BOLT IR task, phase 3
Evaluation guidelines

(version 2.5, September 30, 2014)

Summary of changes from Phase 2

- Citation retrieval task unchanged in format and document set.
  - Topic development will strive to make topics with fewer than 100 relevant citations.
  - Performers will submit unusual and/or baseline submissions that will not be scored, but that will help expand the pool.
  - Some changes in metrics will improve recall estimation.
- No grouping task.
- A new interaction task will allow systems to automatically solicit clarifying feedback from the user after the initial query and retrieval.

Task Overview

The user has a complex information need and a collection of informal documents (in this case, forum posts) where the answers may be found. The user formulates and issues an ad hoc, natural language query in the form of a single English sentence. The system returns a list of relevant, short citations of text with pointers back to their origin location in the document collection. These passages are assigned into groups where each group addresses a different aspect of the topic.

This task models a real-life intelligence analysis scenario where the analyst is confronted by informal textual sources of a social nature, and would like to study relationships among people involved in the discussion, points of view expressed regarding a specific event, and their relative weight or frequency.

For example, imagine the analyst is looking to study the range of opinions expressed about the Israel-Gaza war in October 2012. The analyst might initially expect to find expressions of pro-Hamas and pro-Israel sentiment, and indeed the initial retrieval finds numerous passages taking both sides of the conflict. Additionally, there are expressions of general ennui with respect to the Palestinian-Israeli conflict, theories that the conflict was staged to coincide with and influence the U.S. elections, and formal international responses which should really be considered separately from public opinions. In the end the analyst also constructs a catch-all group containing relevant passages that don’t fit into the other groups, but are interesting enough to want to keep.

To accomplish this task, systems must retrieve relevant citations from documents. The task will be split into a **retrieval** subtask and an **interaction** subtask. In the retrieval subtask, systems will return short, relevant citations in response to the query. The system responses will be pooled and judged by an assessor at LDC. In the interaction subtask, systems will conduct a single-round interaction with a user to clarify, refine or disambiguate the user's search intent.

Document collection

The collection for phase 3 is a large set of online discussion forum threads collected by the LDC. The threads are available in the original HTML and also in a cleaned XML format. The threads are not a single holistic collection but rather come from a number of different forums on different subjects. The threads come from three different identified language sources: English, Egyptian Arabic, and Mandarin Chinese. The posts in each thread may contain text in other languages (MSA, for example).

The entire forum collection comprises roughly 3 billion words of text. For phase 2, LDC identified a subcollection of roughly 700 million words from each language source. For phase 3, we will use this same subcollection. Teams may perform generic preprocessing of the collection, including machine translation, up to when the evaluation topics are released. The preprocessing activities must be documented in a form that ensures reproducibility and that records wall-clock time and resources used. An example form would be a Unix Bourne shell script containing commands that were run over the collection, and comments inline indicating how long each stage took to process.

Teams may use the phase 1 and 2 topics, relevance judgments, and groupings to train their systems. They should keep in mind that the phase 1 task was somewhat different, that phase 1 only used 400 million words of the collection per language. Because of evaluation issues as well as system performance in phase 1 which affected pooled assessments, the phase 1 data probably is not optimal as training data.

Teams may also annotate reasonable portions of this data, but teams **must share all annotations** in a documented format. Comprehensive, corpus-wide annotation is not permitted. If it is infeasible to share an annotation set, for example because it is intimately tied into the details of the system, NIST will allow exceptions upon detailed request. Training and annotation may be done up until the evaluation topics are to be released (see schedule below).

Topics

Topics describe the information need of the user, including example relevant citations found during topic development and any rules of interpretation required for performing relevance judgments. At evaluation time, teams will only receive the 'query' and 'language-target' fields of the topics; the full topic content will be released with the relevance judgments.

Topic format

```
<bolt-ir-topics eval=”BOLT-IR-P3” contact=”Ian Soboroff
ian.soboroff@nist.gov”>
<topic number=”BIR_300001”>
<query> The query sentence. </query>
<language-target lang=”arz”/> <!-- cmn, eng, arz, or none -->
<!-- the fields below will not be available until the conclusion
of the evaluation -->
<description>
A short description of the information desired by the user and
its important facets. The description presents the user’s task
as embodied by this topic, and is the basis for and governs the
rules of interpretation.
</description>
<properties>
  <asks-about target=”abstract-entity”/>
  <asks-for response=”statements/opinions”/>
  <languages eng=”T” arz=”T” cmn=”F”/>[1]
</properties>
<rules>
Any formal rules of interpretation, as identified by the topic
creator, which determine how to judge relevance for this topic.
</rules>
</topic>
…
</bolt-ir-topics>
```

There will be approximately 100 new evaluation topics targeting combinations of three experimental conditions of interest:
1. relevant information found in a single language vs. in multiple languages.
2. topics explicitly targeted at a specific language (indicated by the language-target tag).
3. different topic types.

During topic development, LDC will search the document collection manually for relevant citations. This is not intended to collect complete citations, but to scope the topics to a limited expected relevant set size, to provide a manual-search perspective on relevant citations, and to expand the total relevant set beyond those retrieved by the systems.

[1] The “threads” tag used in phase 1 has been dropped. All topics in phase 1 were multi-thread.

LDC will attempt to develop a subset of topics that have 100 relevant citations or fewer. Because this is impossible to ensure without exhaustive searching of the collection (and the LDC will not be exhaustively searching the collection), there will likely be some topics with more than 100 relevant citations.

Topic types

Topics for the BOLT IR evaluation will fall into several categories. However, unlike topics in GALE, they do not follow a template format for the query. Each query will include a “properties” section defining the types.

```
<asks-about target=”abstract-entity”/>
```

Target can be:

- person
- location
- organization
- movement
- event
- abstract entity (belief, ideology, ...)
- etc.

```
<asks-for response=”statements/opinions”/>
```

Response can be:

- statements or opinions about
- relationships between
- effects of
- information about
- participated in
- etc.

The above two sections serve to organize the query set for purposes of averaging scores among common conditions. They do not supersede the natural language query. Only a subset of the target/response possibilities will be contained in the evaluation topic set.

```
<languages eng=”T” arz=”F” cmn=”F”/>
```

The languages tag indicates where the user creating the query expects relevant citations to be found. Note that this is not a definitive statement that relevant information is **only** found in those languages, just that based on the relevance judgments, these are the languages represented. “eng”, “arz”, and “cmn” refer to

English, Egyptian Arabic, and Mandarin Chinese as they are denoted in the LDC data distributions. Values are either “T” (for true; relevant information is expected to be found in this language) or “F” (for false).

(Note this is different from the <language-target> tag that follows the query, which indicates whether the user is only interested in responses from a certain language.)

NIST and LDC will provide several example topics meant to be illustrative of the topics in the evaluation set. Teams should not assume that the example topics fully cover the space of topics planned.

Subtask 1: Citation Retrieval

Teams will receive the <topic number=”X”> , <query>, and <language-target> portions of the phase 3 evaluation topics, and will return a ranked list of at most 1000 translated passages per topic from the document collection. Each passage may be at most 250 characters in length. Each passage will include a pointer to its span of origin in the XML version of the forums collection. This combination of a short passage with its source pointer is called a “citation”.[2]

The format will be as follows (see the appendix for an authoritative DTD):

```
<bolt-ir-submission team=”my team name”
                    date=”the due date”
                    eval=”BOLT-IR-P3”
                    subtask=”citations”
                    run=”evaluation”
                    contact=”John Doe johndoe@foo.org”>
<response number=”BIR_300001”>
  <cite score=”0.876” thread=”thread-id” post=”post-id”
  offset=”start” length=”in chars”>
  This is a translated passage of some post in some thread.
  </cite>
  <cite score=”0.82” thread=”thread-id” post=”post-id”
offset=”start” length=”in chars”>
  This text came from another post in another thread.
  </cite>
  <cite score=”0.5” thread=”thread-id” post=”post-id”
```

[2] Occurrences of the word “passage” or “bullet” in this document, except where the context implies the text passage or the unit of retrieval in phase 1 respectively, should be read as “citation”.

```
     offset=”start” length=”in chars”>
  Each citation has a score between 0 and 1.  This score imputes
  a ranking over all citations for a response.  Tied scores will
  be broken arbitrarily.
  </cite>
  ...
</response>
<response number=”BIR_300002”
...
</bolt-ir-submission>
```

The “subtask” in the submission header should be “citations”. See below for information on the “run” field. Citation scores should be system generated and be numbers between 0 and 1 inclusive, where 1 is the highest possible score. Tied scores will be broken arbitrarily.

Offsets and lengths in the <cite> tag are in UTF-8 characters in the original, source-language post, in the XML version of the document collection, starting with the first character coming after the <post> tag as character 0. Offsets and lengths should count XML entities as untranslated characters (that is, for example, “&amp;” is five characters.) A citation points to a single citation in the corpus. The English text in a citation may not exceed 250 characters. During assessment, the assessor will be able to view passages in their original document context, so there is no reason to include context in the passage, for example to resolve pronouns.

Teams may return up to 1000 citations per topic. This is designed to solve a problem in phase 2 where systems were only measured to depth 100, and as a result could not achieve 100% recall for topics with more than 100 relevant citations. While we will only pool citations to depth 100 for assessment by LDC, systems that retrieve relevant citations deeper than rank 100 (that were pooled by other systems) will receive credit for them. Systems may decide to retrieve anywhere from 1 to 1000 citations for a topic; there is no requirement to return 1000 citations.

The top ranked 100 citations by score will be pooled for assessment. Some measures will give systems credit for retrieving known relevant citations anywhere in their ranking. In addition to the character precision, recall, and F measures computed at rank 100, we will compute character precision at rank R (where R is the number of relevant citations known for the topic), and character average precision to depth 1000. By using a measurement depth greater than the possible number of relevant citations, recall won’t be truncated as it was in phase 2.

Citations with duplicate or near-duplicate passage text from the same team should not be retrieved. The common example of where this is likely to occur is with quoted text in posts within a thread. Near duplication is defined as 95% symmetric overlap of word bigrams or greater. The highest-ranked citation in a near-duplicate class will be pooled and judged. The others will be marked as equivalently relevant, but will count as false alarms in the evaluation metrics. **BBN** provided (in phase 2) a standard implementation of the near-duplicate citation heuristic to ensure teams handle this case uniformly.

The phase 3 citation retrieval task has several “run conditions” that teams must participate in. Each run may contain up to 1000 retrieved citations per topic, but only the top 100 will be pooled.

Evaluation run: this is the part of the submission intended for official evaluation. The evaluation results for this run will be reported as team results in the final report. (In phase 2 this was the sole condition.)

Phase 2 run: each team must submit the results of running their phase 2 citation retrieval system on the phase 3 topics. The purpose of this run is to show system improvement from phase 2. (It is understood that the latest version of the phase 2 system may differ from the phase 2 evaluation run, since fixes may have been incorporated prior to system submission.)

Baseline runs: each team must submit at least one and up to ten “baseline” runs. Baselines will not be scored in the evaluation, but will be pooled to enrich the diversity of the assessment pools. Baselines can include simplistic approaches (e.g. tf-idf bag-of-words search using the query field terms), highly experimental algorithms (e.g. parameter settings you don’t want evaluated officially), collection subsets (e.g. run over detected Egyptian Arabic only), or indeed anything else. The goal of the baselines is to include *different* kinds of runs than the official evaluation run, so that the pool diversity is broader than just the evaluation run.

External resources. Teams may make use of external resources so long as those resources are either (a) openly available, or (b) teams commit to making them available to all teams by the training/annotation deadline.

Timing information. Please return to NIST In a separate file the wall-clock time elapsed per query, in seconds. Systems may not take more than 10 minutes to process any single topic.

Evaluation

Assessors will judge citations as relevant or not to the topic according to a three-point scale:

- **Yes**, the citation is clearly relevant to the assessor’s query.
- **Maybe,** the citation is on-topic but is of dubious utility.
- **No,** the citation is not relevant to the assessor’s query.

The LDC will judge all submitted citations, and half the topics will be judged by two LDC assessors. The LDC assessment process will additionally gather from the assessor whether an Arabic citation contains Egyptian; and whether a “yes” or “maybe” judgment was given generously.

The assessor will judge the citation with respect to its original context (since, as indicated above, passages need not be long enough to resolve references). If the assessor has trouble understanding the system-provided translation of the passage, they will refer to the citation in the source language. For “yes” and “maybe” relevant citations, the assessor will also indicate whether the translation is **acceptable** or **not acceptable**.

In phase 2, the LDC assessor had the ability to mark off extraneous parts of a citation using a <relspan> section, but in practice this was not used. LDC assessors will not mark <relspan>s in phase 3.

Official measures for this task will include:

1. Character precision, the fraction of characters returned marked “yes” or “maybe”, at cutoff rank 100.
2. Character recall, the fraction of “yes” or “maybe” characters out of all such citations judged by LDC (including those found by LDC during topic development), at cutoff rank 100.
3. An F-measure (harmonic mean) of the above precision and recall metrics, weighted equally.
4. Character R-precision, the fraction of characters returned marked “yes” or “maybe” at rank R = the number of known “yes”/”maybe” citations.
5. Character-based mean average precision at cutoff rank 1000.
6. Plots of precision at rank cutoffs and recall-precision plots.

NIST will provide a scoring script that computes these measures given a judged submission. As in phase 1 and 2, a number of other measures will be reported for diagnostic purposes.

Pilot

There will be a pilot of the citation retrieval task specifically aimed at testing development and system output for topics targeted at fewer than 100 relevant citations. LDC will develop a small topic set for the dry run and assess pooled system outputs to confirm topic design strategy.

Subtask 2: Interaction subtask

The goal of the interaction subtask is to extend citation retrieval with a single round of interaction with the LDC topic assessor. Each team's system will nominate a number of topics for interaction, and the assessor will interact with an interface designed by the team for each topic. Following the interaction the teams will submit a post-interaction run intended to improve retrieval performance based on the interaction.

The interaction process will be as follows. The evaluation topics will be released at the start of the evaluation period. Teams will then nominate topics for interaction. Topic nominations will be due at the same time as official evaluation runs. Topic nominations must be fully automatically performed by the system. Topic nominations will be ranked by team preference, as the LDC may not have the resources to interact with all nominated topics.

During an interaction period of one week, LDC assessors will interact with team UIs for the nominated topics. UIs must be accessible through a standard web browser to be specified. UIs will not be hosted at LDC but teams may host them at their own site or using a third-party platform such as Amazon AWS. Interaction UIs may not do simple relevance feedback (i.e., they may not show citations to the user and ask him or her to make relevance judgments for them).

The interaction period will allow for 30 minutes of training and warm-up time per assessor for each system. This training may be conducted remotely via videoconference and/or a shared desktop. Teams will need to have staff available to the LDC during the interaction period both to schedule training and to resolve system issues.

The LDC assessor will interact with the system for a maximum of two minutes per topic. Systems must enforce the interaction period limit, and alert the LDC user when the time is up. Teams may collect any available interaction data, implicit or explicit, during this interval. Interaction times will be recorded and reported with an emphasis on short interactions (i.e. change in effectiveness plotted against interaction time used).

After the interaction period, teams will submit both their gathered interaction data as well as a post-interaction evaluation run making use of the interaction data to improve retrieval effectiveness. The post-interaction run will process all phase 3 topics for which an interaction took place. The primary goal of the system interaction is to improve citation R-Precision for the nominated interaction topics, but we will also report the change in R-Precision over the entire query set.

Post-interaction runs will be included (to depth 100) in the assessment pools for the citation retrieval task.

Todo: determine a format for interaction data.

Evaluation

Evaluation of the interaction task is meant to illustrate improved effectiveness as a result of the interaction. We will report:

- change in R-precision.
- Delta R-precision as a function of interaction time used.

Schedule

February 20 Draft Evaluation Plan
June 2: relevance pilot

- Goal: test topic development for topics with <100 relevant citations
- NO INTERACTION!
- LDC will create 6 pilot topics (2 eng, 2 arz, 2 cmn)
- Topics will be sent to teams on 6/2
- Citations retrieval results due to NIST 6/4
- LDC will send relevance assessments to NIST and teams by 7/1
- Scores on pilot topics will be returned to teams by 7/17

Sep 29: interaction dry run

- Goal: test interaction setup with LDC assessors
- NO RELEVANCE ASSESSMENT!
- LDC will create 50 topics (40 E, 5 A, 5 C)
- topics will be sent to teams on 10/1
- topic interaction nominations due to NIST 10/3
- LDC assessors will interact with systems during the week of 10/6

Evaluation: Nov 10-26

- 100 evaluation topics sent to teams on Nov 10
- pre-interaction runs due to NIST Nov 12
- topic nominations for interaction due to NIST Nov 12
- interaction week: Nov 17-21
- post-interaction runs due to NIST Nov 26
- baseline runs due to NIST Nov 26 (can be sent earlier)

December 8 – February 5 assessment period at LDC
February 20 results returned to teams.

DTD for IR task topic files

```
<!-- DOCTYPE bolt-ir-topics SYSTEM "bolt-ir-topic-schema.dtd" -->

<!ELEMENT bolt-ir-topics (topic+)>
<!ATTLIST bolt-ir-topics eval CDATA #REQUIRED>
<!ATTLIST bolt-ir-topics contact CDATA #REQUIRED>
<!-- This is header information for a topic set.
     'eval' is an identifier for the evaluation. It should be
        BOLT-IR-P2.
     'contact' is a contact name and email address, for example,
        Ian Soboroff ian.soboroff@nist.gov
  -->

<!ELEMENT topic (query,description,language-target,properties,rule+)>
<!ATTLIST topic number ID #REQUIRED>

<!ELEMENT query (#PCDATA)>
<!ELEMENT description (#PCDATA)>
<!ELEMENT language-target EMPTY>
<!ATTLIST language-target lang CDATA #REQUIRED>
<!-- lang is “arz”, “cmn”, “eng”, or “none” -->

<!ELEMENT properties (asks-about,asks-for,languages)>
<!ELEMENT asks-about EMPTY>
<!ATTLIST asks-about target CDATA #REQUIRED>
<!ELEMENT asks-for EMPTY>
<!ATTLIST asks-for response CDATA #REQUIRED>
<!ELEMENT languages EMPTY>
<!ATTLIST languages eng (T|F) "F">
<!ATTLIST languages arz (T|F) "F">
<!ATTLIST languages cmn (T|F) "F">

<!ELEMENT rule (#PCDATA)>
<!ATTLIST rule number CDATA #REQUIRED>

<!-- <cite> elements in the topics are citations discovered by LDC annotators
during topic development, and are meant as examples only. CMN and ARZ
citations will be untranslated. -->
<!ELEMENT cite (#PCDATA|relspan)*>
<!ATTLIST cite id CDATA #REQUIRED>
<!ATTLIST cite thread CDATA #REQUIRED>
<!ATTLIST cite post CDATA #REQUIRED>
<!ATTLIST cite offset CDATA #REQUIRED>
<!ATTLIST cite length CDATA #REQUIRED>
<!ATTLIST cite rel value (yes|no|maybe) "yes">
```

DTD for IR task submission files (citation subtask)

```
<!-- <!DOCTYPE bolt-ir-submission SYSTEM "bolt-ir-schema.dtd"> -->

<!ELEMENT bolt-ir-submission (response+)>
<!ATTLIST bolt-ir-submission team CDATA #REQUIRED>
<!ATTLIST bolt-ir-submission date CDATA #REQUIRED>
<!ATTLIST bolt-ir-submission eval CDATA #REQUIRED>
<!ATTLIST bolt-ir-submission run CDATA #REQUIRED>
<!ATTLIST bolt-ir-submission subtask CDATA #REQUIRED>
<!ATTLIST bolt-ir-submission contact CDATA #REQUIRED>
<!-- This is header information for the submission.
     'team' is the name of your team.
     'date' is the date that you submitted it.
     'eval' is an identifier for the evaluation. It should be
        BOLT-IR-P2.
     ‘subtask’ identifies the subtask and should be ‘citations’.
     'contact' is a contact name and email address, for example,
        Ian Soboroff ian.soboroff@nist.gov
  -->

<!ELEMENT response (cite+)>
<!ATTLIST response number CDATA #REQUIRED>
<!-- The response number refers to the topic number in the topic file
  -->

<!ELEMENT cite (#PCDATA|relspan)*>
<!ATTLIST cite key CDATA #IMPLIED> <!-- generated by NIST for LDC during
pooling -->
<!ATTLIST cite score CDATA #REQUIRED>
<!ATTLIST cite thread CDATA #REQUIRED>
<!ATTLIST cite post CDATA #REQUIRED>
<!ATTLIST cite offset CDATA #REQUIRED>
<!ATTLIST cite length CDATA #REQUIRED>
<!ATTLIST cite original CDATA #IMPLIED> <!-- optional: original post text
from above pointer -->
<!ATTLIST cite rel (yes|no|maybe) "no">
<!ATTLIST cite chksrc (yes|no) “no”>
<!ATTLIST cite trans (accept|problematic|notaccept|na) “notaccept”>
<!ATTLIST cite dialect (ARZ|no-ARZ) "no-ARZ">
<!-- Passages must have a score between 0 and 1 inclusive.
     Teams should not include “rel”, “chksrc”, “trans”, or dialect attributes
in their submissions.
  -->
```