The Query of Everything: Developing Open-Domain, Natural-Language Queries for BOLT Information Retrieval

Kira Griffitt, Stephanie Strassel
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA. 19104 USA
Email: kiragrif@ldc.upenn.edu, strassel@ldc.upenn.edu

Abstract

The DARPA BOLT Information Retrieval evaluations target open-domain natural-language queries over a large corpus of informal text in English, Chinese and Egyptian Arabic. We outline the goals of BOLT IR, comparing it with the prior GALE Distillation task. After discussing the properties of the BOLT IR corpus, we provide a detailed description of the query creation process, contrasting the summary query format presented to systems at run time with the full query format created by annotators. We describe the relevance criteria used to assess BOLT system responses, highlighting the evolution of the procedures used over the three evaluation phases. We provide a detailed review of the decision points model for relevance assessment introduced during Phase 2, and conclude with information about inter-assessor consistency achieved with the decision points assessment model.

Keywords: language resources, information retrieval, queries, question-answering, open-domain, natural-language, informal text

1. Introduction

This paper describes the resources, procedures, and adaptations developed by the Linguistic Data Consortium (LDC) in support of the Information Retrieval (IR) evaluation within DARPA's Broad Operational Language Translation (BOLT) program.

Within the context of BOLT's overarching goal of improving machine translation capabilities in informal data genres, the BOLT IR task focused on advancing the state of the art of information retrieval over these genres (DARPA, 2011). In particular, BOLT IR seeks to support development of systems which could: 1) take as input a natural language English query sentence, 2) return relevant responses to that query from a large corpus of informal documents in the three BOLT languages (Arabic, Chinese, and English), and 3) translate relevant responses into English where necessary (i.e. if those responses came from non-English documents). These objectives were chosen because they closely modeled the information retrieval needs of monolingual English intelligence analyst (NIST 2014).

LDC developed queries and assessed system responses for the BOLT Phase 1, Phase 2 and Phase 3 evaluations. The National Institute for Standards and Technology (NIST) was responsible for designing the BOLT IR evaluation task and measuring system performance.

2. The BOLT IR Corpus

The BOLT program focused on English, Mandarin Chinese and Egyptian Arabic covering three genres: discussion forums (Garland et al., 2012); SMS/Chat (Song et al., 2014), and Conversational Telephone Speech. While the BOLT MT evaluations covered all three genres, IR evaluations focused exclusively on discussion forum data. This genre exhibits the challenges of informal language while still containing the kind of news-focused content required for multilingual query development in BOLT.

The BOLT Phase 1 IR Corpus comprised 400 million words of discussion forum data per language. The large corpus size was necessary to ensure that multiple query sets could be developed from the same data pool without exhausting all possible topics, and that systems had a sufficiently large pool of data over which to do retrieval.

In Phases 2 and 3 the corpus was expanded to approximately 700 million words per language, such that the Phase 1 data was a strict subset of this expanded data pool. In all phases, we developed dry run and/or pilot queries to support system training and development, as well as evaluation queries for testing system performance.

3. BOLT IR Queries

3.1 GALE Distillation and BOLT IR

Both BOLT IR and the earlier the GALE Distillation task (Florian et al., 2011) have the needs of the monolingual English-speaking analyst in mind.

No.	Template
1	List facts about [EVENT]
2	What connections are there between [EVENT1/TOPIC1] and [EVENT2/TOPIC2]?
7	Describe the relationship of [PERSON/ORG] to [PERSON/ORG]
14	What [PEOPLE/ORGANIZATIONS/ COUNTRIES] are involved in [EVENT] and what are their roles?
15	Describe involvement of [PERSON/ORGANIZATION/ COUNTRY] in [EVENT/TOPIC]

Table 1: Some GALE Distillation query templates

As shown in Table 1, GALE Distillation queries were template-based, with uniform content and structure requirements. Distillation templates restricted both the form and content of queries, providing a set of 17 uniform

English sentences with placeholders for query arguments, which could only be completed with entities or events of the type specified in the argument placeholder.

While GALE templates allow for the expression of broad, underspecified information needs (e.g. Template 2) and complex information needs (e.g. Template 14), they do not require systems to do natural language understanding beyond the restricted English template sentences. Distillation also differs from BOLT IR in that possible argument types are directly specified in the templates, and are restricted to those specified types.

Building on the experience of GALE Distillation, BOLT IR expands on template-based queries, requiring an extension of system capabilities in natural language understanding and argument typing. In contrast to Distillation, BOLT IR queries: a) Require systems to use natural language understanding to interpret the English query sentence, successfully identify query target (argument) types, and successfully identify desired query response types without the aid of templates; b) Require systems to translate responses from non-English source data (including Egyptian Arabic) into English; and c) Require systems to work exclusively within the informal genre of discussion forums.

The informal, uncontrolled language of discussion forums introduces a number of additional challenges to BOLT IR, including a larger range of expressions and idioms than those found in formal, controlled language (e.g. newswire data), non-standard linguistic and typographic phenomena, and long anaphora chains in threads of arbitrary thematic and structural complexity (Garland et al., 2012).

An additional challenge stemmed from BOLT's focus on dialectal Egyptian Arabic, rather than the Modern Standard Arabic (MSA) variety targeted in GALE. The diglossia situation in the Arabic speaking world means that informal text harvested from the web often contains a mixture of both Egyptian and MSA. This was certainly true of the BOLT IR discussion forum corpus, and the query design and assessment procedures had to account for this.

3.2 BOLT Query Format and Structure

In order to promote a consistent approach to query development, ensure sufficient variety in query topics, and provide a high degree of confidence that developed queries were viable for evaluation, we produced both long form (full) and short form (summary) versions for all queries. The full format required annotators[1] to create not just a natural language query string, but an associated set of metadata that could be used to monitor thematic variety in queries, establish relevance criteria for query responses and provide a set a of sample human answers in accordance with those relevance criteria.

Full queries were formatted as XML, with the following structure and fields shown in Figure 1 below. The content and constraints on full-form query fields are as follows:

```
<topic number="BIR_300054">
  <query>Should the United States
Intervene in Syria?</query>
  <description>This query asks for
statements or opinions about whether or
not the United States should intervene in
Syria.</description>
  <language-target lang="none"/>
  <properties>
    <asks-about target="location"/>
    <asks-for response="statements-or-
opinions"/>
    <languages eng="T" arz="F" cmn="F"/>
  </properties>
  <rule number="1">Answers must be about
whether the United States should
intervene, not just what is happening in
Syria</rule>
  <rule number="2">Answers must be about
intervention by the United States, not
other countries.</rule>
  <rule number="3">Answers must be about
intervening in Syria, not the middle east
in general.</rule>
  <cite number="1" thread="bolt-eng-DF-
312-210461-25161904" post="2" offset="1"
length="71" rel="yes">Mr. Bolton the
Zionist Of course he wants the US to
intervene in Syria.</cite>
  <cite number="2" thread="bolt-eng-DF-
183-195681-7949359" post="21"
offset="2581" length="129" rel="yes">The
US and others should do something but it
should not be military, either direct
military involvement or arming the
opposition.</cite>
```

Figure 1: BOLT Phase 3 query in full form

- *Topic* contains a unique ID for the query
- *Query* contains the actual natural language English query string presented to systems at evaluation time, required to be one sentence in length
- *Description* contains a more formal restatement of the natural language query. Crucially, it cannot contain any information not reasonably inferable from the query itself
- *Language-target* indicates whether systems must do retrieval in a specific language
- *Properties* contains three subfields, used to track trends in query development:
  - *Asks-about* indicates the type of entity the query is targeting
  - *Asks-for* indicates the type of information the query author is seeking
  - *Languages* indicates the language in which the query author found sample human answers to the query
- The numbered *Rules* fields enumerate basic, commonsense pieces of information that a citation (i.e. query response) must contain to be considered relevant. Annotators were restricted to a maximum of three rules per query.

[1] In this document, query developers are also referred to as 'annotators'. However, 'assessors' (see Section 4.2 below) are only ever referred to as 'assessors'.

- The numbered *Cite* fields contain sample human answers (citations) to the query, intended to demonstrate the viability of a query for evaluation; each evaluation query was required to have at least two observed answers in the source corpus. The sample human citations could come from any language in the corpus, whereas system citations originating in Arabic or Chinese had to be translated into English.

In order to push IR system capabilities in query understanding and query argument interpretation, full format queries were provided only after the conclusion of the evaluation. At run time, systems were provided with an abbreviated summary form of the queries. As shown in Figure 2, the summary form contains only the topic number, query string, and (in Phases 2 and 3) the language-target for each query.

```
<topic number="BIR_300054">
  <query>Should the United States Intervene
in Syria?</query>
  <language-target lang="none"/>
</topic>
```

Figure 2: BOLT Phase 3 query in summary form

3.3 Query Development Procedure

LDC annotators developed pilot, dry run and evaluation queries for all phases of BOLT. Formal guidelines described requirements for query creation, including examples of suitable and unsuitable queries. Although annotators were not restricted in their query topics, a special effort was made to ensure that some were applicable to intelligence analysis scenarios. To support this goal, annotators were provided with suggested query target (i.e. "asks-about") and response (i.e. "asks for") types to use during query development. Suggested target types were: persons, organizations, locations, facilities, events, movements, practices-or-customs, products, publications, laws, awards, diseases, abstract entities, or other. Suggested response types were: statements-or-opinions, causes of, effects of, relationship-between, or other.

Annotators were limited in how much time they could spend developing each query. The amount of time varied from 60-90 minutes per query, depending on the evaluation phase. For each candidate query, annotators supplied basic information using a custom web and then searched the corpus using a language-specific, phrase-based, Boolean search tool. Once at least two relevant answers were found in the corpus[2], annotators began full query development by writing the query as a natural language English sentence, creating a query description (formal restatement) of the query, optionally indicating a language-target, selecting applicable query target and response type categories, and writing a set of the rules for how relevance would be determined for this query. Note that in order to ensure linguistic variety, annotators were allowed and encouraged to use synonyms and paraphrases when writing a query in English sentence form, as long as these did not make the language of the query overly informal. Candidate queries without at least two relevant answers in the discussion forum corpus were dropped from further development.

The resulting combination of query string, metadata, rules and sample citations produced a full form topic that was then reviewed by a senior annotator for conformance to query guidelines before being selected by task managers into a query data set.

3.4 Query Development Results

The query development procedures described above resulted in diverse set of information retrieval queries for each phase of BOLT, spanning a variety of query forms and themes. Examples include:

- What would happen if the U.S. president had line item veto?
- What do people think of Mohamed Morsi as a candidate for Egyptian presidential elections?
- How can you protect yourself from identity theft?
- Are there weapons stockpiled in Coptic churches?
- What did people say when Arlen Specter switched parties?
- What are the effects of catching Ebola?
- How can one reduce exposure to formaldehyde after a home renovation?
- Do tattoos affect employment?

Across BOLT Phases 1-3, LDC produced a total of 512 natural-language, open-domain queries, with distribution across the languages, phases and partitions as shown in Table 2.

	English			Arabic			Chinese		
	pilot	dry run	eval	pilot	dry run	eval	pilot	dry run	eval
Phase 1	0	5	60	0	2	26	0	2	60
Phase 2	1	40	34	0	5	33	0	5	33
Phase 3	2	40	50	2	5	50	2	5	50
	Total: 232 queries			**Total: 123 queries**			**Total: 157 queries**		

Table 2: Count of queries per language in each phase

4. Responses and Assessment

4.1 System Responses

While sample human citations could come from any language, BOLT systems were required to return citations in English only, applying BOLT Machine Translation technology to passages returned from Arabic or Chinese documents. In order to constrain response length, a 250-character limit was imposed on system citations during Phases 2 and 3.

[2] Given the prevalence of both MSA and Egyptian dialectal Arabic (EA) in the Arabic discussion forum data, both varieties were allowable as responses to queries with Arabic as a language-target. This was the case for both human-generated sample citations and for the original source text underlying (translated-into-English) system citations. In Phase 3, annotators were also required to flag system citations whose underlying, untranslated source text was primarily EA, to enable analysis of relative performance on EA vs. MSA data.

System runs were submitted to NIST, who then produced anonymized pools for each query consisting of the top-ranked citations from each system. The anonymized, pooled citations were then distributed to LDC for assessment.

4.2 Assessment Procedures and Criteria

LDC further grouped pooled system citations for each query by language, so that citations translated from non-English source documents could be assigned to bilingual (Chinese-English or Egyptian Arabic-English) assessors. This further grouping of citations by language was necessary so that assessors could check the non-English source text underlying the English citation, for instance in cases where the machine-translated English citation was unclear.

During assessment, assessors were presented with the query string, its rules, and the pooled set of English system citations for that query that were extracted from documents in the assessor's native language. To be considered relevant, a system response had to satisfy all rules of interpretation for the query and provide at least some new information (i.e. not simply restate the query). If a citation did not meet all these criteria, it was not considered relevant.

In addition to these relevance judgments, assessors also provided translation acceptability judgments for relevant citations that came from non-English source documents. These judgments indicated how well a system preserved relevant information from the underlying non-English source text when translating the citation into English.

Where possible, the annotator who developed the query also assessed system responses for that query, although this was not a requirement.

4.3 Changes to Assessment Procedure and Guidelines

A number of adaptations were made to the assessment procedures and criteria after the Phase 1 evaluation.

In BOLT Phase 1, assessors were required to perform coreference on relevant system citations, in the interest of reducing redundancy for the end user. In practice coreference was problematic, in query responses comprised complex predications and were very rarely (if ever) truly coreferential. Thus coreference of citations was eliminated in Phase 2 and beyond.

It was observed in the Phase 2 dry run that assessors used varying standards of strictness in assessing relevance. To address this concern, assessment guidelines and training were revised to provide additional guidance on this question, with the intention of encouraging annotators to err on the side of generosity when judging system citations for relevance.

For instance, consider the example in Figure 3 below. In this example, the citation (translated into English by the BOLT system) discusses the Euro crisis and its effects, so it clearly satisfies rules 1 and 3. While the citation doesn't mention China explicitly, it is reasonable to infer that China is one of the "Asian economies" mentioned in the citation, thus satisfying rule 2 and allowing the citation to be assessed as relevant.

```
<query>What are the influences of Euro
financial crisis on China?</query>
<rule number="1">Answers must be about Euro
financial crisis rather than any other
country's economic crisis.</rule>
<rule number="2">Answers should be about the
effects of Euro financial crisis on China
rather than other countries.</rule>
<rule number="3">Answers must be the effects
instead of any other things about Euro
economic crisis.</rule>
<citation> Due to the spread of the European
debt crisis has intensified, the us economic
recovery is sluggish, further deterioration
of the external environment in the
development of the Asian economies.
</citation>
```

Figure 3: Query citation judged as relevant

4.4 Decision Tree and Decision Points

To encourage greater overall assessor consistency, the assessment procedure was revamped after Phase 1 to make use of the notion of decision points, in which each individual component of the relevance decision making process is broken out into a separate question for assessors to answer directly; the final relevance judgment is automatically derived from the finer-grained decisions.

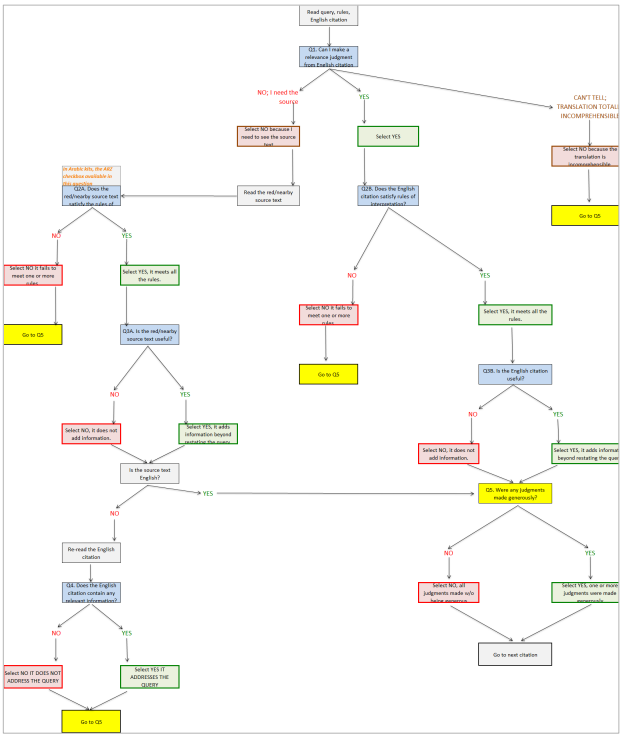

Figure 4: Phase 3 Relevance Assessment Decision Tree

In this model, assessors answered up to five questions for each citation. The assessment user interface was designed to present questions dynamically, so that answers to earlier questions determined which (version of) later questions would be presented. Underlying this model is a decision tree, capturing all of the decision points facing an assessor. The Phase 3 decision tree is

shown in Figure 4[3].

The first tier in the decision tree concerns the possible need to see the citation in the context of its source document. The second tier concerns the citation's conformance to the rules of interpretation for its associated query. The third tier concerns the utility of relevant information in the citation. The fourth tier concerns the preservation of relevant information in the translations of citations from non-English source documents. The fifth tier concerns the level of generosity assessors used to make their judgments. The specific questions for each tier are described in detail below.

In Question 1 (Q1), assessor were presented with the following question and potential answers (judgments):

- Q1: Can you answer relevance questions based on this English citation without looking at the source text?
    - YES.
    - NO, because the translation is incomprehensible.
    - NO, because I need to see the source text to resolve pronouns, get more context and/or clarify the translation.

If the assessor responded "No, incomprehensible" to Q1, they were asked whether any of their assessments were made generously (Q5, see below) and then moved to the next citation in their kit[4]. If the assessor responded "No, -need the source", the assessment interface would then display the source document in its original language, and all subsequent questions were answered with respect to the citation in the context of its original (untranslated) source text. Assessors then moved on to Question 2A. If the assessor responded "Yes" to Q1, they were required to make all subsequent judgments based on the English citation alone, and the assessment interface would not display the source document; assessors then moved on to Question 2B.

Taken together, Questions 2A and 2B comprise Tier 2 of the decision tree, since they both concern the citation's conformance to its rules of interpretation. Q2A asks for a judgment based on the source/surrounding text:

- Q2A: Does the source/surrounding text satisfy the rules of interpretation?
    - YES, it meets all the rules.
    - NO, it fails to meet one or more rules.

In contrast, Q2B asks for a judgment based on the English citation:

- Q2B: Does the English citation satisfy the rules of interpretation?
    - YES, it meets all the rules.
    - NO, it fails to meet one or more rules.

If the assessor responded "No" to Q2A or Q2B, they moved to Q5 to indicate whether any assessments were made generously, and then moved to the next citation in their kit. If they responded "Yes" to Q2A or Q2B, they continued assessment and moved to Question 3A or 3B, respectively.

Taken together, Questions Q3A and Q3B comprise the third tier of the decision tree, since they both concern the utility of the information in the citation. Q2A is based on the source/surrounding text:

- Q3A: Is the source/surrounding source text useful?
    - YES, it adds information beyond restating the query.
    - NO, it does not add information.

While Q3B based on the English citation:

- Q3B: Is the English citation useful?
    - YES, it adds information beyond restating the query.
    - NO, it does not add information.

Note that after this third tier, the "A" branch of the decision tree (the portion of the decision tree where the object of assessment is the source/surrounding text, containing questions Q2A, Q3A, etc.) and the "B" branch of the decision tree (the portion of the decision tree containing Q2B, Q3B, etc.) differ. This is because the "A" branch takes into account whether the citation under assessment comes from an English or non-English document: If the assessor responded "No" to Q3A and the source for the citation was an English document, they moved to Q5 to indicate whether any assessments were made generously, and then moved to the next citation in their kit. If the assessor responded "Yes" to Q3A, and the source for the citation was a non-English document, they moved to Question 4:

- Q4: Does the English citation above contain any relevant information?
    - YES, it addresses the query.
    - NO, it does not address the query.

Regardless of whether the assessor responded "Yes" or "No" to Q4, they moved to Q5 and then onto the next citation in their kit.

Contrastively, the "B" branch of the decision tree did not present assessors with Question 4, since the only object of assessment in this branch of the decision tree is the English citation. Thus, regardless of whether the assessor responded "Yes" or "No" to Q3B, they moved directly to Q5 and indicated whether any assessments were made generously, and then moved to the next citation (without being presented with the tier 4/Q4 question).

Finally, assessors were presented with Question 5 (Q5), with the following potential judgments:

- Q5: Were any of the judgments made generously?
    - YES, one or more of the judgments was made generously.
    - NO, all of the judgments were made

[3] Figure 4 is intended to illustrate the complexity and overall flow of the decision tree and is not expected to be fully legible.

[4] A kit comprised the set of citations for a particular query that were extracted from source documents in an assessor's native language.

without being generous.

As discussed above, in Phase 2 and beyond assessors were instructed to err on the side of generosity when facing difficult decisions about relevance. In many cases, it was not necessary for assessors to invoke this "generosity" standard, since system citations were clearly relevant or not relevant. Q5 was introduced for Phase 3 to help provide additional information to system developers about which of their returned citations required a generous interpretation from assessors.

Once the assessor had responded to the decision tree questions for every citation in their kit, the kit was marked completed and the assessor moved onto a new kit.

4.4.1. Assessor Consistency on Decision Points

Qualitative feedback on the decision tree model of assessment and the generous assessment standard was positive. Some portion of the Phase 2 queries were dually assessed using a double-blind assessment procedure. Overall agreement was computed by NIST and is summarized in Figure 5 below. Note that Q5 was introduced to the decision tree after Phase 2 and so no results for this question are available for Phase 2.

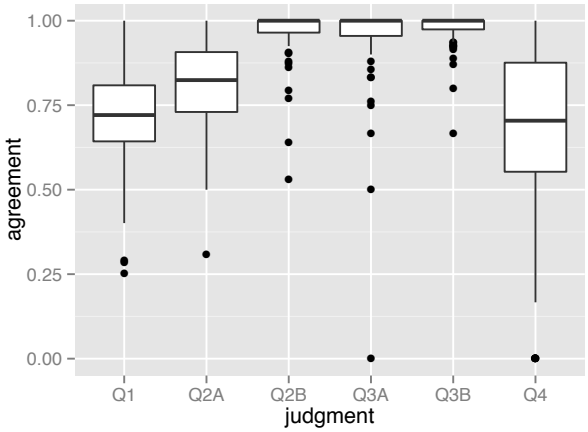

Figure 5: Inter-Assessor agreement in Phase 2 evaluation

Each question column in Figure 5 directly corresponds to a decision point from the Phase 2 assessment procedure:

- Q1 indicates assessor agreement on whether a citation could be assessed based on its English citation alone
- Q2A and Q2B indicate assessor agreement on whether a citation fit the rules of interpretation for its associated query
- Q3A and Q3B indicate assessor agreement on whether a citation added information beyond restating its associated query
- Q4 indicates whether assessor agreement on whether the translation of citations from non-English source documents preserving relevant information in the source text.

Inter-assessor agreement for Q2A averaged over 75%. Agreement for the Q2B, Q3A, and Q3B decision points was at or near 100% for the dually assessed queries. Given these results and feedback on the Phase 2 assessment process, the same assessment procedure was kept in place for the Phase 3 evaluation.

5. Conclusions

The BOLT Information Retrieval evaluation required systems to answer open-domain natural language English queries, returning short English answers from a large multilingual corpus of informal discussion forum text. Compared to the earlier GALE Distillation task, BOLT queries required systems to demonstrate a greater degree of natural language understanding as well as the ability to handle the challenges of informal text in three languages, including dialectal Arabic. Over the course of the BOLT program, our approach to query development and assessment changed to reflect emerging requirements as well as challenges inherent to the assessment task. In particular, we introduced a decision points model for query assessment that allowed us to achieve improved inter-assessor consistency.

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.

6. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-11-C-0145. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. We also acknowledge National Institute for Standards and Technology (NIST) for their role in designing and running the BOLT Information Retrieval evaluations.

7. References

DARPA. 2011. Broad Agency Announcement: I2O Broad Operational Language Translation (BOLT). Defense Advanced Research Projects Agency, DARPA-BAA-11-40.

Radu Hans Florian, Joseph Olive Caitlin Christianson, and John McCary eds. (2011). Chapter 4: Distillation. In Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation. Springer.

Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, Haejoong Lee. (2012). Linguistic Resources for Genre-Independent Language Technologies: User-Generated Content in BOLT. In *Proceedings of LREC 2012*, Istanbul, Turkey.

Julie Medero, Kazuaki Maeda, Stephanie Strassel, Christopher Walker. (2006). An Efficient Approach for Gold-Standard Annotation: Decision Points for Complex Tasks. In *Proceedings of LREC 2006*, Genoa, Italy.

National Institute for Standards and Technology (2012). BOLT IR task, phase 1 Evaluation guidelines. http://www.nist.gov/itl/iad/mig/upload/bolt-ir-guidelines-v5-0-April-15-2012.pdf. Retrieved March 16, 2016.

National Institute for Standards and Technology (2013). BOLT IR task, phase 2 Evaluation guidelines.

http://www.nist.gov/itl/iad/mig/upload/BOLT_P2_IR-guidelines-v1-3.pdf. Retrieved March 16, 2016.

National Institute for Standards and Technology (2014). BOLT IR task, phase 3 Evaluation guidelines. http://www.nist.gov/itl/iad/mig/upload/BOLT-P3-A-IR-guidelines_v2-5.pdf. Retrieved March 16, 2016.

Zhiyi Song, Stephanie Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainor, Preston Cabe, Thomas Thomas, Brendan Callahan, Ann Sawyer. (2014). Collecting Natural SMS and Chat Conversations in Multiple Languages: The BOLT Phase 2 Corpus. In *Proceedings of LREC 2014*, Reykjavik, Iceland.