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

Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, Haejoong Lee
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810
Philadelphia, PA 19104 USA
E-mail: garjen@ldc.upenn.edu, strassel@ldc.upenn.edu, safa@ldc.upenn.edu, zhiyi@ldc.upenn.edu, haejoong@ldc.upenn.edu

Abstract

We describe an ongoing effort to collect and annotate very large corpora of user-contributed content in multiple languages for the DARPA BOLT program, which has among its goals the development of genre-independent machine translation and information retrieval systems. Initial work includes collection of several hundred million words of online discussion forum threads in English, Chinese and Egyptian Arabic, with multi-layered linguistic annotation for a portion of the collected data. Future phases will target still more challenging genres like Twitter and text messaging. We provide details of the collection strategy and review some of the particular technical and annotation challenges stemming from these genres, and conclude with a discussion of strategies for tackling these issues.

Keywords: Linguistic resources, collection, annotation, data centers

1. Introduction

The DARPA BOLT (Broad Operational Language Translation) Program has among its goals the development of genre-independent machine translation and information retrieval systems. While earlier DARPA programs including GALE (Olive, 2011) made significant strides in improving natural language processing capabilities in structured genres like newswire and broadcasts, performance degrades rapidly when systems are confronted with data that is less formal or whose topics are less constrained that what is typically found in news reports. BOLT is particularly concerned with improving translation and information retrieval performance on informal genres, with a special focus on user-contributed content in the early phases of the program. In the first phase of BOLT, currently underway, Linguistic Data Consortium is collecting and annotating threaded posts from online discussion forums, targeting at least 500 million words in each of three languages: English, Chinese and Egyptian Arabic. A portion of the collected data is manually "triaged" for content and linguistic features, with an optional annotation pass to normalize orthographic and linguistic variation that may prove particularly challenging for downstream (human or automatic) annotation processes. The triage process results in a selection of approximately one million words per language; this data is then tokenized and segmented into sentences with English translations produced where required. The resulting parallel text is manually aligned at the word level, and approximately half of the source data selected for translation is further annotated for morphological and syntactic structure (via Treebanking) for predicate argument structure (via PropBanking), and for entity co-reference.

Later phases of the program target similar data volumes in still more challenging genres including text messaging, chat and micro-blogs like Twitter. The data goals and performance targets for BOLT pose intensive demands, with several key factors that add appreciable risk to the endeavor, most notably an aggressive schedule for collection and annotation combined with the need to develop robust collection and annotation methods to address the inherent variation and inconsistency reflected in the informal genres that are targeted. In this paper we describe the current collection effort, review several of the linguistic and content challenges that are pervasive in this data, and discuss some of the solutions we have adopted.

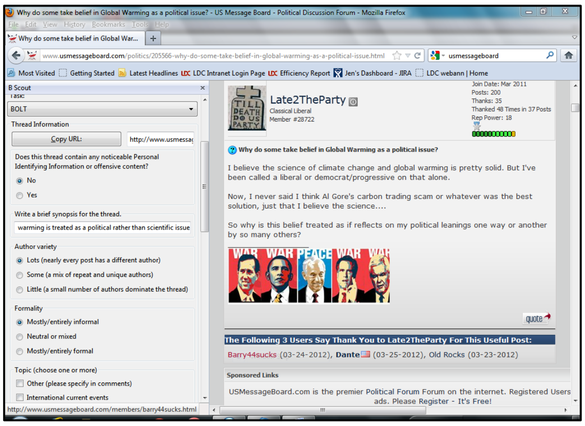
2. Collection

2.1 Data Scouting

In order to create a corpus with both a high volume of data and a reasonable concentration of threads that meet content and language requirements, we are pursuing a two-stage collection strategy: manual data scouting seeds the corpus with appropriate content, and a semi-supervised harvesting process augments the corpus with larger quantities of automatically-harvested data.

Collection of discussion forums begins with native speaker annotators who are trained in the BOLT data scouting process. These trained data scouts search for individual threads that meet BOLT requirements. Formal guidelines define basic concepts and provide detailed instructions for evaluating the appropriateness of candidate threads. For BOLT, appropriate threads contain primarily original content (as opposed to copies of a published news article, for instance); primarily informal discussion in the target language; and a primary focus on discussion of dynamic events or personal anecdotes. The data scouting guidelines also specify what types of threads or forums should be avoided.

In addition to formal guidelines, data scouting is facilitated through BScout, a customized user interface developed by LDC for BOLT. BScout is a Firefox browser plug-in that records judgments for each scouted thread, including the thread URL, a brief synopsis and an assertion that the thread contains no sensitive personal identifying information or other problematic content. Data scouts also record additional information about thread and forum properties including the level of formality and (for Egyptian scouts) the use of Egyptian Arabic versus Modern Standard Arabic. This meta-information informs the automatic harvesting process.


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
quote

Task: BOLT

Thread Information

Copy URL: http://www.usmessag

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?

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!

Figure 1: Data Scouting with BScout

The resulting 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. Similarly, when multiple forums are targeted from a single host site, that entire site is added to the harvesting whitelist.

2.2 Intellectual Property and Privacy Issues

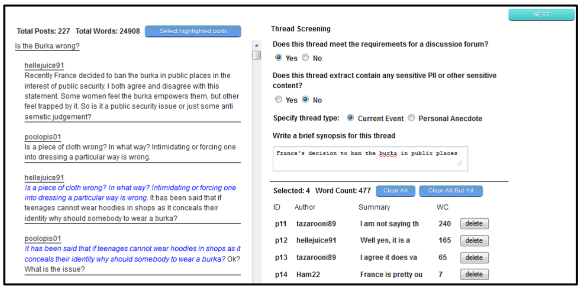
The type of data targeted presents particular challenges in the domains of copyright and contract law, privacy and objectionable content. Although web content may originate from anywhere in the world, our conservative default assumption is that all content is copyrighted, and we take additional steps to ensure that collected data can be redistributed for research, education and technology development. To further protect the privacy of data creators and to ensure that the corpus does not contain problematic content, data is manually screened for sensitive personal identifying information or other sensitive content prior to inclusion in the annotated corpus. For instance, discussion forums contain numerous credited and uncredited copies of published materials such as newspaper articles. Data scouts are instructed to exclude such content.

2.3 Triage and Segmentation

While our data scouting and automated harvesting approach supports the data volume requirements for BOLT, it also results in a certain amount of unsuitable material making its way into the corpus. While all harvested data is made available to BOLT performers, only a small subset is selected for manual translation and annotation to create BOLT training, development and evaluation sets. It is important that the data selected for annotation meets requirements for language and content; it is also highly desirable that the selected data is high-value; i.e. that it does not duplicate the salient features of existing training data. For these reasons data scouting is followed by a manual triage process. 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. Additional threads may be selected for triage based on meta-information provided by data scouts as well as other factors like number of posts, average post length and the like.

The triage task has two stages: post selection and sentence segmentation/labeling. During post selection, a native speaker annotator first confirms that the candidate thread generally meets content and language requirements and that it does not contain offensive material or sensitive personal identifying information; problematic threads are discarded from subsequent stages. The annotator then selects individual posts from the thread that are suitable for translation and downstream annotation, following selection guidelines developed with input from BOLT research sites, evaluators and sponsors. For instance, a post that consists solely of the poster agreeing or disagreeing with a previous poster, or a post that contains primarily quoted text, adds little novel content to translation training models and is therefore less appropriate for translation when compared to a post that contains novel linguistic content about an event or entity.

LDC's customized BOLT data triage user interface displays each thread in its entirety, with posts clearly separated and quoted text displayed in blue font. Annotators click on a post to select it; the list of selected posts and associated post metadata appears on the right side of the interface.



Total Posts: 227 Total Words: 24908

Is the Burka wrong?

hellejuice91
Recently France decided to ban the burka in public places in the interest of public security. I both agree and disagree with this statement. Some women feel the burka empowers them, but other feel trapped by it. So is it a public security issue or just some anti semetic judgement?

poolopis01
Is a piece of cloth wrong? In what way? Intimidating or forcing one into dressing a particular way is wrong.

hellejuice91
Is a piece of cloth wrong? In what way? Intimidating or forcing one into dressing a particular way is wrong. It has been said that if teenages cannot wear hoodies in shops as it conceals their identity why should somebody to wear a burka?

poolopis01
It has been said that if teenages cannot wear hoodies in shops as it conceals their identity why should somebody to wear a burka? Ok? What is the issue?

Thread Screening

Does this thread meet the requirements for a discussion forum?

● Yes ○ No

Does this thread extract contain any sensitive PII or other sensitive content?

○ Yes ● No

Specify thread type: ● Current Event ○ Personal Anecdote

Write a brief synopsis for this thread

France's decision to ban the burka in public places

Selected: 4 Word Count: 477

ID	Author	Summary	WC
p11	tazarooni89	I am not saying th	240
p12	hellejuice91	Well yes, it is a	165
p13	tazarooni89	I agree it does va	65
p14	Ham22	France is pretty ou	7

Figure 2: Selecting Posts for Annotation

The second stage of data triage, sentence segmentation/labeling, requires the annotator to identify and label individual Sentence Units within each selected post. A Sentence Unit (SU) is a natural grouping of words written or spoken by a single person. SUs have semantic cohesion—that is, they can have some inherent meaning when taken in isolation; and they have syntactic cohesion—that is, they have some grammatical structure. The goal of SU annotation is to provide a stable basis for later linguistic annotation activities including translation and syntactic analysis. Annotators first identify SU boundaries by marking the last word of each sentence in the post; they then classify each SU as Keep or Exclude, to indicate which sentences should be excluded from subsequent translation and annotation tasks. Excluded content may include 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.

Where possible, annotators correct automatic segmenter output rather than generating Sentence Unit boundaries from scratch. While automatic sentence segmentation is fairly accurate for more formal genres like newswire, discussion forums and other user-generated content is much more challenging. Use of punctuation and white space is highly variable; for Arabic in particular even long posts may lack punctuation entirely. This makes manual SU segmentation, let alone automatic segmentation, quite challenging. Formal SU annotation guidelines provide specific rules for locating sentence boundaries, and for handling common features like strings of emoticons.

2.4 Automatic Harvesting and Processing

In addition to the front end user interfaces designed to support manual data scouting and triage, LDC has developed a 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 presents its own unique challenges for automatic harvesting in terms of structure and formatting, so the framework assumes a unique configuration for each site.

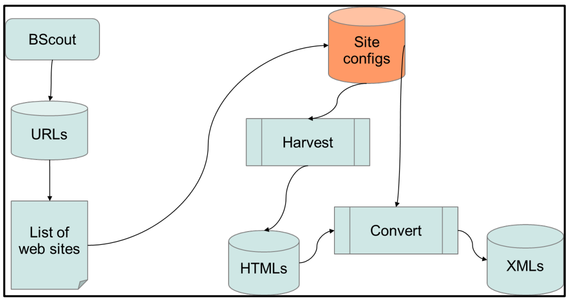
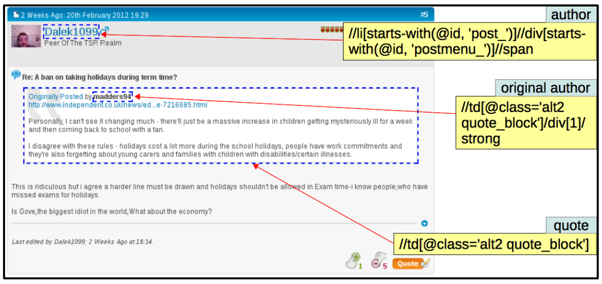


Figure 3: Harvesting and Conversion Process

URLs submitted by data scouts using BScout are first grouped by host site. 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, given a home page for a particular forum, an XPath expression is written to identify individual thread URLs contained within that page. Similarly, given a thread page, an XPath expression is written to identify the specific HTML element that contains the body text of posts. Regular expressions are used to clean up target strings. For example, when extracting the post date from the byline, extraneous strings such as "*This post was written on"* are cleaned up using regular expressions.

Once site configuration files have been developed, a harvester processes downloads individual threads, and a converter processes transforms the downloaded HTML files to an XML format. The XML format for BOLT was designed with input from research sites, and consists of a series of post elements including author, post date and post body, with additional markup to identify quoted material (to the extent that such material is consistently marked in the source HTML).


2 Weeks Ago: 20th February 2012 19:29 #5

Dalek1099
Peer Of The TSR Realm

Re: A ban on taking holidays during term time?

Originally Posted by madders94
http://www.independent.co.uk/news/ed...e-7216685.html

Personally, I can't see it changing much - there'll just be a massive increase in children getting mysteriously ill for a week and then coming back to school with a tan.

I disagree with these rules - holidays cost a lot more during the school holidays, people have work commitments and they're also forgetting about young carers and families with children with disabilities/certain illnesses.

This is ridiculous but i agree a harder line must be drawn and holidays shouldn't be allowed in Exam time-i know people,who have missed exams for holidays.

Is Gove,the biggest idiot in the world,What about the economy?

Last edited by Dalek1099; 2 Weeks Ago at 18:34.

author
//li[starts-with(@id, 'post_')]//div[starts-with(@id, 'postmenu_')]//span

original author
//td[@class='alt2 quote_block']/div[1]/strong

quote
//td[@class='alt2 quote_block']

Figure 4: XPath Expressions in Harvesting

Site configuration is often quite challenging. Many site configuration difficulties require a careful examination of the source HTML file in order to identify the problem and achieve the correct configuration. For example, 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.

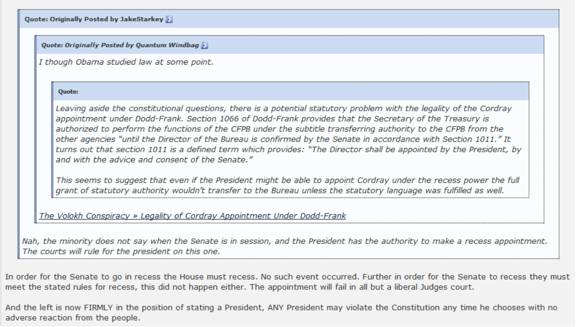
A particularly difficult (and increasingly common) challenge is harvesting host sites that use AJAX. For such sites, the downloaded HTML contains no content; i.e., there is no body text. Instead, the contents are downloaded dynamically to the web browser when the Javascript code embedded or linked on the HTML page is executed. The use of AJAX among host sites appears to be increasing over time. So far in BOLT, these sites have been dealt with outside of the standard site configuration and harvesting framework, but work is in progress to account for this emerging pattern in the generalized framework.

3. General Challenges

3.1 Quoted Text

The prevalence of quoted material in discussion forums poses challenges in both formatting and content. Quotes in discussion forums often consist entirely of content copied directly from a third party data provider, e.g. an entire newspaper article. It is also very common for forum posts to quote content from prior posts within the same thread. Setting aside issues of copyright, external quotes are undesirable for BOLT annotation because the language is primarily formal and non-interactive, while internal quotes are undesirable because the same content is likely to have been annotated previously, as part of the original post. As such, the presence or absence of quoted text is an important consideration during data triage. While quoted text is not itself an annotation target, quotes can nonetheless provide important context during annotation. Accurate representation quoted text is also important when establishing provenance during information retrieval tasks.

Posters themselves exhibit considerable variety in choosing to quote entire posts from earlier in the thread or only relevant portions. Additionally, posters may engage in complex quoting in which Poster A quotes a post from Poster B, which in turn contains a quote from Poster C and/or some external source (Figure 5).



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.

Figure 5: Multiple Embedded Quotes in a Post

Because of the importance of quotes for various parts of the BOLT data pipeline, it is highly desirable for the processed XML version of harvested threads to preserve markup for quoted text. Simply detecting the presence of quoted text in the original source data can be quite difficult given the wide range of HTML representations for quoted text, and there will be a certain number of cases in which the quote markup is missed. However, the majority of well-formed quote markups are preserved in the official XML format, including the possibility of embedded quotes-within-quotes.

3.2 Threading, Post Selection and Annotation

The threaded nature of discussion forums is of particular interest to BOLT, given the program's emphasis on informal and interactive discourse. The content of a forum thread covers multiple posters' perspectives on a topic, and individual posts are best understood in the context of the previous posts within the thread. At the same time, while the unit of collection is full threads, the unit of annotation is individual posts and sentences within those posts. This reality presents some difficulties for downstream annotation, particularly for co-reference.

The co-reference task identifies different mentions of the same entity (person, organization, etc.) within a post; this primarily consists of linking definite referring noun phrases and pronouns to their antecedents. In threaded messages, the pronoun "you" will often be used to refer to a previous poster, while that poster's name does not appear explicitly in the body text for any message. Moreover, in a long or complex thread it can be very difficult to tell which previous poster "you" refers to.

Co-reference annotation is made still more difficult by the BOLT practice of selecting individual posts rather than full threads for annotation. While post sub-selection is necessary given resource constraints and other factors, this does lead to cases where the co-reference chain is broken for a given entity. For instance, in Example 1 the second post would likely be labeled "Exclude" during triage due to the prevalence of quoted text (in italics), but ideally this post should be available for co-reference annotation since it is the only post in the thread where the entity's full name is stated.

Example 1

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.

While triage annotators are encouraged to consider such issues during post selection, such problems may only be apparent after the downstream annotation tasks have begun. To overcome this challenge, annotators for all downstream tasks are given two versions of the BOLT data to work with: an official version of each file that contains just the selected posts, and a full thread version containing all posts. Annotators can make use of the full thread version for context, and in cases like Example 1 where unselected posts contain information that is crucial for annotation, posts can be provisionally annotated and flagged for later inclusion.

3.3 Non-Standard Language Usage

Discussion forum data is of interest to BOLT largely because of its highly informal nature. Posters do not aim to produce carefully edited prose with standard spelling and punctuation. Non-standard variants, slang and internet abbreviations are common, as are typographical errors and misspellings. Some intentional misspellings have become part of standard internet language (examples from English include *kitteh* for *kitty* and *pwned* for *owned*). These non-standard uses of language present particular challenges for downstream annotation, in particular translation. Translators must preserve something of the stylistic flavor of the source text while creating a literal, meaning-accurate translation suitable for training MT systems. Other non-standard language features like special text formatting and emoticons have potential complications for other tasks including information retrieval. For example, a poster may follow a statement with a winking smiley emoticon to indicate a non-serious stance. Annotation guidelines for each BOLT task specify how such challenges are handled.

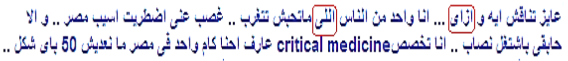
4. Language-Specific Challenges

Beyond the general challenges presented by discussion forums, a number of language-specific issues require special attention.

4.1 Egyptian Orthographic Variation

A general pattern of diglossia in Arabic leads to the use of MSA (Modern Standard Arabic) in formal settings and writing, while dialectal Arabic varieties are primarily used in informal or spoken interactions. But while colloquial varieties like Egyptian Arabic are prevalent in social media such as discussion forums, Twitter and text messaging, there is a lack of commonly accepted orthographic standards for dialectal varieties, and inconsistencies in the way people spell the same words or sounds are to be expected. An example of the orthographic variation in Egyptian Arabic is the frequent use of *alif maqsura* for *yaa* and *ta marbuta* for *haa*, which would both be considered typos or misspellings in MSA, as depicted in the boxed words in Example 2.

Example 2



عايز تناقش ايه و ازاى ... انا واحد من الناس اللى ماتحبش تتغرب .. غصب عنى اضطريت اسيب مصر .. و الا
حابقى باشتغل نصاب .. انا تخصصcritical medicine عارف احنا كام واحد فى مصر ما نعديش 50 باى شكل ..

We want to talk about what and why…. I am one of those who do not like to migrate… but I had to leave Egypt not by choice, otherwise I would continue to be a thief. I am specialized in critical medicine. Do you know how many of us are there in Egypt? We are 50 at the most.

Additionally, Egyptian Arabic is frequently written using a Romanized script, as in Example 3.

Example 3

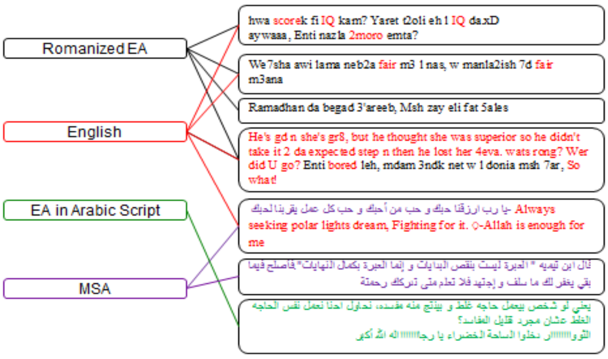
ana s2alt 3an ezay w fen a2dar aktb so2aly w 2ab3ato le2ostaz mustafa w no one answer me untill now.rabena ysam7km.

I asked how and where I can write my question and send it to Mr. Mustafa, and no one answer me until now. May God forgive you.

This reality poses an additional challenge for consistency throughout the BOLT annotation pipeline. In order to avoid the likely scenario in which annotators at different phases of the pipeline make different decisions in dealing with nonstandard representations of the language, an additional level of semi-automated annotation to normalize the Egyptian data has been designed. During this optional normalization stage, Romanized text is converted to Arabic script and all text is normalized to a single, standardized representation that is propagated down through the rest of the annotation pipeline.

4.2 Codeswitching

Along with use of multiple orthographic representations of dialectal Arabic, an additional challenge is presented by the frequent use of foreign language(s) including English and other varieties of Arabic, especially Modern Standard Arabic. Codeswitching may occur in isolation, or more commonly, in combination with the orthographic variation described above. Figure 6 below shows a portion of a typical Egyptian Twitter feed, in which English, Romanized Egyptian Arabic, Egyptian written in Arabic script, and Modern Standard Arabic are freely utilized by a single author.



hwa scorek fi IQ kam? Yaret t2oli eh l IQ da xD
aywaaa, Enti nazla 2moro emta?

We 7sha awi lama neb2a fair m3 l nas, w manla2ish 7d fair m3ana

Ramadhan da begad 3'areeb, Msh zay eli fat 5ales

He's gd n she's gr8, but he thought she was superior so he didn't take it 2 da expected step n then he lost her 4eva. wats rong? Wer did U go? Enti bored leh, mdam 3ndk net w l donia msh 7ar, So what!

يا رب ارزقنا حبك و حب من أحبك و حب كل عمل يقربنا لحبك- Always seeking polar lights dream, Fighting for it. ○-Allah is enough for me

قال ابن تيميه " العبرة ليست بنقص البدايات و إنما العبرة بكمال النهايات".فأصلح فيما
بقي يغفر لك ما سلف و إجتهد فلا تعلم متى تتركك رحمتة

يعني لو شخص بيعمل حاجه غلط و بينتج منه مفسده، نحاول احنا نعمل نفس الحاجه
الغلط عشان مجرد نقليل المفاسد؟
الثوووا!!!!!!ر نخلوا الساحة الخضراء يا رجا!!!!!!الله الله أكبر

Figure 6: Variation in a Single Egyptian Twitter Feed

English content embedded in a post that is otherwise written in Arabic orthography is simple to detect and exclude from downstream annotation. However, many Egyptian Arabic posts are written using a Romanized script, making it considerably more difficult to distinguish real English borrowings from Arabic words whose transliteration is English-like. It can be even more difficult to clearly distinguish mixing among Egyptian Arabic and other dialects or MSA given lack of diacritics in written text.

4.3 Chinese Word Substitution

Orthographic variation in Chinese is also prevalent in discussion forums due to the informal nature of the data. Common uses of nonstandard orthography include number substitutions and homophones. Example 4 shows the use of a number substitution, which is prompted by the sound similarity between the pronunciation of the numbers and the pronunciation of the words of the intended meaning. In this case, the pronunciation of 520 sounds like the Chinese for *I love you,* normally written as 我爱你.

Example 4

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!

In other cases the character for a commonly used homophonous word is substituted for the intended meaning. In Example 5, 萝卜丝 literally means radish slice, but in this context it is understood as a transliteration of Roberts.

Example 5

明明就是**萝卜丝**抓了刘翔的手、什么叫互相的拉拽？还你妹的拳击与动员、这个主持人，你是不是脑子有问题啊？

Obviously it is [**Roberts | radish slice**] who grasped Liu Xiang's hand. Where does the push and pull come from? And what is the nonsense of boxer about? Hey Anchor, are you out of your mind?

Sometimes such variations are induced by intentional substitutions of characters in order to circumvent censorship in the discussion forums. These often involve substitution via homophones for the controversial term, where the homophomes themselves have an innocuous meaning. In Example 6 below, the characters for *Li Yue Yue Niao* and *Wen the Best Actor award winner* are substituted for the potentially censorable *Li Peng* and *Wen Jiabao*, respectively.

Example 6

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???

These orthographic issues cannot be fully addressed by normalization, particularly because the current approach limits that annotation task to only a portion of the Egyptian Arabic data. Instead, annotation guidelines for each downstream task (translation, word alignment, Treebanking) provide explicit guidance on how such variants must be treated.

4.4 Topicalization in Threaded Posts

The practice of topicalization in Chinese allows the noun representing the topic or subject of a sentence to remain implicit once the topic has been established. Topicalization produces threads in which later posts may contain no explicit reference to the people, places, or events under discussion. In Example 7 below, the subject Wang Lijun is introduced in the first post; his name is not explicitly mentioned in subsequent posts. When another name, Bo, is introduced several posts later, that name also becomes implicit in following posts. In the final post in the thread, both individuals are understood to be participants but neither is mentioned explicitly. In this example, DROP-WL represents an implicit mention of Wang Lijun while DROP-BO represents an implicit mention of Bo.

Example 7

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!

There are several annotation challenges associated with topicalization. For translation, the full thread context must be carefully reviewed in order to understand the implied topic/subject(s). Word alignment and co-reference annotation also must account for the empty subject on the source side and the explicitly stated subject on the translation side.

5. Conclusion

To support the BOLT Program's goal of improved machine translation and information retrieval technologies for informal genres, Linguistic Data Consortium is engaged in collection and annotation of discussion forums and other user-generated content in three languages. The BOLT corpora described here have been designed for variety, breadth and volume. The collection target is unconstrained, real-world data, reflecting the full spectrum of quality and content of such data on the web. The scale is very large, ultimately comprising over a billion words per language. These demands have required new approaches and new frameworks for both collection and annotation.

These resources described here will initially be distributed to BOLT performers as training, development and evaluation data. We will wherever possible distribute the data more broadly, for example to our members and licensees, through the usual mechanisms including publication in the LDC catalog.

6. Acknowledgements

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.

7. References

Olive, J.; Christianson, C. and McCary, J. (Eds.) (2011). *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer New York.