BOLT IR task, phase 2
Evaluation guidelines

(version 1.3, 29 May 2013)
(final before dry run)

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, and then group them into non-redundant groups. Accordingly, the task will be split into a **retrieval** subtask and a **grouping** 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 grouping subtask, systems will group the citations produced by their system, and provide each group with a descriptive textual label.

Document collection

The collection for phase 2 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 will identify a subcollection of roughly 700 million words from each language source. The phase 2 collection will be a superset of forum threads used in phase 1. 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 topics, relevance judgments, and facet 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 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.[1]

Topic format

```
<bolt-ir-topics eval=”BOLT-IR-P2” contact=”Ian Soboroff
ian.soboroff@nist.gov”>
<topic number=”1.001”>
<query> The query sentence. </query>
<language-target lang=”arz”/> <!-- cmn, eng, arz, or none -->
```

[1] Topic facets are completely removed from the topics and topic development process.

```
<!-- 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”/>[2]
</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 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 simply to provide a manual-search perspective on relevant citations and to expand the total relevant set beyond those retrieved by the systems.

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”/>
```

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

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 evaluation topics, and will return a ranked list of at most 100 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”.[3]

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-P2”
                    subtask=”citations”
                    contact=”John Doe johndoe@foo.org”>
<response number=”1.001”>
  <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”
     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=”1.002”>
...
</bolt-ir-submission>
```

The “subtask” in the submission header should be “citations”. 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.

[3] 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”.

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 100 citations per topic. All returned passages will be pooled for assessment.

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% 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** has provided a standard implementation of the near-duplicate citation heuristic to ensure teams handle this case uniformly.

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 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. Additionally, members of the teams will also judge all submitted citations for relevance, for purposes of comparison.

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**.

As the assessor grades a retrieved passage, it may be that only a substring of the passage will be determined to be relevant. In these cases, the relevant portion of the passage will be identified with a <relspan> section. 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 the passage that are not relevant to the topic.

Official measures for this task will include:

1. Character precision, the fraction of characters returned marked “yes” or “maybe”,
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).
3. An F-measure (harmonic mean) of the above precision and recall metrics, weighted equally.
4. Character-based mean average precision.

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

Subtask 2: Thematic Grouping

For the thematic grouping task, each team will produce up to three clusterings (each called a **view**) of the citations retrieved by their system for each topic. A view consists of a set of at most 10 **groups**, each of which have a group **label** and one or more citations. Citations may be members of more than one group in a view.

The goal of thematic grouping is to minimize redundancy across groups. From an interface standpoint, we imagine that the user of a BOLT IR system would only see a representative citation from each group, unless they wished to drill down into the group and explore all the citations there. The notion of redundancy is necessarily specific to an individual topic as well as to the citations retrieved. For example, if a topic concerning opinions in different countries, it might make sense to group the citations by country as well as polarity. Hence, teams may return up to three views per topic.

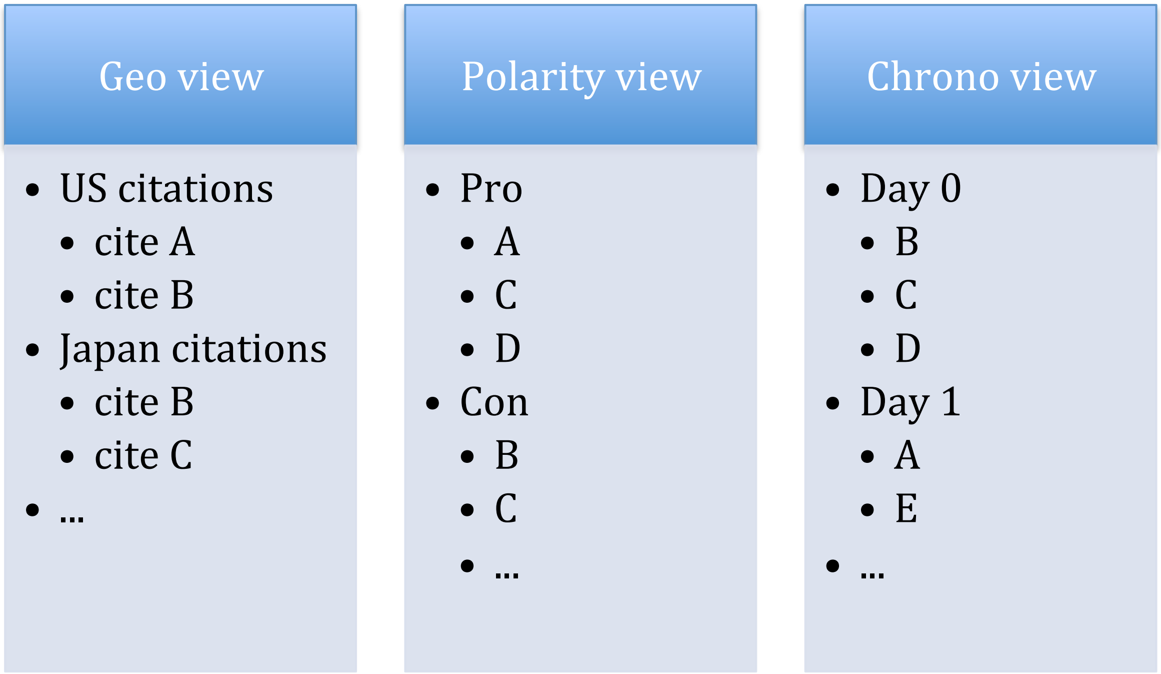

Figure 1: three views of the citations for an imaginary topic.

After the citation retrieval subtask is assessed, each passage will be assigned a unique identifier (see submission format below). Teams will then produce up to three views of the citations output by their system. Each view may have up to 10 groups in it. This grouping should be produced fully automatically given the query, the output of the system from the citation retrieval task, and the document collection. Each passage must be a member of one or more groups. Each group should receive a textual label of at most 250 characters.

A view may have a group with the label "miscellaneous", which can serve as a catchall for citations that don't fit well into other groups. Citations in the "miscellaneous" group are not intended to be related to or redundant with each other. The miscellaneous group does not count towards the 10 group maximum per view.

Group labels will be evaluated as if the labels were extractive summaries of the citations in the group. The exact procedure remains to be determined.

The thematic grouping task will run as a pilot task in phase 2. LDC will examine thematic groupings and judge their appropriateness, scope, and cohesiveness, following guidelines to be discussed later.

Thematic groupings will use the following format (please see the appendix for an authoritative DTD):

```
<bolt-ir-submission team=”my team name”
                    date=”the due date”
                    eval=”BOLT-IR-P1”
                    subtask=”groups”
                    contact=”John Doe johndoe@foo.org”>
<response number=”1.001”>
<view number=”0” label=”Grouping by color”>
<group number=”0” label=”A nice blue grouping”>
<cite id=”aBc12Z” grpscore=”0.995” 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 id=”xYz33h” grpscore=”0.8” score=”0.82” thread=”thread-id”
   post=”post-id” offset=”start” length=”in chars”>
  This text came from another post in another thread.
</cite>
</group>
<group number=”1” label=”Another group of citations”>
<cite id=”fr0tz3” score=”0.5” thread=”thread-id” post=”post-id”
   offset=”start” length=”in chars”>
  Each passage has a score between 0 and 1.  This score imputes
   a ranking over all passages for a response.  Tied scores will
   be broken arbitrarily.
</cite>
...
</group>
…
</view>
<view number=”1” label=”Grouping by event clump”>
…
</view>
</response>
<response number=”1.002”>
...
</bolt-ir-submission>
```

The “subtask” attribute in the submission header should be “groups”. Views and groups should be numbered from 0, but the order is unimportant. Citations in the groups should be literal citations from those emitted in subtask 1, with the addition of a ‘grpscore’ attribute indicating a score or confidence that the citation belongs in this group.

(Keeping the passage text in the submission makes them more readable without a special tool.)

Views will be judged independently of each other. Put another way, one view does not have to take into account the existence or structure of another view for the same topic.

Evaluation

The evaluation for thematic grouping in phase 2 will be exploratory. At a high level, system-produced groupings should help the user find relevant information quickly, understand the range of information presented, and explore the areas of the topic that are of most interest to them.

An LDC assessor will mark the teams groupings according to several qualities:

1. Appropriateness/relevance: (yes/no)
   a. For each view, is it relevant to the topic (yes/no); does this view seem to be a useful way to cluster the citations?
   b. For each group, is the group relevant to the topic?

2. Cohesiveness: do all the parts relate to the whole? (yes/no)
   a. for each group in a view, is the group a reasonable member of the view?
   b. for each citation In a group, does the citation belong in this group?

3. Scope: this group seems at the right level of granularity for this topic?
   a. for each group, a (yes/no) decision.

These assessments will not be applied to a ‘miscellaneous’ group, because doing so would be equivalent to asking the LDC to re-cluster the group.

Metrics are still under discussion, but may include

1. The fraction of appropriate, well-scoped groups in the view (“view precision”)
2. The average cohesiveness of groups in the view (“group precision”)
3. The standard deviation of the precision of citations in a group, averaged over groups in a view (“group skew”)

Grouping annotations and metrics are intentionally limited to those which can be annotated by an assessor looking only at individual groups in a view. Measures that would require the assessor to make cross-group comparisons or to regroup citations are avoided for scale reasons.

Idea from Salim in the 1/30 call: fraction of groups that contain only relevant citations. More broadly, ideas about how to scale group measures by precision of their citations. This was a very unfinished discussion.

As mentioned, labels on groups will be scored by treating the label as an extractive summary of the citations, and follow current practices in summarization evaluation.

Schedule

The schedule below includes two dry runs of the relevance task; the first will have a single topic and is intended to help finalize any issues remaining before the guidelines are finalized. A dry run of 50 topics (5 focusing on Egyptian, 5 on Mandarin, and 40 on English) will take place in June. The official relevancy evaluation will happen in September.

The thematic grouping pilot is not currently reflected in the schedule.

January 15 Draft Evaluation Plan

- Finalized by Feb 15

January 31 Pilot

- **LDC releases document list for phase 2 (700Mw/lang)**
- Jan-31 mini dry run (one query)
- Feb (two week assessment period)

May 20 dry run queries released to teams

- (LDC is planning to develop the eval queries at this time, too)
- Dry run set will include 50 queries (5-A/5-C/40-E)

June 10 dry run results due

- Teams return relevancy and grouping results to NIST by 6/10
- LDC to complete assessment for relevance by 8/5 (grouping assessment due date TBD)
- Scores returned to teams on or about 8/9

PRIOR TO EVALUATION QUERY RELEASE, submit to NIST:

- **Documentation of collection preprocessing stages,**
- **Annotations of the collection,**
- **External resources used,**

September 16 Relevancy Evaluation

- (36 hour period)

October 31

- Relevancy assessments completed

December 2 Eval Reporting

- Reports to DARPA

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 ,Äúarz,Äù, ,Äúcmn,Äù, ,Äúeng,Äù, or ,Äúnone,Äù -->

<!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 (yes|no|maybe) "yes">
<!-- <relspan> sections are for annotators to limit the relevant span
    of a supplied source text. <nonrelspan> sections are for delineating
    passage content which should be ignored for scoring.
  -->
<!ELEMENT relspan (#PCDATA)>
<!ELEMENT nonrelspan (#PCDATA)>
```

DTD for IR task submission files (citation subtask)

```
<!-- <!DOCTYPE bolt-ir-submission SYSTEM "bolt-ir-cite-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 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.
     ,Äòsubtask,Äô identifies the subtask and should be ,Äòcitations,Äô.
     '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
  -->

<!-- citation attributes which teams should return -->
<!ELEMENT cite (#PCDATA|relspan)*>
<!ATTLIST cite score CDATA #REQUIRED>  <!-- between 0 and 1 inclusive -->
<!ATTLIST cite thread CDATA #REQUIRED>
<!ATTLIST cite post CDATA #REQUIRED>
<!ATTLIST cite offset CDATA #REQUIRED>
<!ATTLIST cite length CDATA #REQUIRED>
<!ATTLIST cite original CDATA #REQUIRED> <!-- original post text from
above pointer -->

<!-- citation attributes added by NIST during pooling -->
<!ATTLIST cite id CDATA "-1">
<!ATTLIST cite group CDATA "null">

<!-- citation attributes added by LDC during assessment -->
<!ATTLIST cite rel (yes|no|maybe) "no">
<!ATTLIST cite chksrc (yes|no) "no">
<!ATTLIST cite trans (accept|problematic|notaccept|na) "notaccept">

<!-- relspan tags may be added by LDC during assessment -->
<!ELEMENT relspan (#PCDATA)>
```

DTD for IR task submission files (grouping 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 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.
    ,Äòsubtask,Äô identifies the subtask and should be ,Äògroups,Äô.
    'contact' is a contact name and email address, for example,
      Ian Soboroff ian.soboroff@nist.gov
  -->

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

<!ELEMENT view (group+)>
<!ATTLIST view number CDATA #REQUIRED>
<!ATTLIST view label CDATA #REQUIRED>
<!-- Up to three views per topic are allowed.  The label is a text
    label for the view. -->

<!ELEMENT group (cite+)>
<!ATTLIST group number CDATA #REQUIRED>
<!ATTLIST group label CDATA #REQUIRED>
<!-- Groups should be numbered, but the order is unimportant. The label is
    a text label for the group. -->

<!-- Citations are ALMOST identical to the citation task. These should be
    verbatim from the citation task output file, with the addition of the
    ,Äògrpscore,Äô attribute which is a score value for this citation
within
    this group. -->
<!ELEMENT cite (#PCDATA|relspan)*>
<!ATTLIST cite id CDATA "-1"> <!-- generated by NIST for LDC during
pooling -->
<!ATTLIST group id CDATA "null"> <!-- generated by NIST 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 #REQUIRED> <!-- original post text from
above pointer; if original is English, can put 'n/a' in here. -->
<!ATTLIST cite rel (yes|no|maybe) "no">
<!ATTLIST cite chksrc (yes|no) "no">
<!ATTLIST cite trans (accept|problematic|notaccept|na) "notaccept">

<!ELEMENT relspan (#PCDATA)>
<!ATTLIST cite grpscore CDATA #REQUIRED>
```