PROJECTS

BOLT: Broad Operational Language Translation

BOLT Overview

The DARPA Broad Operational Language Translation (BOLT) Program aims to create new techniques for machine translation (MT) and information retrieval (IR) that can be applied to informal genres of text and speech including those commonly used in online and spoken communication. LDC supports the BOLT Program by collecting and annotating naturally occurring chat, SMS/text messaging, discussion forums and telephone conversations in three languages – Chinese, Egyptian Arabic and English.

For MT system training and development, collected data is translated and richly annotated for a variety of tasks including word alignment, Treebank, PropBank and co-reference. LDC also develops natural language queries and responses for IR system development. In support of BOLT technology evaluations conducted by the National Institute of Standards and Technology (NIST), LDC performs post-editing of MT system output and assesses relevance and utility of IR system responses.

Genre	Language	Data Volume
Discussion Forum	Chinese	1.4 billion words
	Egyptian Arabic	650 million words
	English	1.2 billion words
SMS/Chat	Chinese	2.1 million words
	Egyptian Arabic	350,000 words
	English	2.2 million words
Conversational Telephone Speech	Chinese	120 hours
	Egyptian Arabic	120 hours

Source Data

Through a combination of new data collection, harvesting of online data and re-purposing of previously collected content, LDC supplies training, development and evaluation data for each language. A new genre is introduced with each phase of the program.

- **Phase 1:** online discussion forums collected and post-processed with LDC's WebCol framework
- **Phase 2:** SMS and chat messages captured in real-time from enrolled, consented users, as well as donations of prior message archives with both conversation sides retained for naturalness
- **Phase 3:** informal telephone conversations from prior LDC collections

Large Scale Translation Resources

To satisfy program demands for large volumes of high quality parallel text, LDC produces translations of collected foreign language text into English. Up to 1.2 million words of collected data per genre, per language are selected for manual translation. Selected data is then segmented into sentence units, message units and/or speaker turns; this step ensures that the resulting parallel text is aligned at the segment level.

Translators follow formal specifications developed for BOLT that address both general principles and genre- or language-specific issues. Given the highly informal and conversational nature of the BOLT data sources, special guidance is provided for translating internet slang, abbreviations and the like.

For idiomatic expressions, we provide a literal translation and a translation designed to capture the intended meaning in fluent English.

Example of Translation Alternatives for Idiomatic Expressions

Original Chinese: 猴年马月
Intended meaning: *God knows how long*
Literal meaning: *monkey year and horse month*

Resulting translation in context: Cut off the reporting when the time limit is reached. Otherwise, *[God knows how long | monkey year and horse month]* it will take.

Linguistic Data Consortium 3600 Market St., Suite 810, University of Pennsylvania, Philadelphia, PA 19104-2653 USA
Telephone: +1.215.898.0464 • Fax: +1.215.573.2175 • ldc@ldc.upenn.edu • www.ldc.upenn.edu

All translations undergo several rounds of quality control, with the exact procedure dependent on the intended use of the data. For instance, evaluation data is subject to more thorough review than training data since the human translations serve as a "gold standard" for system evaluation.

During evaluation of MT system output, LDC post-editors compare automatic translations sentence by sentence against manual gold standard translations. The post-editors make changes to the MT output to achieve identical meaning with the human translation, using the fewest possible number of edits.

Rich Annotation

A portion of the data designated for translation is further annotated in collaboration with our data partners.

Transliteration (LDC and Columbia University): produces an Arabic orthographic version of the collected data and normalizes the spelling to facilitate morphological analysis and subsequent annotation, in order to address the prevalence of Romanized script (Arabizi) in the Egyptian Arabic SMS and chat data.

Word Alignment (LDC): captures translation correspondences between parallel sentences, resulting in links between individual words, phrases and groups. Tokens that do not have any match in the parallel sentence are explicitly marked, and links may be categorized for their syntactic or semantic function.

Treebanks (LDC and Brandeis University): fully parsed corpora that are manually annotated for syntactic structure at the sentence level and for part-of-speech or morphological information at the token level. Every token in every sentence is annotated. Treebanks support the creation and training of parsers and taggers, work on machine translation and speech recognition, and research on joint syntactic and semantic role labeling.

PropBank (Brandeis University and University of Colorado): annotating the semantic roles of a given predicate's argument. It creates a corpus of text annotated with information about basic semantic propositions. Annotation is done not only for verbal propositions but also eventive nouns and adjectives. This allows consistent argument labels across a verb sense and also between nominal and adjectival counterparts, such as between *decide* and *decision* in English.

Co-reference Annotation (Raytheon BBN Technologies): captures the part of human language interpretation that links definite references in the text to the respective entities in discourse. Annotators link together names, pronouns and definite descriptions that refer to the same entity, providing information that is crucial for systems doing semantic interpretation. Noun phrase mentions of events are also linked to verb phrases that describe the event. The null pronouns found in Chinese are included in the co-reference annotation and when speaker turn information is available, the speaker names are also included.

Information Retrieval

The BOLT IR task requires systems to process a set of natural language queries written in English; locate and extract short answers from a large collection of multilingual discussion forum threads; and translate foreign language answers into English. LDC produces BOLT IR queries for system training and testing and assesses system output along multiple dimensions including relevance and utility. Pilot assessment tasks also investigate redundancy in query responses and the impact of user interactions on system performance.

Query Examples

- *What do people think about Pope Shenouda III?*
- *What happens if you get addicted to the internet?*
- *Are there weapons stockpiled in Coptic churches?*
- *Does going to private school help you get a good job?*
- *Should smoking be allowed in public areas?*
- *Why do professional athletes get paid so much?*

Publication of BOLT Data

Virtually all of the resources created for BOLT will be published in LDC's catalog.

For more information visit: **www.ldc.upenn.edu/collaborations/current-projects/bolt**