Linguistic Data Consortium

Linguistic Resources for Genre-Independent Language Technologies: User-Generated Content in BOLT

Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, Haejoong Lee

Outline

- Introduction to BOLT
- Data collection and triage
  - Manual processes
  - Automatic harvesting
- Challenges inherent in informal web data
  - General challenges
  - Language-specific
- Summary/conclusion

BOLT

- DARPA BOLT (Broad Operational Language Translation)
  - Goals include development of genre-independent machine translation and information retrieval systems
  - Earlier DARPA programs including GALE (Olive, 2011) focused on structured genres like newswire and broadcasts
  - MT and IR systems perform less well on data that is informal or whose topics are less constrained
  - BOLT is particularly concerned with improving translation and information retrieval performance on informal genres

BOLT

- Phase 1 of BOLT
  - LDC collect and annotate threaded posts from online discussion forums
  - Target collection of at least 500 million words in each of three languages: English, Chinese and Egyptian Arabic
  - Part of the collected data is manually “triaged” for content and linguistic features
  - The triage process results in a selection of approximately one million words per language and segmented into sentences
    - English translations produced where required, with parallel text manually word-aligned
  - Approximately 50% of the source data selected for translation is further annotated
    - Treebanking, PropBanking, Entity co-reference

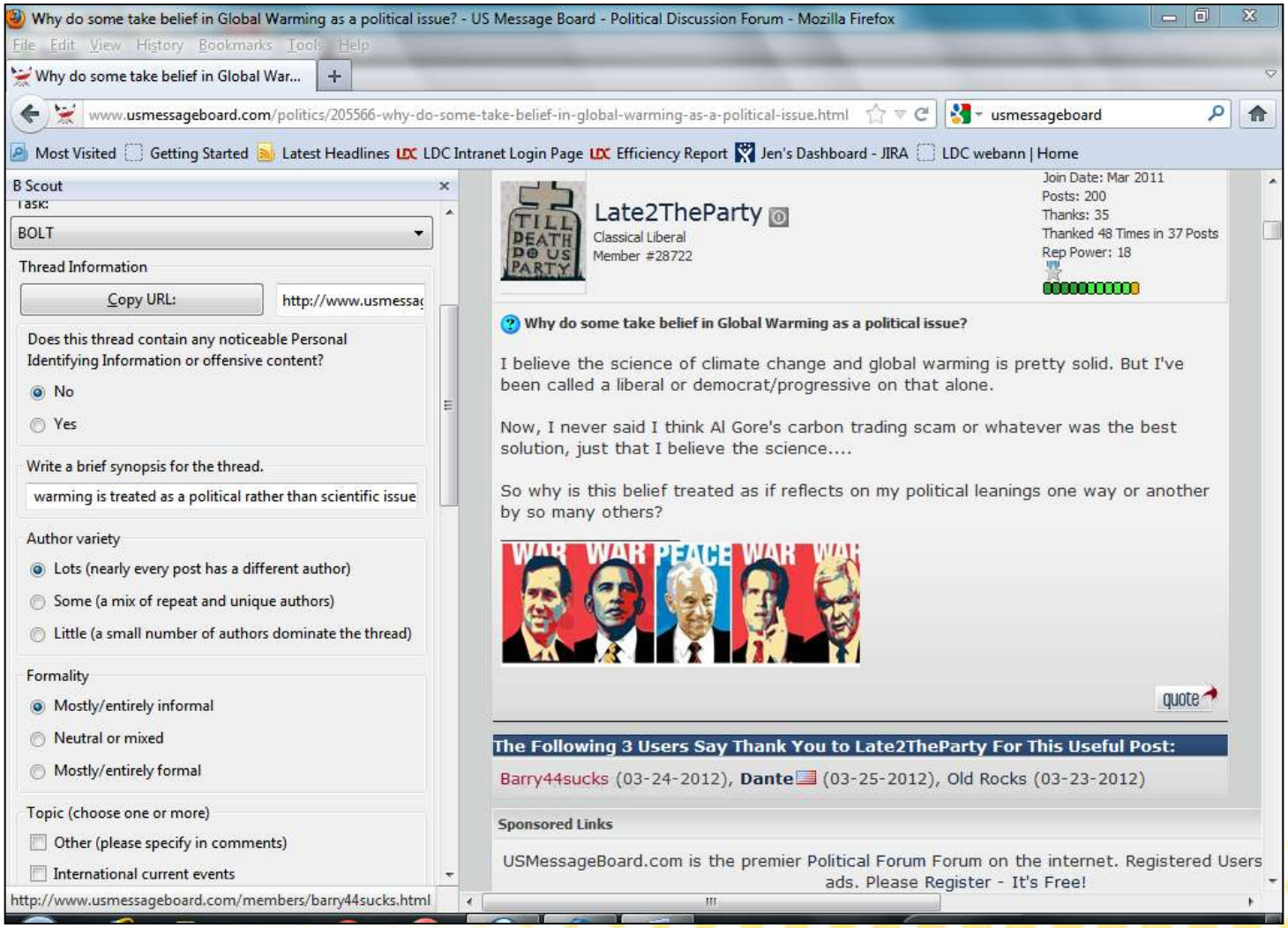
Data Collection

- Two-stage collection strategy: manual data scouting seeds the corpus, semi-supervised harvesting adds larger quantities of automatically-harvested data.
- Begins with native speaker annotators trained in the BOLT data scouting process.
  - Scouts search for individual threads that meet BOLT requirements
    - Primarily original content (as opposed to copies of a published news article, for instance)
    - Primarily informal discussion in the target language
    - Primary focus on discussion of dynamic events or personal anecdotes

Data Scouting Tool

- Data Scouting is facilitated through BScout, a customized user interface developed by LDC for BOLT.
  - Firefox browser plug-in records judgments for each scouted thread
    - Thread URL, screen for problematic content, synopsis, formality, (for Egyptian scouts) the use of Egyptian Arabic versus Modern Standard Arabic
- This information informs the automatic harvesting process
  - Scouted URLs and their corresponding annotations are logged to the BScout database and added to a whitelist for harvesting.
  - When multiple threads are submitted from the same forum that entire forum is targeted for harvesting.

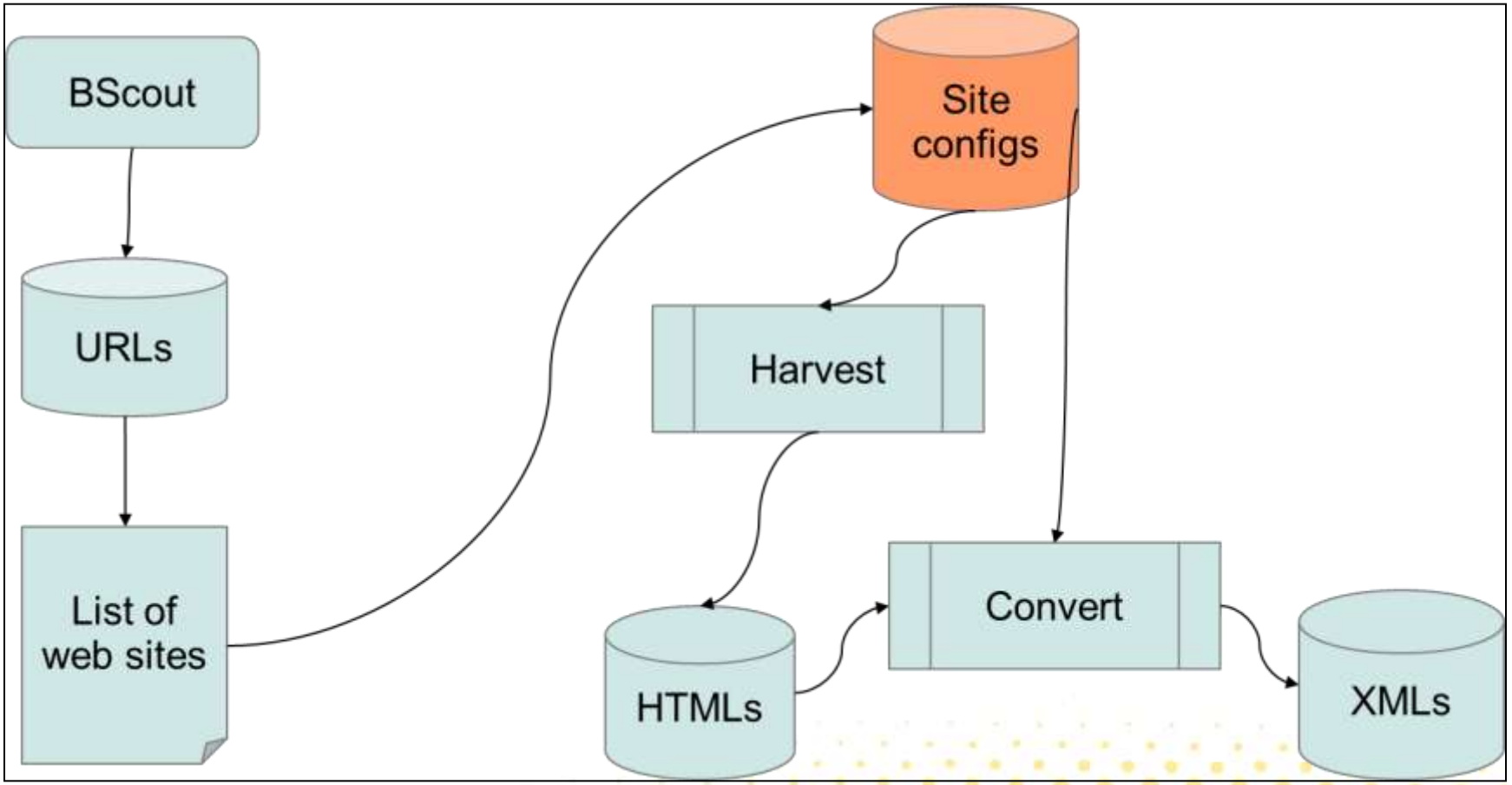
BScout


Why do some take belief in Global Warming as a political issue? - US Message Board - Political Discussion Forum - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Why do some take belief in Global War...
www.usmessageboard.com/politics/205566-why-do-some-take-belief-in-global-warming-as-a-political-issue.html
usmessageboard
Most Visited Getting Started Latest Headlines LDC Intranet Login Page Efficiency Report Jen's Dashboard - JIRA LDC webann | Home
B Scout
Task:
BOLT
Thread Information
Copy URL:
http://www.usmessa
Does this thread contain any noticeable Personal Identifying Information or offensive content?
No
Yes
Write a brief synopsis for the thread.
warming is treated as a political rather than scientific issue
Author variety
Lots (nearly every post has a different author)
Some (a mix of repeat and unique authors)
Little (a small number of authors dominate the thread)
Formality
Mostly/entirely informal
Neutral or mixed
Mostly/entirely formal
Topic (choose one or more)
Other (please specify in comments)
International current events
http://www.usmessageboard.com/members/barry44sucks.html
Late2TheParty
Classical Liberal
Member #28722
Join Date: Mar 2011
Posts: 200
Thanks: 35
Thanked 48 Times in 37 Posts
Rep Power: 18
Why do some take belief in Global Warming as a political issue?
I believe the science of climate change and global warming is pretty solid. But I've been called a liberal or democrat/progressive on that alone.
Now, I never said I think Al Gore's carbon trading scam or whatever was the best solution, just that I believe the science....
So why is this belief treated as if reflects on my political leanings one way or another by so many others?
WAR WAR PEACE WAR WAR
quote
The Following 3 Users Say Thank You to Late2TheParty For This Useful Post:
Barry44sucks (03-24-2012), Dante (03-25-2012), Old Rocks (03-23-2012)
Sponsored Links
USMessageBoard.com is the premier Political Forum Forum on the internet. Registered Users ads. Please Register - It's Free!

Automatic Harvesting

- Backend framework for BOLT to enable efficient harvesting, processing and formatting of large volumes of discussion forums and other user-generated web data
- Each forum host site has unique challenges for automatic harvesting in terms of structure and formatting
  - framework assumes a unique configuration for each site

Harvesting process



Harvesting process

- URLs submitted by data scouts using BScout are grouped by host site.
- Harvester downloads individual threads, converter transforms the downloaded HTML files to an XML format
- For each site, a configuration file is written for both the harvester and converter, consisting of a dozen or more XPath expressions and regular expressions.
  - For example, XPath expression to identify individual thread URLs contained within forum home page, XPath expression to identify the specific HTML element that contains the body text of posts in a thread, regular expressions to clean up target strings.

Site configuration challenges

- Many difficulties require careful examination of the source HTML file in order to identify the problem and achieve the correct configuration.
  - URL navigation (next forum, next thread) may need to be computed from a snippet of Javascript code
  - Illegal characters, control characters and poorly-rendered HTML can cause parse errors, requiring manual review to diagnose and correct problems
  - increasingly common use of AJAX:
    - downloaded thread HTML has no content; i.e., no body text
    - contents are downloaded dynamically to the web browser when the Javascript code embedded or linked on the HTML page is executed
    - work is in progress to account for this emerging pattern in the generalized framework

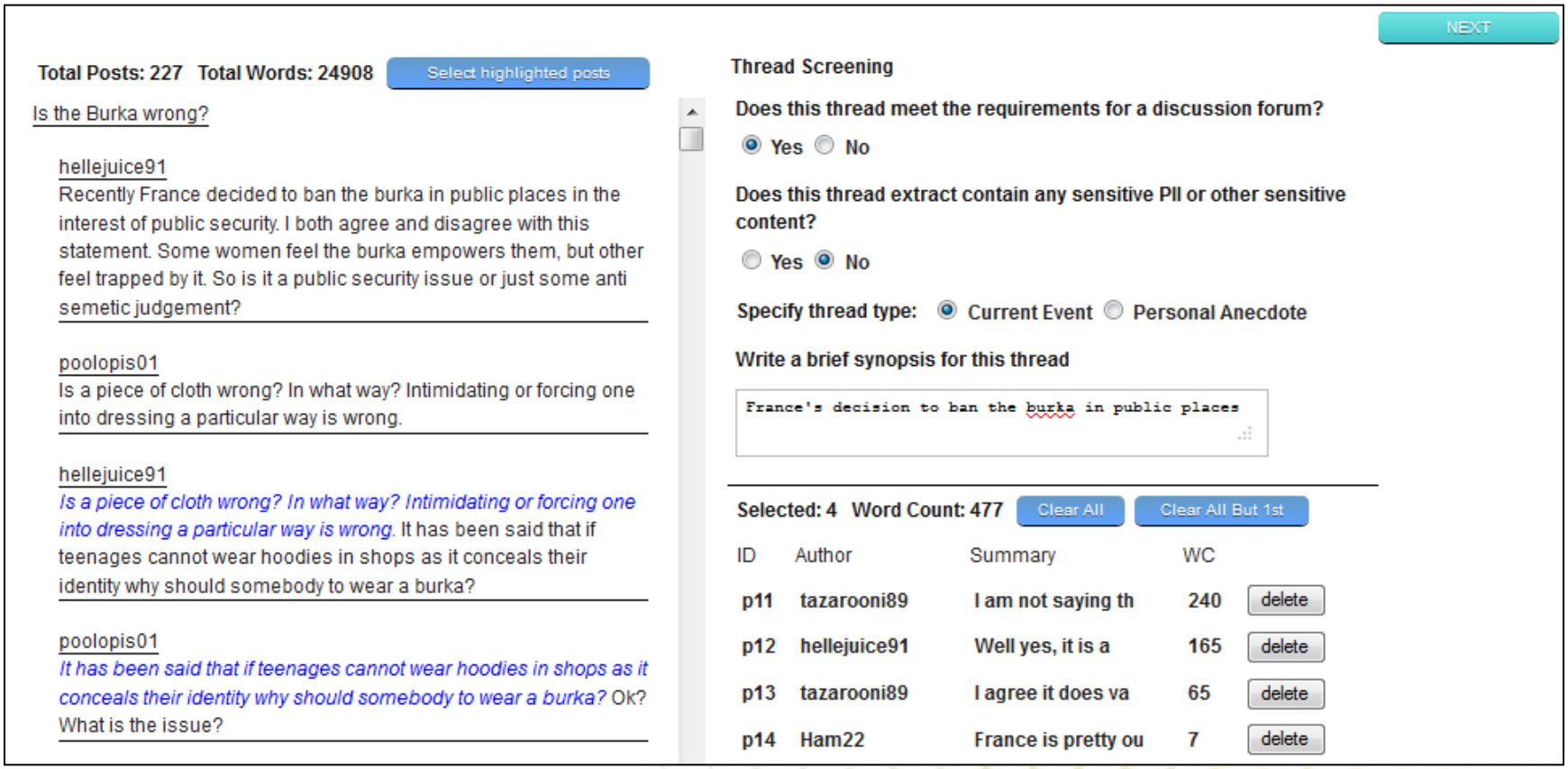
Triage

- All harvested data provided to BOLT performers, small subset selected for manual translation and annotation
  - Data selected for annotation must meet requirements for language and content
  - Selected data should be high-value, i.e. not duplicate the salient features of existing training data
- Manual triage cannot reach all threads in the corpus
  - Threads are selected for triage based in part on the results of data scouting, with manually scouted threads and threads from whitelisted forums having highest priority

Triage stage 1

- Triage task has two stages:
  - 1st stage = thread screening and post selection
    - native speaker annotator
    - confirms that the thread meets content and language requirements, does not contain offensive material or sensitive personal identifying information
    - selects individual posts from the thread that are suitable for translation and downstream annotation
      - Suitable posts: novel on-topic content in target language
      - Unsuitable posts: simple agreement/disagreement, primarily quoted text, not (primarily) in target language

Triage stage I tool



Triage stage 2

- 2nd stage = sentence unit segmentation
- Annotator identifies and labels individual Sentence Units within each selected post.
  - SU annotation provides stable basis for later linguistic annotation
  - Annotators identify SU boundaries, classify each SU as Keep or Exclude
    - Excluded content: sentences that consist entirely of quotes, sentences that are not in the target language, and segments that consist of formulaic greetings, hyperlink text, image labels, or other undesirable material.
    - Sentence Units marked Exclude are dropped from further annotation but are not deleted from the source corpus.

General challenges

- Quoted text
  - External quotes (e.g., newspaper articles)
    - primarily formal and non-interactive
  - Internal quotes (from other posters in a thread)
    - same content may have been annotated previously as part of the original post.
- Quote markup is maintained in harvesting/conversion, presence or absence of quoted text is considered during data triage

Multiple embedded quotes

> **Quote: Originally Posted by JakeStarkey**
>
> > ***Quote: Originally Posted by Quantum Windbag***
> >
> > *I though Obama studied law at some point.*
> >
> > > **Quote:**
> > >
> > > *Leaving aside the constitutional questions, there is a potential statutory problem with the legality of the Cordray appointment under Dodd-Frank. Section 1066 of Dodd-Frank provides that the Secretary of the Treasury is authorized to perform the functions of the CFPB under the subtitle transferring authority to the CFPB from the other agencies "until the Director of the Bureau is confirmed by the Senate in accordance with Section 1011." It turns out that section 1011 is a defined term which provides: "The Director shall be appointed by the President, by and with the advice and consent of the Senate."*
> > >
> > > *This seems to suggest that even if the President might be able to appoint Cordray under the recess power the full grant of statutory authority wouldn't transfer to the Bureau unless the statutory language was fulfilled as well.*
> >
> > *The Volokh Conspiracy » Legality of Cordray Appointment Under Dodd-Frank*
>
> *Nah, the minority does not say when the Senate is in session, and the President has the authority to make a recess appointment. The courts will rule for the president on this one.*

In order for the Senate to go in recess the House must recess. No such event occurred. Further in order for the Senate to recess they must meet the stated rules for recess, this did not happen either. The appointment will fail in all but a liberal Judges court.

And the left is now FIRMLY in the position of stating a President, ANY President may violate the Constitution any time he chooses with no adverse reaction from the people.

Threading

- Discussion forums = threaded, informal, interactive
  - multiple posters' perspectives on a topic
  - individual posts are best understood in the context of the previous posts within the thread
- For BOLT:
  - unit of collection = full threads
  - unit of annotation = individual posts and sentences within those posts
  - difficulties for downstream annotation, particularly co-reference

Co-reference

- Co-reference task identifies different mentions of the same entity (person, organization, etc.) within a post;
  - linking definite referring noun phrases and pronouns to their antecedents
  - In threaded messages, “you” often refers to a previous poster not appear explicitly named in the body text of any post.
  - In a long or complex thread, which previous poster is “you”?
- BOLT practice of selecting individual posts rather than full threads for annotation doesn’t help.
  - Post sub-selection is necessary but leads to cases where the co-reference chain is broken for a given entity

Co-reference example

Post 1: OK guys, I have a new one for you: Billy H. was to Presidents as Pluto is to Planets. Discuss.

Post 2: *OK guys, I have a new one for you: Billy H. was to Presidents as Pluto is to Planets. Discuss.* William Henry Harrison is no longer considered a President?

Post 3: B-to-the-double-H was a small, meaningless President.

Post 4: I disagree. He ran the first modern campaign for president. He had tokens made and ribbons printed up and even slogans we still remember today. "Tippicanoe and Tyler Too" refered to the General winning a battle against the Indians at Tippicanoe and his V.P John Tyler. The log house and hard cider jug on his political tokens was a slap at opponents who tried to portray him as a hard drinker.

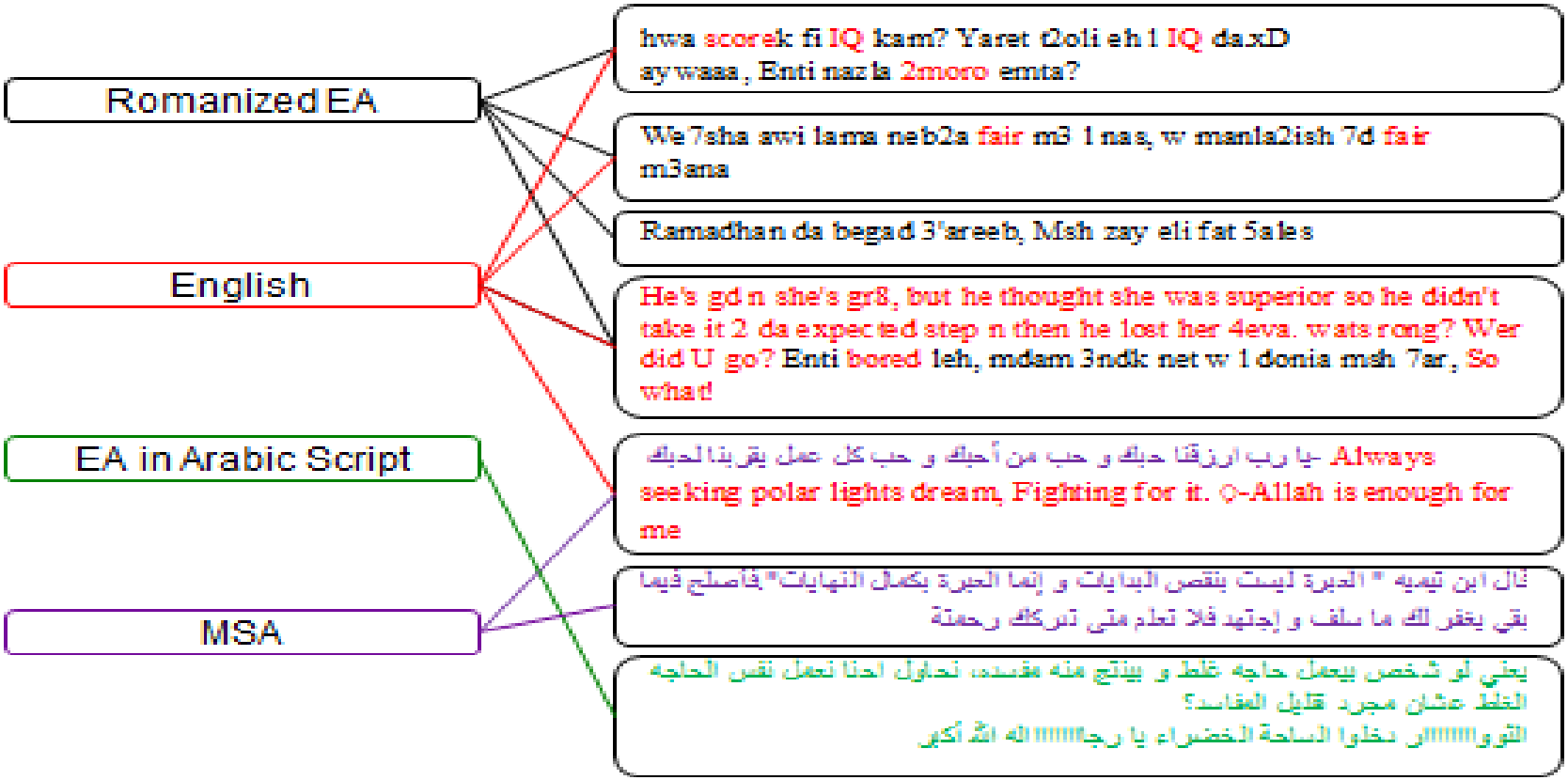
Non-standard language

- Discussion forum data is highly informal, unedited
- Non-standard variants, slang and internet abbreviations, typographical errors and misspellings (intentional or not)
  - Translation: produce a literal, meaning-accurate translation
- Special text formatting and emoticons
  - Information retrieval: a poster may follow a statement with a winking smiley emoticon to indicate a non-serious stance

Egyptian Arabic

- Orthographic variation
  - Due to diglossia, dialectal Arabic is primarily used in informal, spoken interaction. No standardized written form.
  - Multiple spellings for the same words
  - Variation in use of Arabic script and Romanized
- Codeswitching
  - Egyptian
  - MSA
  - English

Egyptian Arabic Twitter Feed



Chinese Word Substitution

- Number substitutions and homophones
- Informal/playful
  - **520**，送给所有亲人，兄弟，朋友，想我的，我想的，还有我下一位女朋友！

  I love you. My love goes to all my family, my brothers, friends, those missing me, those I miss and my next girlfriend!
- Censorship avoidance
  - 李月月鸟和温影帝比，谁家更有钱？？？

  [Li Peng | Li Yue Yue Niao] and [Wen Jiabao | Wen the Best Actor award winner], whose family is richer???

Chinese topicalization

- Once a topic is established, subjects may not be explicit
  - **Post 1:** @重庆市人民政府新闻办公室： 据悉，王立军副市长因长期超负荷工作，精神高度紧张，身体严重不适，经同意，现正在接受休假式的治疗。 转发(4776) | 评论(1429) 8分钟前 来自新浪微博

  It is reported that Deputy Mayor Wang Lijun has agreed to take vacation-style treatment due to unwellness from exhaustion and high pressure, after approval from DROP-WL.

  - **Post 3:** 软禁了哇。

  DROP-WL imprisoned?

  - **Post 7:**他是薄的人？

  Is he (Wang) in Bo's team?

  - **Post 11:** 铁杆头号手下啊！从东北带来的啊！

  DROP-WL die-hard subordinate! DROP-WL accompanied DROP-BO from North East!

Summary and Conclusion

- BOLT data
  - Very large collection of real-world data from web forums
  - Designed for variety, breadth, and volume
  - Requires new approaches for both collection and annotation
  - Initially available to BOLT performers, future publication in LDC catalog

- Thanks!
  - Questions?

This material is based upon work supported by Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-11-C-0145. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.