BOLT IR task, phase 1
Evaluation guidelines

(version 5.0, 15 Apr 2012)

Task

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. (Examples of such queries may be found in the "description" field of the TREC ad hoc topics.) The system returns a set of summary bullets of relevant information, where each bullet addresses one or more facets of the topic and contains a representative, extracted span of text with a list of citations (forum post and text offset) to the sources of evidence for that facet. Redundancy and false positives will be penalized by virtue of the application of a nugget-based metric that takes into account both Precision and Recall. (Redundancy will be treated as a false positive.)

Comment [BJD1]: This scoring approach discourages returning whole passages instead of returning non-redundant, tight coverage of nuggets.

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.

Document collection

The collection for phase 1 is a large set of online discussion forum threads. 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 are in three different languages: English, Egyptian Arabic, and Mandarin Chinese.

Snapshot release 1 of the data contains 13k English threads, 43k Arabic threads, and 49k Chinese threads. LDC plans four releases of the data. The final dataset for the IR task (R4) will be identified and released in late May.

Part of the document collection is excluded from the evaluation; teams may use this held-out subset for training their systems. The subset will be identified as part of the R2/R3 data releases. Note that this subset will not necessarily be representative of the collection as a whole, in topic, source, or style.

Teams may also annotate portions of this data, but teams **must share all annotations** in a documented format. 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 any rules of interpretation required for performing relevance judgments. At evaluation time, teams will only receive the ‘query’ and ‘response-length’ fields of the topics; the full topic content will be released with the relevance judgments when results are returned to teams.

For the BOLT IR task, topics have a collection of facets which system responses are expected to address. An extract of text from the collection is called a nugget. When the assessor develops a topic, they will identify the facets they want to see addressed in system responses, and at least one nugget matching each facet.

Topic format

```
<topic number=”1.001”>
<query> The query sentence. </query>
<!-- 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=”F” cnm=”F”/>
  <threads multiple=”F”/>
</properties>
<facet>
  <desc>A short description of a facet of this topic that system
responses should cover.</desc>
  <nugget post=”a-post-in-the-collection” offset=”35”>An extract
from this post that meets the assessors requirements.</nugget>
</facet>
<facet>
…
</facet>
<rules>
Any formal rules of interpretation, as identified by the topic
creator, which determine how to judge relevance for this topic.
</rules>
</topic>
```

Comment [IS2]: Since the measures now include both precision and recall, a maximum length limit on system responses is not needed.

Comment [IS3]: A facet is an aspect of the topic that the assessor expects system responses to cover. A nugget is an extract of text from the collection that meets the assessors expectation of relevance for this facet.

There will be 150-180 topics targeting combinations of three experimental conditions of interest:
1. relevant information found in a single language vs. in multiple languages
2. relevant information found in a single thread vs. across multiple threads
3. different topic types

Topic types

Topics for the BOLT IR evaluation will fall into several categories. However, unlike topics in GALE, they will 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. In phase 1, only a subset of the target/response possibilities will be contained in the evaluation topic set.

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

The languages tag indicates where relevant information is expected 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 “cnm” 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).

```
<threads multiple=”F”/>
```

The threads tag indicates the number of threads where relevant information is expected to be found. Note that this is not a definitive statement of whether multiple threads do or do not contain the relevant information, just that based on relevance judgments, the relevant information is contained in one or many threads. The value of the “multiple” attribute is either “T” (for true; information is expected to be found across multiple threads) or “F” (for false).

```
<facet>…</facet>
<desc>…</desc>
<nugget>…</nugget>
```

The facet tag describes an aspect of the topic that the assessor expects system responses to address. Each facet tag will have a description section (<desc>) and a nugget section (<nugget>). The description is a natural language description of what this facet of the topic is. The nugget is an example extract from the corpus that meets the assessor’s expectation of relevance. The assessor will not compile an exhaustive set of nuggets for each facet at topic development time, but rather during assessment, the assessor may identify different nuggets and additional facets for the topic based on system responses.

NIST and LDC will provide several example topics meant to be illustrative of the topics in the evaluation set. As with the training subset of the document collection, teams should not assume that the example topics fully cover the space of topics planned.

Results

Teams will return, for each topic, a set of summary bullets of relevant information, each with a list of citations to supporting source documents from the collection for that bullet. A source citation is the identifier of a post and a character offset into the XML version of that post. The goal of the system in producing the response is to cover as many facets of the topic as possible, in a concise fashion. The format will be as follows:

```
<result number=”1.001”>
<response>
  <bullet>
  Responses may contain many bullets.  The goal of a bullet is
  to cover one or more facets of the topic.
  <source post=”ians-email” offset=”22”/>
  <bullet>
  This text came from the post indicated in the source tag.
  <source post=”identifier-of-a-post” offset=”123”/>
  </bullet>
  <bullet>
```

Comment [IS4]: XML tags renamed with respect to the concerns raised by James and on the teleconference, and to reflect BOLT’s vision of the response as a bulleted list.

```
  Note that bullets may have more than one source, and sources
  can point to multiple places in a single document.
  <source post=”another-post” offset=”0”/>
  <source post=”another-post” offset=”37”/>
  </bullet>
  <bullet>
  The evaluation measures will cover both precision and recall,
  so systems must be judicious in their output.
  <source post=”another-another-post” offset=”25”/>
  </bullet>
</response>
</result>
<result number=”1.002”>
...
```

Notes:

- Responses will necessarily be of varying length, as some topics will have greater or fewer essential facets which the response bullets will need to cover.
- Result bullets must be cited back to the collection. No sources in external data resources are allowed. Offsets in the <source> tag are in characters in the encoding of the XML version of the document collection (expected to be UTF-8). Offsets need to be within three words of supporting information in the post, otherwise the citation will not be judged correct by the assessor.

Comment [BJD5]: This is meant to address the concern about how close the source citation offset needs to be to count.

Constraints

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.

Evaluation

Assessors will judge responses as addressing the facets of the topic:

1. During development of the topics, assessors will identify essential facets of information that responses must contain, along with sample extracts of text from the collection which address the facet (these extracts are called “nuggets”).
2. System response bullets must be relevant to one or more facets of the topic. The assessor will match bullets against facets of the topic identified during topic development.
3. System responses will be judged as covering a facet leniently (i.e., the response “President of the US” would be valid for a facet of the topic “Who is Barack Obama?”), even though it is only a phrase and the assessor may have identified the same information with a longer nugget text.

4. While reviewing responses, assessors may identify additional information topic facets based on system responses
5. The source forum post(s) will be judged as to whether they support the relevance of the cited information; at least one source must be relevant for the bullet to count. Citation precision and recall will be reported as a stand-alone measure as well as a component of some measures described below.
6. The assessor will identify the portion of relevant text in each bullet, following their own reasonable judgment as the putative user of the system. The assessor is not trying to find the minimal text that is relevant, or to remove unnecessary terms, but to strike out any significant portions of bullet text that are not responsive to any topic facet.

Comment [IS6]: This addresses the concern that systems may return entire posts or other overlong extracts in their responses. After assessment, we have three lengths: the length of the original response, the length of the edited response, and the average length of response text addressing this specific facet.

Matching of facets to `<bullet>` sections requires human judgments, thus, the evaluation outputs may not be automatically reusable on other response data.

The evaluation will report a range of measures to facilitate analysis. As a starting point, the primary measure will be the GALE distillation F-measure, which is the F_1 of facet precision and facet recall, where facet recall is scaled by a function of the F_1 for source citations:

Comment [BJD7]: The measures below came out of a conversation between DARPA and NIST sparked by the last teleconference, and a desire to have a balance between precision and recall.

$$F_{GD} = \frac{2P_nF_cR_n}{P_n + F_cR_n}$$

$$F_c = \sqrt[4]{\frac{2P_cR_c}{P_c + R_c}}$$

Facet precision (P_n) is the fraction of relevant <bullet> text in the response which match a previously unmatched facet and which are supported by a relevant citation. Note that portions of bullet sections that do not match a facet or that are redundant with a previously matched facet will negatively impact precision.[1] Facet recall (R_n) is the fraction of the topic's facets that are covered by relevant <bullet> section text. The length of a missed facet will be the length of the average nugget (as identified by assessors and systems) for that facet. The quality of source citations is captured by the F_c scaling factor.

Comment [BJD8]: This insertion plus the footnote below spells out the nugget-based notion of precision being applied here.

Another measure, ERR, was designed to handle redundancy and diversity in web search. ERR is a cascade measure, which means that it models a user reading the bullets in order and stopping when they have covered all the facets. With some slight modification, ERR fits the BOLT-IR task well:

Comment [BJD9]: The GALE measure balances precision, recall, and incorporates a penalty for redundancy. The web search community has developed a number of new measures for this identical problem, and further found that they correlate to user behavior. Thus, it would be worthwhile discussing alternative measures such as ERR.

$$Q_i^k = q_i^k \prod_{j=1}^{k-1} 1 - q_i^j$$

$$S_i = \sum_{k=1}^{K} Q_i^k$$

$$ERR = \sum_{i=1}^{M} p_i S_i$$

[1] Precision of a system answer will be measured in terms of the number of correct nuggets divided by the sum of the number of correct nugget matches and the number of "estimated incorrect nuggets." Estimated incorrect nuggets are computed by dividing the number of words in the redundant or irrelevant portion of the system response by the overall average number of words per nugget.

q(i,k) is the probability that a user interested in nugget i will be satisfied by the k’th bullet. The value of this will be the residual recall of relevant source citations, in other words, what fraction of correct uncited sources are brought by this bullet, scaled by the fraction of the bullet that is actually responsive. Q(i,k) then represents the probability that the user will stop reading at bullet k. The cumulative score for nugget i is the sum of the Q(i,k) values; in a ranked list, these would be discounted as the user reads further, but for BOLT, no discount is needed as the bullets are unranked. ERR is then the weighted sum across nuggets; in phase 1, all nuggets will have equal weight, so ERR is the average of the per-nugget scores.

Other measures may be reported as possibly informative for future phases. For example, precision of supporting citations; fallout of the summary (fraction of citations not relating to any nugget); number of citation sections (as a measure of length), and nugget redundancy (nuggets covered by more than one citation - increasing summary length but not nugget recall).

Examples

Here is an example topic about Bain Capital (from initial topic ideas and ASTRAL’s example):

```
<topic number=”…”>
<query>What did Bain Capital do in the 1990s?</query>
<description>
Systems should provide information on the major actions of Bain Capital in
the 1990s, including major acquisitions and personnel changes.
</description>
<properties>
  <asks-about target=”organization”/>
  <asks-for response=”information-about”/>
  <languages eng=”T” arz=”F” cnm=”F”/>
  <threads multiple=”T”/>
</properties>
<facet>
  <desc>Acquisition of Ampad from Mead Corporation.</desc>
  <nugget post=”bolt-eng-DF-199-192783-6834284” offset=”…”>One of these
transactions involved a company called Ampad, which Bain Capital purchased in
1992.</nugget>
</facet>
<facet>
  <desc>Acquisition of Experian</desc>
  <nugget>…</nugget>
</facet>
<facet>
  <desc>Actions taken to form GST Steel</desc>
  <nugget post=”bolt-eng-DF-170-181103-8882806” offset=”…”> Soon after, in
October 1993, Bain Capital became majority shareholder in a steel mill that
had been operating since 1888.</nugget>
</facet>
<facet>
```

```
  <desc>Mitt Romney takes a leave of absence to head the Salt Lake Organizing
Committee bid for the 2002 Winter Olympics</desc>
  <nugget>…</nugget>
</facet>
<rules>
According to Wikipedia, Bain was involved in XX leveraged buyouts and mergers
during the 1990s.  Any one of these can be a valid facet.
This topic is focused on the 1990s, so actions related to its founding, or
recent news about Bain relating to Mitt Romney, is not relevant.
</rules>
</topic>
```

The topic as composed by the assessor could be longer; in the course of their development research, they may want to identify as many facets as they can. Since assessor time is limited and only a few teams are participating, the notion of facet recall in this task is constrained to what the assessors and teams identify.

Here is an example system response (from ASTRAL's example):

```
<result number="…">
<response>
<bullet>
Soon after, in October 1993, Bain Capital, co-founded by Mitt Romney, became
majority shareholder in a steel mill that had been operating since 1888.
<source post="bolt-eng-DF-170-181103-8882806" offset="…"/>
</bullet>
<bullet>
One of these transactions involved a company called Ampad, which Bain Capital
purchased in 1992.
<source post="bolt-eng-DF-199-192783-6834284" offset="…"/>
</bullet>
</response>
</result>
```

This response has bullets that cover two facets out of four identified in the topic. The assessor decides that the bullet text is reasonable, and so does not identify any sections as not responsive to the facet. The fact that the sentences read out of context ("Soon after", "One of these transactions") does not penalize the response, because the assessor is not expecting a flowing summary and because the bullets are situated in the context of their originating posts.

This example is from a contributed example topic by IBM:

```
<topic number="…">
<query> What assistance was sent by other countries to Japan after the
earthquake?</query>
<description>
Systems should indicate what countries sent assistance to Japan following the
2011 earthquake, as well as what kind of assistance was sent.
</description>
<properties>
  <asks-about target="country"/>
  <asks-for response="relationships"/>
  <languages eng=”T” arz=”T” cnm=”T”/>
  <threads multiple=”T”/>
</properties>
<facet>
  <desc>The US delivered humanitarian aid and mobilized a large number of
military troops in Operation Tomodachi</desc>
  <nugget post=”bolt-eng-NG-170-181123-87085” offset=”…”> Speaking at a
Washington news conference, President Obama says the U.S. is marshaling
forces to help deal with the aftermath of the magnitude 8.9 earthquake in
Japan.</nugget>
  <nugget post=” bolt-eng-NG-170-181123-87085” offset=”…”> U.S. ships
carrying aid are en route, and the Air Force has delivered coolant to a
damaged nuclear plant.</nugget>
</facet>
<facet>
  <desc> Afghanistan city of Kandahar donated $50,000 to Japan</desc>
  <nugget>…</nugget>
</facet>
<rules>
According to Wikipedia at
http://en.wikipedia.org/wiki/Humanitarian_response_to_the_2011_T%C5%8Dhoku_ea
rthquake_and_tsunami, 47 countries sent some form of assistance; probably
only a subset are evidenced in the corpus; these should be listed as facets
here
</rules>
</topic>
```

Here is a sample response (also from IBM):

```
<result number=”…”>
<response>
      <bullet><source post=”bolt-eng-NG-170-181123-87085”
offset=”…”/>Speaking at a Washington news conference, President Obama says
the U.S. is marshaling forces to help deal with the aftermath of the
magnitude 8.9 earthquake in Japan.</bullet>
      <bullet><source post=”bolt-eng-NG-170-181123-87085” offset=”…”/>U.S.
ships carrying aid are en route, and the Air Force has delivered coolant to a
damaged nuclear plant.</bullet>
      <bullet><source post=”bolt-eng-NG-170-181123-87085” offset=”…”/>Japan
earthquake: Obama offers quake-ravaged Japan any assistance needed -
latimes.com</bullet>
      <bullet><source post=”bolt-eng-NG-170-181123-87085” offset=”…”/>Have
you heard anything on what China is offering? Yea, they've sent their
emergency response teams and doctors.... just like most other countries.
</bullet>
      <bullet><source post=”bolt-eng-NG-170-181123-87085” offset=”…”/>The
Brits who arrived back from NZ yesterday headed straight back out to Japan.
</bullet>
      <bullet><source post=”bolt-eng-NG-170-181123-87085”
offset=”…”/>Thousands of countries are helping. </bullet>
</response>
</result>
```

In this response, the first two bullets correspond to a single topic facet – the US – and so only the first would count as relevant for this response. The third bullet is not relevant. The fourth and fifth bullets are relevant to two other facets. The assessor might choose to trim the ending from bullet four (“just like most other countries”). The last bullet is not relevant to any facet (and is factually incorrect).

Schedule

- R1 data released
- guidelines and example topics released
- R2, R3 data released
- annotations delivered, external resources identified, training stops
- R4 data released
- evaluation topics released
- results returned to NIST/LDC
- NIST/LDC evaluates results
- results returned to participants