulated a query for their topic, and surveyed the relevant threads that were returned by a basic text search tool. Based on this survey, they made a qualitative judgment about the prevalence of relevant information in the corpus, and developed topics where their judgment was that the topic was not likely to be highly productive for the corpus. As part of the query, the annotator could indicate that they only wanted information in Arabic, Chinese, or English, or would accept results from any language. This factor was a constraint on relevant information.

After arriving at a right-sized topic, the annotator wrote a textual description and rules of interpretation for assessing relevance. The topics were classified according to two informal taxonomies: the subject of the topic ("asks-about", which could be a person, location, organization, movement, event, or abstract entity, ...), and the kind of information desired ("asks-for": which could be statements or opinions, relationships, effects, information about, participants or members of, ...). The annotator also noted from which languages they expected the majority of relevant information to come, based on their searching. In contrast to the language-target restriction in the query, this expectation was not a restriction on relevance; relevant information could come from any language, as long as the query allowed it. Lastly, the annotator included one or more examples of relevant passages from their search. Figure 1 shows an example topic. Systems only had access to the topic number, the `query`, and the `language-target` information, but the complete topic statement was developed ahead of time and used for reference during assessment.

There were three BOLT IR evaluations over the course of three phases of the program. The first phase was highly exploratory, whereas the latter two, which we call P2 and P3, resulted in comparable collections following roughly the same rules for topic development and relevance assessment. In phase 3, topics tended to have more relevant passages due to the pooling procedure used (see below), but they actually have a greater variance in relevance because effort was put towards creating topics in P3 with small ($< 100$ passages) relevant sets.

# 5. POOLING

In response to the topic, evaluation systems automatically retrieved a ranked list of up to 1000 citations from the document collection. A citation is defined to be an English passage of no more than 250 characters, with a pointer back to the original source text. In Figure 1, the `thread`, `post`, `offset`, and `length` is the pointer for first example citation. System outputs followed this format, with threads and posts counting from 1 and offsets starting at 0 within the cleaned XML formatted collection. Citations were required to be English, and so Arabic or Chinese passages needed to be translated before they could be returned to the user; systems conceivably could do this at any time prior to returning results, including automatically translating the entire corpus at the start. Citations could not cross post or thread boundaries.

We combined the top 100 citations from each evaluation system to form a pool.[8] Pooling usually drops duplicated documents, but in this case, we planned for the assessors to judge the citation text as well as the pointer, so many close duplicates needed to be kept in the pool. To make it easier for the assessors to be consistent across near-duplicate cases, we combined all citations with greater than 95% token bigram overlap into equivalence classes. Neither near-duplication nor relevance are actually transitive, so assessments made to equivalence classes were hand-checked to make sure we didn't propagate judgments too far.

The pooling method can only be as effective as the diversity of the constituent system outputs, and indeed in phase 2 we were concerned that because the evaluation only has a limited number of participants, the estimates of recall we computed were biased high. For phase 3, we solicited a range of system outputs from each evaluation participant, including baselines which would not be measured but which served to enrich the pool. This resulted in nearly twice the number of relevant citations compared with phase 2. Because this occurred despite the aforementioned effort to limit the number of relevant citations in phase 3, we are hopeful that the phase 3 relevance judgments represent a more reusable collection with better recall estimates. Section 7 discusses recall further.

```
<topic number="BIR_300158">
   <query>What do people say about self-publishing?</query>
   <description>This query asks for statements and opinions about self-publishing.</description>
   <language-target lang="none"/>
   <properties>
      <asks-about target="practice-or-custom"/>
      <asks-for response="statements-or-opinions"/>
      <languages eng="T" arz="F" cmn="F"/>
   </properties>
   <rule number="1">Answers must be about self-publishing.</rule>
   <cite number="1" thread="bolt-eng-DF-275-201910-15807137" post="5" offset="133" length="125" rel="yes">
    Meaning that if they ask you to pay them, you're also getting scammed (or self-publishing, which also
    tends to be a scam...).</cite>
</topic>
```

**Figure 1: An example topic from phase 3. Systems were presented with the query and language-target clauses.**

## 6. ASSESSMENT PROCESS

Rather than making a simple relevance judgment for each citation, the assessor followed a decision-points annotation model [5] derived from the application scenario described in section 2, in order to tease apart relevance from translation quality. [3] First, the assessor decided whether the citation was clear enough to make a relevance judgment, or if they felt they probably understood the citation but wanted to refer to the original language text, or if the citation was completely incomprehensible. Incomprehensible citations were not assessed further.

For comprehensible citations, the assessor judged whether the citation satisfied the topic's rules of interpretation for relevance, and further whether the citation provided useful information to the user. Citations were judged to be not useful if for example they merely restated the query. If the assessor needed to refer to the original language citation, either because they wished to see greater context or wanted to confirm their understanding of the translation, they decided whether the original Mandarin or Arabic text was relevant and useful as above, and then returned to the translated citation to make a judgment of relevance for the citation itself. In phase 3, the assessor further noted whether they felt they were being generous in their judgment.

This multi-stage process allowed the assessor to separately consider translation acceptability from relevance. Resources did not permit us to do deep assessment of translations, as might be expected of an MT corpus, but by folding translation acceptability into the relevance assessment process, it is possible to subset the data to use only the most acceptable translations or to allow relevance within the original citation context.

In the BOLT evaluation, systems received credit for citations that were understandable, relevant and useful, even if the source language needed to be checked to make that judgment. Having these fine-grained assessments with a rule for giving credit means that later users of this data can choose to give credit for stricter or looser scenarios.

We presented a topic's citation pool to the assessor in random order to avoid system bias effects. A single citation was judged for all citations in a near-duplicate equivalent class, and equivalence classes were checked for consistency in a second pass.

The assessor judging the citations for a particular topic was not necessarily the topic's creator. In particular, citations that came from Mandarin or Arabic were assessed by a native speaker. A subset of the topics were selected to be fully assessed by a second assessor.

In phase 3, the assessor additionally noted whether the source text for Arabic citations included Egyptian Arabic or not. This annotation was made following the realization that Egyptian writers frequently code-switch between dialectal and standard Arabic, and that the line between a passage being in Modern Standard Arabic or Egyptian dialect was hard to draw. Although the assessor did not mark the code-switch boundaries, we hope that this data might be useful for training systems to differentiate the two dialects of Arabic.

## 7. NOTES ON USAGE

The BOLT IR test collections are straightforward to use to measure the effectiveness of post and thread retrieval. Measuring passage retrieval is more complex, because retrieved passages may not directly correspond to the passages that were judged by our assessors. Also note that the assessor did not judge the source language text when the translated citation was sufficient. In general, the reusability of passage-level assessments is not well understood and deserves more study.

A more significant issue is assessment coverage and recall estimation. Because our pools were limited to DARPA contractors being evaluated on the IR task, it's likely that there are many relevant posts that were not pooled and hence not judged. Experimenters should be cautious in handling unjudged retrieved posts, especially for systems that do find judged relevant posts around those same ranks. Retrieval measures that are robust to missing judgments, such as those proposed by Sakai [6] or Yilmaz et al [10], may be useful here.

## 8. BASELINE PERFORMANCE

While this paper is not an evaluation report, we would like to give an indication of reasonable baseline performance which may be expected on this collection. We do this here by anonymizing the BOLT submissions and reporting precision and recall of retrieved forum posts.

The official measures in BOLT included variants of precision and recall based on character counts for citations. Since

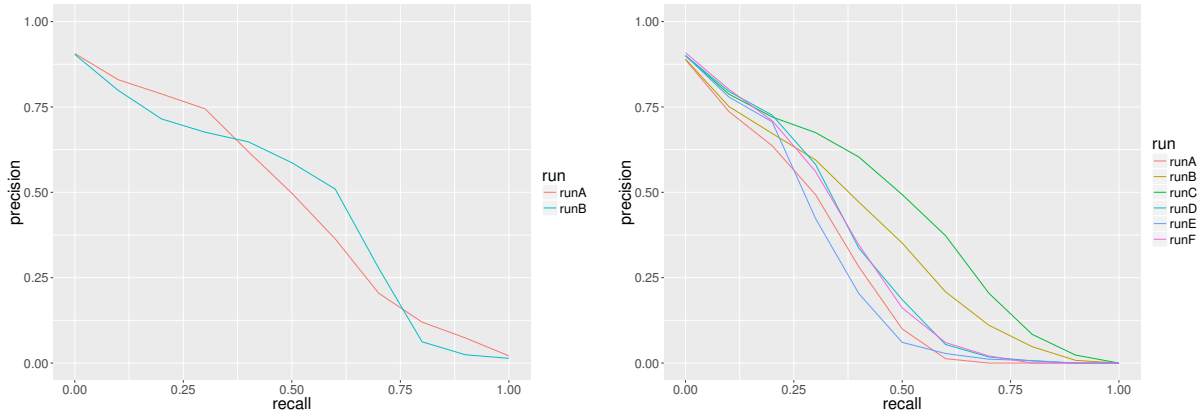


**Figure 2: Recall-precision plots for post retrieval performance of several anonymized submissions in the BOLT program, phase 2 (left) and phase 3 (right).**

as we note in the previous section, reusing this collection for passage retrieval evaluation requires some significant effort, we convert the original BOLT submissions to simpler post retrieval runs by ranking posts according to the highest ranked passage retrieved from that post. A post is considered to be relevant if any relevant passage is known from that post. Because all these systems were pooled, we are not affected by unjudged posts.

Please note that these systems were not optimized for this measure, and while these measures are reasonable to compute for this collection, they are not the best comparisons of these specific runs, which were developed to solve the passage retrieval task rather than the post retrieval task. The purpose of these scores is solely to provide indicators of reasonable state-of-the-art performance on a post retrieval task using this data. Figure 2 plots interpolated precision at standard recall points, as reported by the `trec_eval` tool[1]. Mean average precision for the P2 runs is 0.45 and 0.46, and for the P3 runs ranges from 0.25 to 0.42.

## 9. CONCLUSION

The BOLT IR collections, built as part of DARPA's Broad Operational Language Translation program, are new test collections that can be used to measure multilingual retrieval from informal discussion forum text. Furthermore, as the annotations are at the passage level, the collection provides a basis for considering how to create reusable collections for passage retrieval evaluation. The corpora described in this paper have been distributed to performers in the DARPA BOLT program, and are expected to be published in LDC's catalog in 2016.

Although three collections were built over the course of the program, we only consider the latter two, from phases 2 and 3, to be reusable as IR test collections. The collection built for phase 1 represented a learning curve in designing the passage retrieval task and assessment process, and while that data is likely quite useful to researchers it should be regarded with caution as an IR test collection.

Finally, this paper has been submitted as part of a new SIGIR call for short papers on datasets and test collections, and as such we hope it can be regarded as a proposed template for such papers. In particular, we recommend that papers describing IR test collections provide specific guidance on usage and report baseline effectiveness scores.

[1]https://github.com/usnistgov/trec_eval/

## 10. REFERENCES

[1] J. Allan. HARD track overview in TREC 2005: High accuracy retrieval from documents. In *Proceedings of TREC 2005*, pages 52–68. NIST, 2006.

[2] BOLT program home page. http://www.darpa.mil/program/broad-operational-language-translation, retrieved on Feb 1, 2016.

[3] K. Griffitt and S. Strassel. The query of everything: Developing open-domain, natural language queries for BOLT information retrieval. In *Proceedings of LREC 2016*, Portorož, Slovenia, 2016.

[4] M. Kaszkiel and J. Zobel. Passage retrieval revisited. In *Proceedings of SIGIR 1997*, Philadelphia, PA, 1997.

[5] J. Medero, K. Maeda, S. Strassel, and C. Walker. An efficient approach for gold-standard annotation: Decision points for complex tasks. In *Proceedings of LREC 2006*, Genoa, Italy, 2006.

[6] T. Sakai and N. Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, 11(5):447–470, March 2008.

[7] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Martin. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of SIGIR 2003*, Toronto, Canada, 2003.

[8] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[9] W. Xu, R. Grishman, and L. Zhao. Passage retrieval for information extraction using distant supervision. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, 2011.

[10] E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of SIGIR 2008*, Singapore, 2008.