NEW MEASURES OF EFFECTIVENESS FOR HUMAN LANGUAGE TECHNOLOGY

The field of human language technology (HLT) encompasses algorithms and applications dedicated to processing human speech and written communication. We focus on two types of HLT systems: (1) machine translation systems, which convert text and speech files from one human language to another, and (2) speech-to-text (STT) systems, which produce text transcripts when given audio files of human speech as input. Although both processes are subject to machine errors and can produce varying levels of garbling in their output, HLT systems are improving at a remarkable pace, according to system-internal measures of performance. To learn how these system-internal measurements correlate with improved capabilities for accomplishing real-world language-understanding tasks, we have embarked on a collaborative, interdisciplinary project involving Lincoln Laboratory, the MIT Department of Brain and Cognitive Sciences, and the Defense Language Institute Foreign Language Center to develop new techniques to scientifically measure the effectiveness of these technologies when they are used by human subjects.

Research in human language technology (HLT) has made great progress in the past few years. The general field of HLT encompasses a wide range of algorithms and applications dedicated to processing human speech and written communication. The two specialized HLT fields we address here, from the perspective of the needs of the Defense Department, are (1) machine translation systems, which convert text and speech files from one human language to another, for example, from Arabic into English, and (2) speech-to-text (STT) systems, which produce text transcripts when given audio files of human speech as input. Both processes are subject to machine errors and may produce varying levels of garbling in their output.

Speaker-independent automatic STT systems are currently capable of producing English text transcripts of conversational telephone speech at a word-error rate of 15.2%, an error reduction of 53% over the past five years. Arabic-to-English machine translation systems are capable of producing English text output at a precision rate of 51% for a weighted word-sequence recognition measure known as the BLEU score [1], an increase in performance of over 300% over the past three years [2]. These measures of performance are defined in technology-centric terms to help guide research and development. As important as these measures are, they do not say what the technology can do for us. In military parlance, measures of performance refer to system-internal tests, whereas measures of effectiveness refer to tests that measure how the technology is used in real-world settings. A major goal of current Department of Defense–sponsored research in machine translation of human speech and text is

رأي الأهرام
يبدو أنّ تصريحات وزير خارجية الحكومة العراقية المعينة، التي هدد فيها بإمكان السماح للقوات الأمريكية التي تحتل العراق بشن هجمات على البلدان المجاورة للعراق، هي ردّ على دعم تلك البلدان لعمليات المقاومة العراقية.

Al-Ahram Al-view
Foreign minister's statements appear to be designated the Iraqi government, which threatened the American forces, which could be allowed to occupy Iraq to launch attacks on neighbouring countries to Iraq in retaliation for its support for Iraqi resistance.

Al-Ahram Opinion column
Declarations by the Minister of Foreign Affairs in the "appointed" Iraqi government threatening to allow the American forces that occupy Iraq to launch attacks on neighboring countries seems to be in response to their support of the Iraqi resistance.

FIGURE 1. Arabic-to-English translation of a moderately difficult Arabic text: machine translation system (left) versus human translation (right). Two sample questions that a competent reader of Arabic could be expected to answer, after reading the full passage, are "list one of the author's expectations for the new Iraqi regime," or "why does the author consider the minister's remarks dangerous for the Iraqi government?"

to build automatic systems that translate Chinese and Arabic texts into English texts that can be used by English native readers to perform pertinent foreign language tasks.

Figure 1 shows a fragment of an Arabic Level 3 text (moderately difficult for non-native speakers), rated according to the Interagency Language Roundtable skill levels. Sample questions that a competent Arabic reader could be expected to answer after reading the full text include "list one of the author's expectations for the new Iraqi regime," or "why does the author consider the minister's remarks dangerous for the Iraqi government?"

The translation on the left in the figure is produced by a state-of-the-art machine translation system; the translation on the right in the figure was produced by professional human translators. The machine translation contains a variety of disfluencies and mistakes. These errors impair our ability to understand facts explicitly stated in the Arabic article, and severely degrade our ability to infer ideas that are not specifically stated but that an educated reader would be expected to make after reading the article in the original language. The professional human translation on the right contains enough nuances for us to begin to understand the point of view of the original author and to make relevant inferences. The decision of whether to use machine translation or human translation depends upon factors such as the cost, the availability, and the level of quality needed for the task at hand.

Likewise, a major goal of Defense Department sponsored research in speech recognition is to provide high-quality automatic speech-to-text (STT) transcripts of news broadcasts and conversational telephone speech. Unlike dictation tasks, for which an STT system may be trained and tuned to a specific speaker's voice to reduce word recognition errors, these systems need to operate independently of the speaker. The goal is that the transcripts are good enough to use for tasks that would ordinarily be accomplished by listening to the speech broadcasts or recordings. However, ordinary STT transcripts not only contain word recognition errors, but they are typically produced in single case, lack punctuation, and include verbatim every spoken word or word fragment, including fillers such as "um," "uh," repeats, false starts, and so on.

Figure 2 illustrates the two potential ways we might view the audio signal from a telephone conversation: as a transcript produced by an experimental STT system (left), and as a reference transcript (right) produced by trained human readers and cross-checked with quality controls (i.e., there are no transcription errors in the texts, standard punctuation and capitalization have been applied, and disfluencies have been removed).

actually uh i belong to a gym down here a gold jim uh i exercise so i tried exercise five days a week uh i usually do that what took said can you imagine

A: Yeah I belong to a gym down here. Gold's Gym. And I try to exercise five days a week. And I usually do that.
B: What type of exercising do you do in the gym?

FIGURE 2. Speech-to-text (STT) transcripts from the audio signal of a telephone conversation: experimental system (left) versus human transcript (right). The human transcript is the gold standard by which the experimental results are measured.

The human transcription on the right in Figure 2 is the gold standard by which the automatic SST system output is scored. It also serves as a target standard for technology research and development. The benefits of improving automatic speech transcripts fall into two categories: (1) making the transcripts more readable for human readers and (2) improving automatic downstream processes that use these transcripts as input. Our focus here is on the human readers.

For both technologies—machine translation and STT systems—the human transcripts and translations are clearly easier to read. What is not clear is the level at which these technologies can enable their intended consumers (e.g., analysts) to perform real tasks. In order to quantify the effectiveness of these technologies and provide feedback for research programs such as the Defense Advanced Research Projects Agency (DARPA) EARS Program (Effective, Affordable Reus-

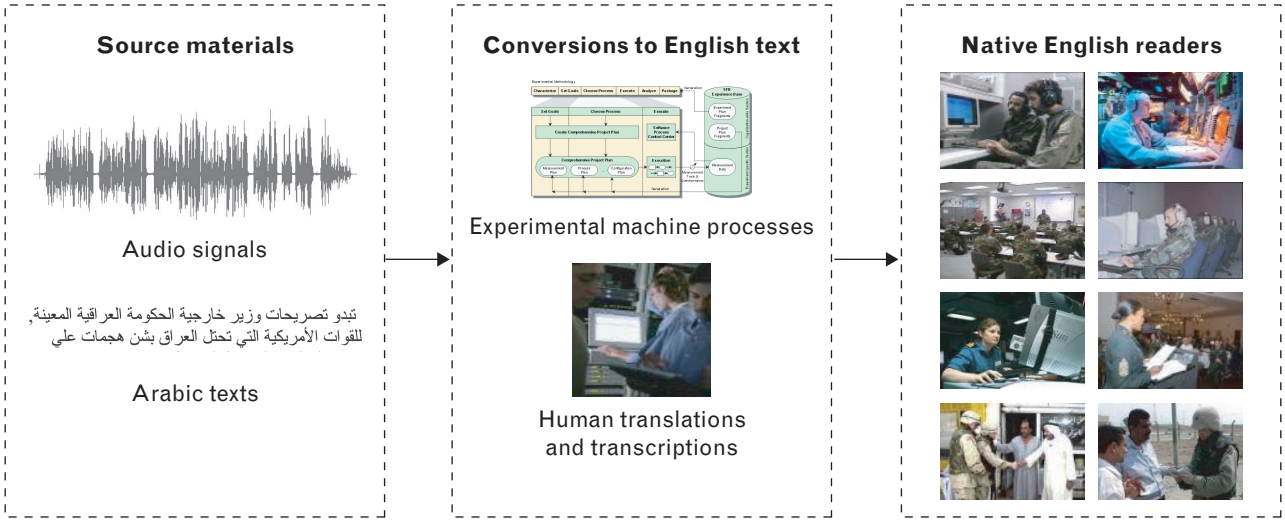

FIGURE 3. The framework for measuring human language technology output effectiveness for English readers. Source materials in audio or text forms are converted to English text by using machine translation or human translation. The English text output is then supplied to a variety of English readers. We contrast the ability of people to process the output of experimental language processing algorithms with baselines that use manual processing of human language data.

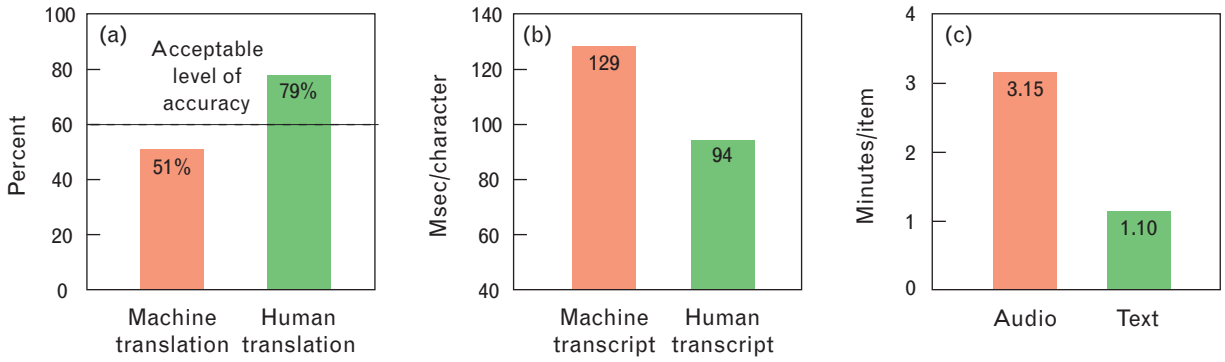

FIGURE 4. Quantitative effects of comprehension for machine translation and human translation. Three types of experiments are shown, defined in terms of (a) translation comprehension based on the Defense Language Proficiency Test, (b) speech transcript reading time, and (c) speech transcript scanning time.

able Speech-to-text) and the DARPA TIDES Program (Translingual Information Detection, Extraction, and Summarization), we launched a related project that applies rigorous psycholinguistic experimentation and government-standard proficiency evaluation techniques to the methodology of these research efforts.

Figure 3 shows our general framework for determining measures of effectiveness for HLT output. Source materials exist in many forms: English audio signals from broadcast news, conversational telephone speech, foreign language texts from newspapers, and so on. These materials are converted to standard English text (either by translation or transcription) and then presented to human subjects who are asked to read and answer comprehension questions. We present texts that were generated by machines and those generated by humans, and then we measure a human reader's ability to answer questions about the text and the speed at which that reader is able to process the text. With gold-standard texts originally processed by humans, we expect a higher level of performance and faster reading times from our human subjects. With experimental machine system output, however, errors can occur, yielding texts that are garbled in places, resulting in a lower level of performance and higher reading times.

By using results from both machine- and human-generated output we can quantify the degradation that these automatic translation and transcription techniques introduce, in a variety of test conditions (such as ranges of errors, types of transformations, and so on). Figure 4 shows results from three different experiments. The graph on the left shows the results of a machine translation experiment in which machine translation output fails to enable people to pass a utility threshold (in this case, 70%) in question-answering accuracy on a standardized test. The graph in the center shows that speech transcripts full of errors are read more slowly (129 msec/char versus 94 msec/char, a slowdown of 37%). The graph on the right shows that a simple text-classification experiment can be accomplished approximately three times faster when readers skim a gold-standard transcript, compared with when they listen to the audio files.

Each experiment employs a different measure of effectiveness. The machine translation experiments are based on test results using a modified Defense Language Proficiency Test (DLPT) for Arabic. The DLPT is a standardized test used in the U.S. Defense Department to measure foreign language skills. It has been administered for decades to assess the suitability of personnel for missions requiring foreign language skills, and it has undergone rigorous scrutiny in its design. We modified the DLPT test in the following way: instead of presenting the human subjects with the original Arabic test materials, we substituted English translations produced by machine translation systems (the test condition) and by professional translation services (the control condition).

Our modified DLPT contained texts rated at Interagency Language Roundtable (ILR) levels 1, 2, and 3. At least 70% of the questions must be answered

correctly to pass a given level. We found that subjects generally passed Levels 1 and 2, meaning that they demonstrated language survival skills and could understand basic facts, but they generally failed Level 3, meaning that they were not able to comprehend abstract linguistic formulations, read between the lines, and make inferences. The other experiments were based on more generic measures of effectiveness (how fast subjects read the texts, how accurately they answer general questions, and how strongly they prefer materials in different conditions).

The next steps for our project include extending our measures of effectiveness to other areas of HLT evaluation (for example, how effectively Mandarin Chinese and Arabic audio and text files can be distilled for particular tasks) and establishing relationships between different modalities (for example, measures of effectiveness for speech-to-speech machine translation devices to be used to accomplish tasks that require interactive dialog between people who do not speak a common language).

This work is a joint, interdisciplinary effort between the Defense Language Institute Foreign Language Center, Lincoln Laboratory's Information Systems Technology group, and the MIT Brain and Cognitive Sciences Department. The principal investigators include Neil Granoien and Martha Herzog and their colleagues at the Defense Language Institute who have served as the principal architects for foreign language proficiency tests for the U.S. Defense Department; Professor Edward Gibson, an established leader in the field of psycholinguistics and human sentence processing; and Douglas Jones and Wade Shen and other members of the technical staff at Lincoln Laboratory. We gratefully acknowledge Jurgen Sottung, Michael Emonts, Osaila El Khatib, and Hussny Ibrahim at the Defense Language Institute, as well as John Tardelli and Paul Gatewood at ARCON Corporation for their assistance in conducting experiments. We also thank our colleague Douglas Reynolds at Lincoln Laboratory for helpful advice, and Charles Wayne and Joseph Olive at DARPA for guidance and sponsorship.

This research project represents an opportunity to bring the technology development community into better contact with the two neighboring fields of psycholinguistics and foreign language training. Since 2002 we have conducted experiments with the participation of hundreds of human subjects at MIT and the surrounding communities, and have written several papers describing our results [3–7].

REFERENCES

1. K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, N.Y. (17 Sept. 2001).
2. NIST 2005 Machine Translation Evaluation Official Results, 1 Aug, 2005 <http://www.nist.gov/speech/tests/mt/mt05eval_official_results_release_20050801_v3.html>
3. D.A. Jones, W. Shen, N. Granoien, M. Herzog, and C. Weinstein, "Measuring Translation Quality by Testing English Speakers with a New Defense Language Proficiency Test for Arabic," *Proc. 2005 International Conf. on Intelligence Analysis, 2–4 May 2005, McLean, Va.*
4. D.A. Jones, E. Gibson, W. Shen, N. Granoien, M. Herzog, Do. Reynolds, and C. Weinstein, "Measuring Human Readability of Machine Generated Text: Three Case Studies in Speech Recognition and Machine Translation," *2005 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* **5,** *18–23 Mar. 2005, Philadelphia, Pa.,* pp. 1009–1012.
5. D.A. Jones, W. Shen, E. Shriberg, A. Stolcke, T. Kamm, and D.A. Reynolds, "Two Experiments Comparing Reading with Listening for Human Processing of Conversational Telephone Speech," *Proc. Interspeech 2005, 9th European Conference on Speech Communication and Technology, Lisbon, 4–8 Sept 2005.*
6. D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman, "Measuring the Readability of Automatic Speech-to-Text Transcripts," *EUROSPEECH 2003, 8th Conf. on Speech Communication and Technology, Geneva, 1–4 Sept. 2003.*
7. R. Clifford, N. Granoien, D. Jones, W. Shen, and C. Weinstein, "The Effect of Text Difficulty on Machine Translation Performance—A Pilot Study with ILR-Rated Texts in Spanish, Farsi, Arabic, Russian and Korean," *Proc. Int. Conf. on Language Resources and Evaluation, Lisbon, 26–30 May 2004.*

Contributed by Douglas Jones, Wade Shen, and Clifford Weinstein of Lincoln Laboratory; Edward Gibson, Florian Wolf, Rina Patel, and Charlene Chuang of the MIT Department of Brain and Cognitive Sciences; and Neil Granoien, Ray Clifford, and Martha Herzog of the Defense Language Institute.