Machine Translation for Government Applications

Douglas Jones, Wade Shen, and Martha Herzog

The idea of a mechanical process for converting one human language into another can be traced to a letter written by René Descartes in 1629, and after nearly 400 years, this vision has not been fully realized. Machine translation (MT) using digital computers has been a grand challenge for computer scientists, mathematicians, and linguists since the first international conference on MT was held at the Massachusetts Institute of Technology in 1952. Currently, Lincoln Laboratory is achieving success in a highly focused research program that specializes in developing speech translation technology for limited language resource domains and in adapting foreign-language proficiency standards for MT evaluation. Our specialized research program is situated within a general framework for multilingual speech and text processing for government applications.

» **Thousands of languages are spoken in** the world. The commercial sector provides significant translation capabilities for a few dozen of these, both in the form of trained professionals and in the form of machine translation (MT) systems. However, the U.S. government has many unmet foreign-language requirements that are likely to remain unmet by the commercial sector. Limited resources have affected military operations in Iraq and Afghanistan, where people speak several dialects of Arabic and Kurdish, as well as Pashto, Dari, and Farsi, among other languages. A reasonably short list of languages required for accomplishing the military's needs more broadly would include Chinese, French, Hindi, Indonesian, Japanese, Portuguese, Punjabi, Russian, and Urdu. The list of other language translation needs is potentially very long and depends on specific missions and needs. Tactical missions may require face-to-face communication, whereas strategic intelligence operations may involve processing large volumes of language data.

The Defense Advanced Research Projects Agency (DARPA) provides substantial funding for MT. The DARPA Global Autonomous Language Exploitation (GALE) program has developed media-monitoring systems for foreign languages, focusing on Arabic and Chinese. These systems can monitor an Arabic Al-Jazeera news program and provide automatic translations into English of high enough quality that output can be examined by English-speaking personnel and routed for subsequent detailed processing by a linguist if necessary. The DARPA Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program is developing portable two-way speech-to-speech translation systems that operate on

laptops and handheld devices. The translations contain errors, but the systems show strong promise in specialized domains (e.g., operating in a laboratory environment using vocabulary for which the systems have been trained). Outside the specific and perhaps narrow band for which MT systems have been trained, performance degrades substantially for all types of systems.

Lincoln Laboratory has been involved in MT programs since the late 1990s, beginning with DARPA and later also the Air Force Research Laboratory. We have been involved in rigorous evaluation of MT systems since the Defense Language Institute initiated our research program in 2002, and we have extended the effort for the Department of the Army and for the Army Sequoyah program. The Army Sequoyah program was formed to serve as the source of MT technology for the Army as well as for the joint services, and is scheduled to have a fifteen-year timeline, from FY2009–2024. Lincoln Laboratory has been called upon to help establish evaluation measures for Sequoyah, primarily in terms of evaluation techniques needed to assess the utility of MT for military tasks. In addition, we are developing a consistent procedure for the training cycle, which is described in the sidebar "Translation Research Cycle" on page 44.

Lincoln Laboratory's Niche

To address some of the challenges inherent in the current state of the art in MT, Lincoln Laboratory is creating specialized, custom systems for our government sponsors. High-quality translation in unlimited domains remains a hard challenge. Many important languages lack the data resources needed for statistical modeling. Our response at Lincoln Laboratory is to create smaller domains, in other words, to aim for custom solutions for specific missions. That is a more tractable approach, especially for languages for which large-scale bilingual training data is simply not available.

To meet these challenges, we have built several custom speech translation systems. Our sponsors provide specialized parallel corpora that pertain to defense department needs. We specialize in techniques that work well for small-domain, custom systems, with an emphasis on speech translation. Our systems place particular emphasis on (1) training models with less data and (2) MT from noisy inputs, in particular, error-full speech recognizers.

DARPA has invested over eight years of effort and millions of dollars into creating parallel corpora adequate for translation of Arabic broadcast news reports as part of the Translingual Information Detection Extraction And Summarization (TIDES) and GALE programs. Instead of approaching the wide domain of broadcast news, our system attempts to learn domain-specific language using much smaller amounts of data (typically, one to two orders of magnitude less). The methods we've developed can be used to build MT capability faster and more cost-effectively for situations where the technology is most needed. In coming journal articles, we will describe methods that extend current statistical approaches that deal with limited training data.

We have also focused on the problem of translation from noisy inputs (e.g., speech translation). Typically, speech recognizers are given a speech signal as input and return a hypothesis: a textual representation of the words that are spoken. If the recognizer's hypothesis is error free, we can simply give its output to an MT system to be translated. Unfortunately, state-of-the-art speech recognition is typically quite error full (a typical lower bound: one out of every ten words is wrong). We have been working on methods that make use of multiple hypotheses to improve performance. In multiple international evaluations, we have shown that these methods can significantly improve speech translation performance.

The Challenge of Translation

Translation is an extremely challenging problem, even when people perform the task. Not only does a foreign language have a different word for practically every lexical item in English, the word order may be completely different, or word order may be unimportant with more significance resulting from inflections or choice of words. The multiple meanings attached to a word in one language will rarely overlap in another; and, more generally, words will be ambiguous in different ways in different languages. Furthermore, most languages are constantly changing at different rates and in different ways. The precise meaning of a sentence may depend on whether it was written yesterday or during the previous decade. People require many years of professional training to be good translators.

Defining a mechanical process for translating one human language into another has remained an unsolved problem since Descartes expressed the concept in 1629 [1, 2]. It is perhaps worth mentioning that Descartes's

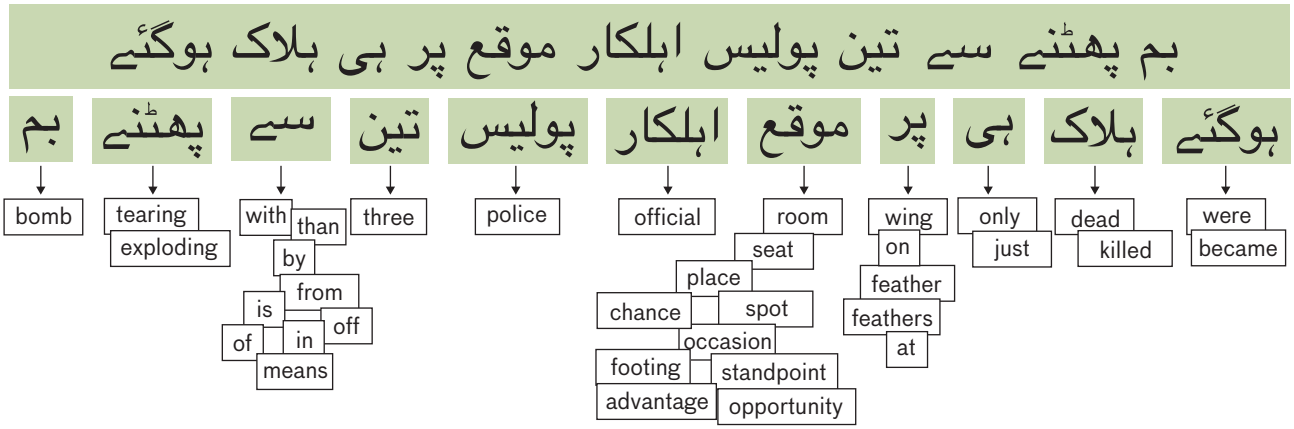


FIGURE 1. The first pass in written translation (here from Urdu to English) is to assign word translations to the individual components. The magnetic poetry concept will list all possible translations. Note the Urdu is read right to left, so the individual Urdu words on the second line have been transposed.

exploration could consider human translation as a subset of MT, in the sense that he asks whether humans could be considered to be machines, a question outside the scope of this article.

Since the advent of digital computers, researchers have hoped to leverage stronger and stronger computing power to solve the translation problem. On the one hand, machines have many advantages: they can operate without fatigue, they are great at memorizing, and they can be duplicated in order to divide and conquer translation workloads. On the other hand, machines are at a great disadvantage: they do not have the general intelligence, cultural awareness, and innate linguistic abilities that people have. In any case, the essential task for MT systems is to take input text or audio from one human language and to provide text or audio in another language, along with some degree of error.

Consider the Urdu sentence in Figure 1. The "literal" translation of each Urdu word was found in an Urdu-English dictionary and presented in the white boxes. Even though these words are all English, and they are possible translations on a word-by-word basis, the translation is still garbled. In Figure 2, the various possible translations are disambiguated, and the sentence is starting to make sense in the middle of the figure. However, the relationship between the words is not clear. We still cannot tell whether the police exploded the bomb—perhaps they were part of a bomb squad and that was their job—or did the bomb explode and kill the police? At the bottom of the figure, the words are in order and make sense: "Three police officers were killed on the spot by a bomb exploding." The challenge of MT is to perform all of these steps in an automated fashion.

What would be required in order to produce the output in Figure 2 automatically? If each Urdu word were labeled with an index indicating which position it should occupy in an English sentence, the challenge would be easier. The problem is that people constantly use new sentences. It is not feasible or even possible to store all possible sentences in advance. Even if it were possible

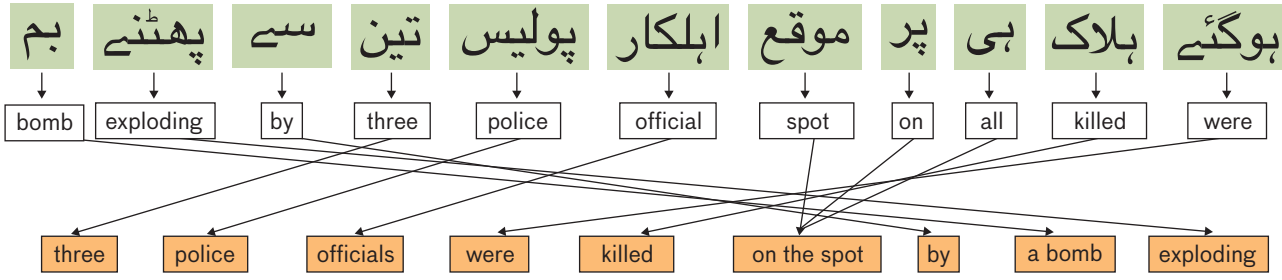


FIGURE 2. In many cases, the order of a sentence in another language is not the same as in English (subject-verb-predicate). The final steps in the translation are selecting the proper words from the magnetic poetry and then placing them in their proper order.

Translation Research Cycle

Complete the loop of training and processing

We have structured our effort at Lincoln Laboratory as a feedback loop in the research cycle, as in Figure A, which separates the data flood and the military operators with several stages of multilingual language processing. We view the multilingual language processing in terms of these three classes of technology:

- Metadata processing
- Content processing
- Analysis processing

Once we have constructed an application, we subject the technology to evaluation, followed by deployment, and the cycle starts again, with the benefit of lessons learned. We have found it very informative to connect the technology evaluation methods to methods used to assess human abilities to do jobs as they are currently being performed. Following are examples of each of these types of technology.

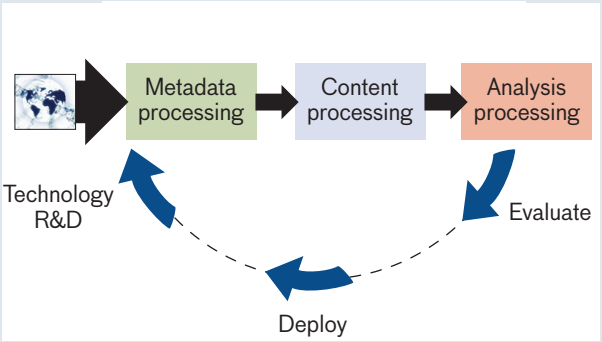


FIGURE A. The research cycle of multilingual language processing is a continuous processing loop of training, evaluating, and revising.

Metadata Processing. Metadata means "information about data." An important example of speech metadata is the identity of the human language being used in the communication. Being able to distinguish between Arabic and non-Arabic spoken language is one example. Although it is simple to state the problem set, highly advanced technology is needed to perform this function. Lincoln Laboratory has been the lead research site for speaker identification, language identification, and dialect identification technology for over a decade [a].

Content Processing. Another class of technology deals directly with the content of the message. One example is automatic speech recognition, which produces written transcripts of written or spoken language. Another is machine translation (MT), which consumes a foreign-language input and produces English text or audio, for example. It is important to know that the output is going to be garbled to some extent because the correspondence between human languages is not exact and machine processing of human languages will always have errors. We should be wary of anyone who says that this problem has been solved.

Analysis Processing. Although skepticism toward performance claims is generally a good idea, a technology-planning approach we like to take is to envision that a particular problem actually has been solved: what happens next? How would that success affect workflow downstream? Assume for the moment that the translation problem really has been solved and that now vast libraries of translated data are available. There is still nontrivial work that needs to be done. An advanced decision support program would help in the intelligence analysis work itself. For example, upstream processes could provide language identification, as we have seen, and this output goes to MT. Once the system gets the data in good, usable English, it extracts the critical elements of information needed for analysis. One component could perform a social network analysis and other advanced information processing. However, for the foreseeable future, the translation problem itself is likely to require significant resources and work in order to advance the state of the art. The main article focuses on the challenges of MT.

a. W. Shen, N. Chen, and D. Reynolds, "Dialect Recognition using Adapted Phonetic Models," *Interspeech 2008: 9th Annual Conf. Int. Sp. Comm. Assoc.*, Brisbane, Australia, Sept. 22–26, 2008.

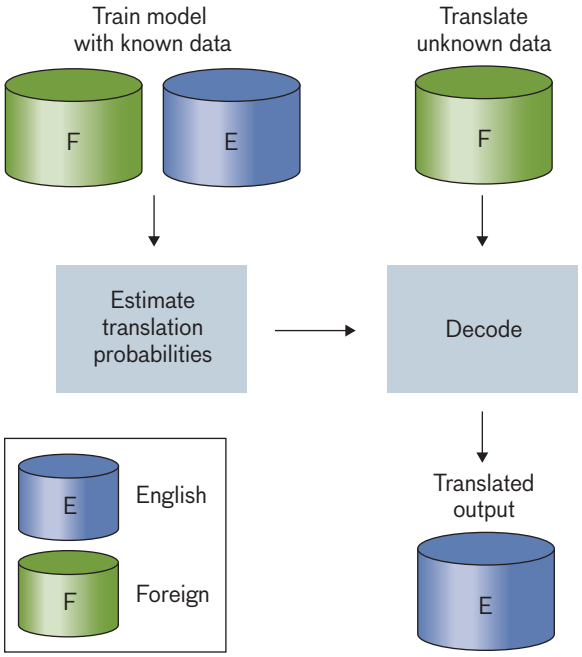


FIGURE 3. Accumulating a large volume of known language translation data provides a statistical basis for improving the translation probabilities of unknown data.

to order the words correctly in translation, each of the source words is ambiguous, which means that additional work is required in order to disambiguate the correct word needed for translation. The word order used in one language can control or add to the meaning, but it often creates confusion when replicated in a second language. The general observation is that languages express ideas differently, and that language is ambiguous. Because of these factors, we must be willing to tolerate some degree of error in the MT process.

Statistical Machine Translation

The current state of the art in MT is to train statistical models that give the most likely translation for each sentence. With enough training data, the number of errors can be reduced. The diagram in Figure 3 illustrates how MT systems work at a very high level of abstraction. MT starts out with a large set of sentence pairs, with each pair consisting of one English sentence and a foreign-language translation of that sentence, represented by the E and F drums in the figure. A typical system may be trained on millions of such sentence pairs, known as a parallel aligned translation corpus. The first such large-scale statistical translation system was based on millions of sentences from the Canadian parliamentary proceedings. The system uses statistical modeling to discover correlations between the source and target sentences at the word level in a stage known as the training process. For the translation process, the correlations are used to decode the foreign-language input into English output.

Parallel Corpora

The parallel corpus is a very simple concept. Figure 4 shows a sample with three sentences out of the millions that an actual system needs. The system counts how often each English word occurs in association with each Chinese word at the sentence level in translation. Notice the three instances of *border*, on the English side. On the Chinese side, there are three corresponding instances of 旁邊, pángbiān, which means border. After the system is trained, it will associate the Chinese word 旁邊 with the English word *border*. Notice, however, that if all of the sentences had included an equal number of the words 中国, Zhōngguó which means China, and 旁邊 (*border*), the word counts would not have been able to distinguish the likelihood of the correct translation from the incorrect one, i.e., translating *China* as *border*, or vice versa. By relying upon vast quantities of parallel texts, the chances of such accidental associations are reduced.

Mathematical Model

The mathematical model is simple to state. The object is to find the most probable English translation, given a foreign-language sentence. The model is constructed based on observations made in advance. The system searches for the most probable foreign sentence, given a known English sentence in the parallel aligned corpus. It takes into account the likelihood of the input sentence. In the usual fashion, probabilities are estimated by simply counting the relevant events, in this case, the phrase alignments that have been discovered in the parallel corpus.

Sample parallel sentences

Target sentences	Source sentences
中國在印度的旁邊	China is at India's border.
越南在泰國的旁邊	Vietnam is at Thailand's border.
中國在越南的旁邊	China is at Vietnam's border.

FIGURE 4. This example of a set of known data, with the parallel nature of the sentence structures, provides clues to individual word translations.

Machine Translation Evaluation

As challenging as high-quality MT systems can be to build, it is also challenging to evaluate them reliably. In engineering a system, researchers usually have some method of testing the system to see if it gets better over time. System-internal testing methods are known as measures of performance. How can we evaluate MT output? A binary decision of good versus bad is too coarse and subjective. A common method of evaluation a decade ago was to maintain a list of test sentences and to rely upon human judgments to determine whether the proportion of good translations increased or decreased after a substantial system change. Current methods are much faster: rather than manually judging translation outputs, a set of trusted reference translations is prepared in advance, and the MT output is mechanically scored against the reference translations. The sidebar titled "Machine Translation Evaluation" on page 47 describes some of the more recent evaluation tools.

BLEU Scores

One of the most widely used automatic scoring techniques is the BLEU scoring algorithm. BLEU, which stands for Bilingual Language Evaluation Understudy, was developed by researchers at IBM in 2000 and made available for public use [3]. Beginning in 2002, it was adopted by the National Institute of Standards and Technology (NIST) in their annual MT evaluations for Arabic and Chinese. The idea is to compare the MT output with a trusted reference translation produced by qualified professional translators and count the number of matching words and connected word sequences. This process does not involve a direct comparison with the original foreign-language source material. Figure 5 shows the number of matches in two sample MT outputs as measured against the reference translation at the top of the table. The BLEU formula rewards longer matches with a higher score: a four-word sequence counts for more than four one-word sequences. See the reference section for a more detailed discussion.

Measures of Effectiveness

The problem with count-based measures of performance is that they do not automatically provide a sense of what can be done with the MT output. If half of the words are correctly translated, is the output useful? There is no single answer to this question because it is too broad. Referring again to the MT output in Figure 5, we note that the difference between the 2002 output and the 2004 output seems like more than a purely quantitative difference in the number of correctly translated word sequences. For the 2002 output, the words seem to belong together, but they do not make much sense. The 2004 output is more comprehensible.

English Control Case

Figure 6 illustrates the comparison underlying our experimentation. A reference translation, produced by a skilled human translator, can be used to test a control group. If the Defense Language Proficiency Test–Standardized Translation Assessment with Reference (DLPT-STAR) test questions are given to a soldier using a reference translation, the soldier will be required to perform the linguistic task, abstracted away from the actual foreign-language component. If the same test questions are given to a soldier using MT output of the same original text, we can compare performance of the test case with the control case. Is the MT output, despite some degree of imperfec-

Reference translation	
Cairo, April 6 (AFP)–An Egypt Air official announced, on Tuesday, that Egypt Air will resume its flights to Libya as of tomorrow, Wednesday, after the UN Security Council had announced the suspension of the embargo imposed on Libya.	
Machine translation (2002)	**Machine translation (2004)**
Cairo 6-4 (AFP)–an official announced today in the Egyptian lines company for flying Tuesday is a company "insistent for flying" may resumed a consideration of a day Wednesday tomorrow her trips to Libya of Security Council decision trace international the imposed ban comment.	Cairo, 4-6 (AFP)–said an official at the Egyptian Company for aviation company today that Egypt Air may resume as of tomorrow, Wednesday flights to Libya following the decision of the Security Council to suspend the embargo imposed on Libya.
Matches	**Matches**
20 1-grams 7 2-grams 3 3-grams 2 4-grams	26 1-grams 19 2-grams 12 3-grams 7 4-grams

FIGURE 5. The correctness of a translation can be evaluated by matching translated word concepts, or grams, to the original. A perfect translation would be a single *n*-gram of the length of the passage.

Machine Translation Evaluation

New testing methods rate MT capabilities

Current methods of MT do not rely on comparing word-for-word matches as done with BLEU. As mentioned in the text, a high rating number from BLEU does not ensure a satisfactory translation. A *good* translation is one that provides the reader with material that can be understood and used. The following describes the method we developed for determining whether readers could understand a translation.

DLI and the ILR Scale. We were able to make this observation more precise by conducting an experiment. In 2003, Ray Clifford, the Chancellor of the Defense Language Institute Foreign Language Center (DLIFLC), had a fundamental insight: tests that are designed to evaluate the proficiency of language learners could be used to measure MT quality. Working with the DLIFLC provided unique advantages: it is the largest foreign-language training center in the world. Its mission is to produce operationally proficient military linguists. The DLIFLC has over 3000 students in residence in Monterrey, California, and over a thousand faculty members in twenty-six languages. But perhaps more importantly for our work, it has nearly fifty years of experience in test development. Like all U. S. government agencies teaching and/or testing foreign languages, DLIFLC adheres to the Interagency Language Roundtable (ILR) standards for evaluating language proficiency. The ILR scale ranges from level 0 to level 5 and covers the four skills—speaking, listening, reading, and writing. The descriptor for level 0 refers to no usable skill, while the level 5 descriptor for each skill indicates competence indistinguishable from that of a highly educated native. Informal descriptions for ILR levels 1 to 3 are shown in Figure A.

DLPT. The DLIFLC developed the Defense Language Proficiency Test (DLPT) to measure the reading and listening competence of military language specialists according to the ILR scale. The DLPT is the standard instrument for testing the language proficiency of military personnel trained in foreign languages by DoD and several other U. S. government agencies. Developed over the course of several decades, the DLPT is focused on the proficiency standards recognized and used throughout the government.

DLPT-STAR. In order to take advantage of the testing construct and design of the DLPT, we could have tried to use existing test items with MTs of the original source documents. However, the DLPT is an expensive, high-stakes test that cannot be overexposed for other purposes. As an alternative, we elected to construct a new test, built according to the principles of the DLPT and fully meeting the standards set by the ILR scale. Creating a new test provided flexibility in adapting it to specific aspects of MT technology. The modified test is called the Defense Language Proficiency Test–Standardized Translation Assessment with Reference (DLPT-STAR). The hypothesis behind the function of the DLPT-STAR is very simple: better performance on the DLPT-STAR means a more usable MT product.

Level 1 texts are simple announcements, notes, and advertisements consisting of a few sentences. A level 1 speaker can participate in simple conversations and ask and answer questions on everyday survival topics.
Level 2 texts are concrete, factual descriptions; instructions; narrations of current, past, and future events. A level 2 speaker can produce paragraph-length language of this type.
Level 3 texts are reports, editorials, and essays including hypothesis, supported opinion, argumentation, and abstract linguistic formulations. A level 3 speaker can produce this type of complex, extended discourse.

FIGURE A. The levels of expertise in language translation cover the broad spectrum of basic knowledge (word-to-word translation), comprehension of concepts, and expression.

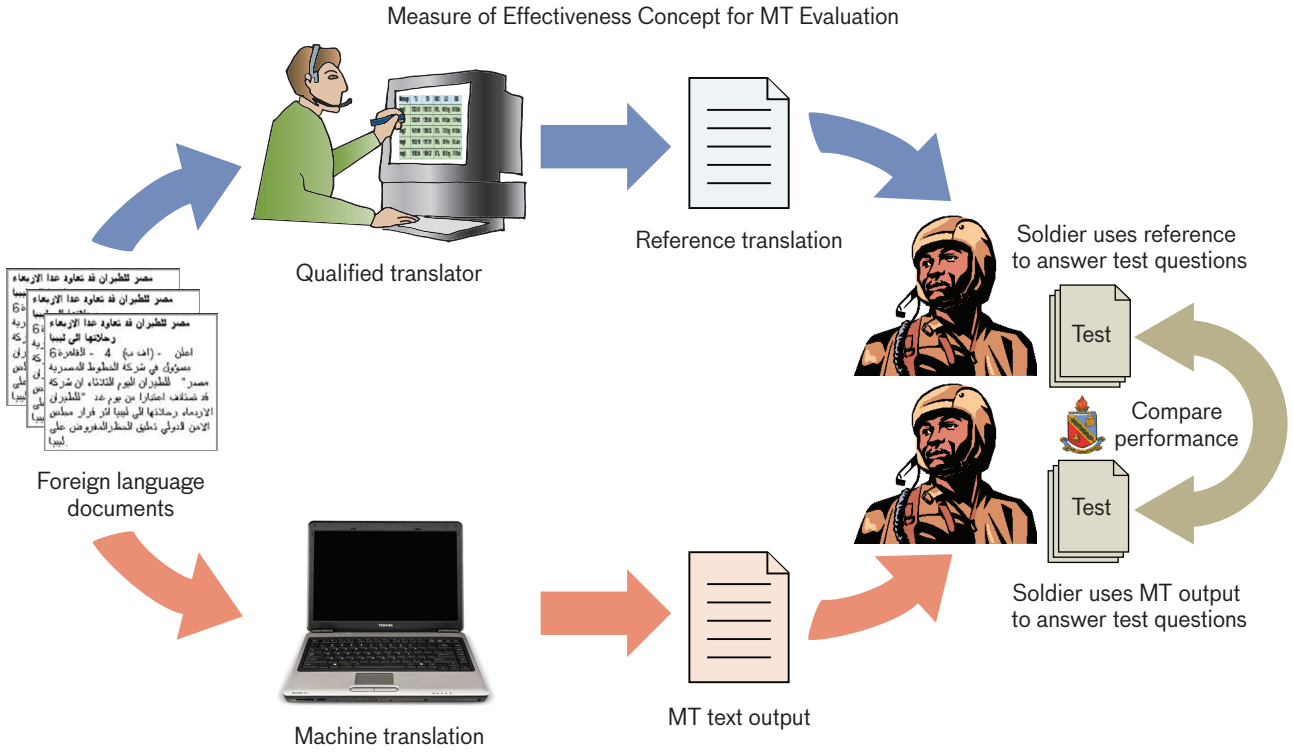


FIGURE 6. Another method of evaluating machine translation (MT) is to compare it directly with a qualified translator. As opposed to the *n*-gram matches shown in Figure 5, here the evaluation is a comprehension test by the users of such MT equipment, the soldiers who will use it in the field.

tion, adequate to enable the soldier to answer questions about the most important information in the original text? Better MT means better test scores. If the disparity between the control case and the test case is minimal, the MT output is usable. If the disparity is great, the MT output is likely to be much less useful.

There have been other MT evaluations using comprehension tests [4] and task-based measures [5]. However, nobody had ever done it using the ILR standards before, and this was a game-changer for MT evaluation. The DLPT-STAR test items measure the most important information in the original text. Specifically, they measure the information that would be important to a consumer of MT output. In addition, our use of the standardized ILR scale and descriptors as a test construct meant that the results of our experimentation were readily interpretable by users familiar with the scale who could apply that knowledge to their own specific questions about the effectiveness of an MT system.

Sample DLPT-STAR Test Item

A sample test item is shown in Figure 7. The first part is the original foreign-language document, in this case, an Arabic document. On the right, there are questions to answer, at different ILR levels. For example, a level 1 question asks for the site of this news broadcast. Someone who has learned Arabic to level 1 should be able to answer that question. At level 2, the reader has to understand a whole paragraph. And at level 3, the reader has to connect thoughts across paragraphs and make appropriate inferences, in this case, to understand emphasis.

Comprehension Results

Our first experiment, conducted in 2004, compared comprehension rates for text-to-text translations in two conditions: a reference translation and an MT output. Figure 8 shows that people were able to correctly answer the level 1 questions 95% of the time when they read the gold-standard reference translations, whereas for the MT output, they were correct 75% of the time. There are similar results at level 2. At level 3, both MT and gold-standard reference comprehension drop. Level 3 materials correspond to carefully constructed newspaper editorials and essays supporting an opinion or arguing a point of view. It

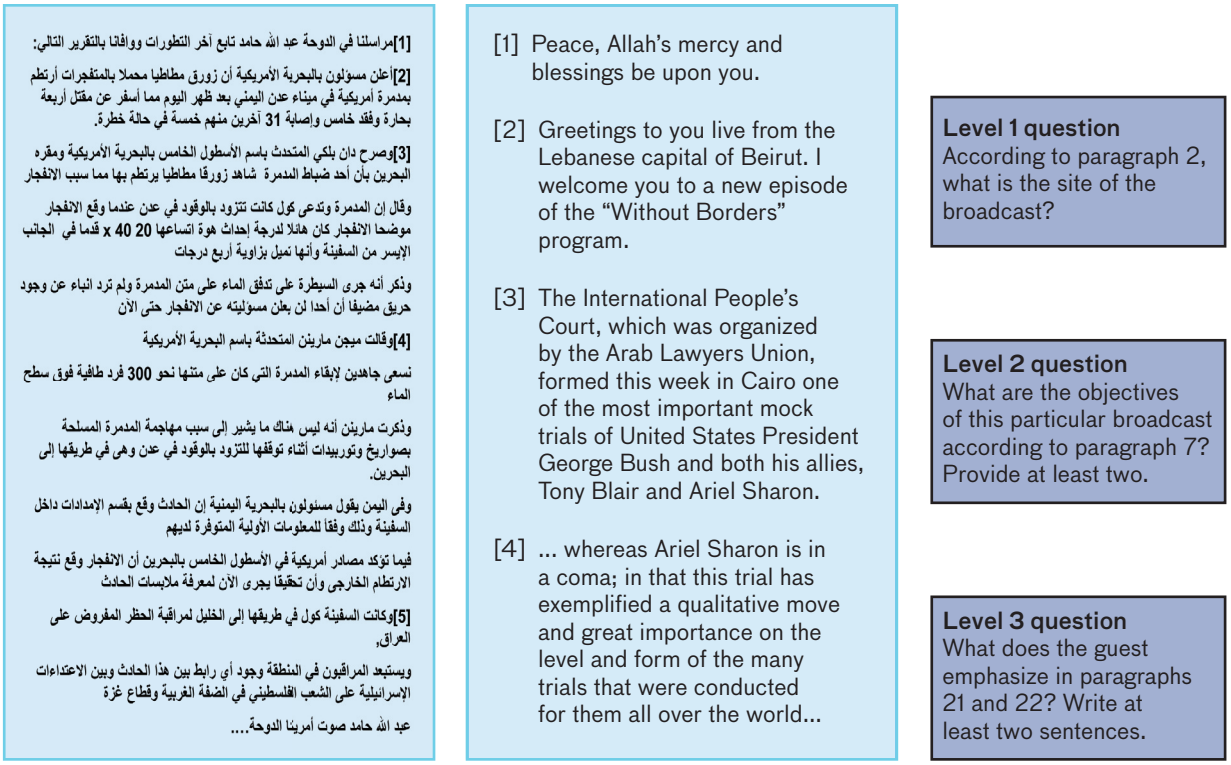


FIGURE 7. Sample test questions are provided for each Interagency Language Roundtable (ILR) standards level. Again, the quality of the MT is evaluated by the ability of the users to respond correctly to the questions.

is understandable that performance may be poorer at this level, even when a reference translation is used because examinees may not frequently read this kind of prose. The 51% performance for MT falls below the passing threshold of 70% used in many standardized criterion-referenced tests, which is what we use for the DLPT-STAR.

GALE Phase 1 Results. In a series of experiments conducted in 2005 and 2006, we added a new type of test condition for MT of audio speech files, shown in Figure 9. For this condition, in addition to performing translation, the computer also has to recognize the Arabic words. That process introduces additional errors since every fifth or sixth word is likely to be recognized incorrectly.

DLPT-STAR versus BLEU

In 2007, we built two additional tests, one for Arabic and one for Mandarin Chinese, which were administered as part of the NIST 2008 MT Evaluation. Each of the tests had ten MT conditions in addition to a reference condition. The results are shown in Figure 10. The experiment included ten Arabic-to-English MT systems and ten Chinese-to-English systems. The MT performance of the worst of ten, fifth of ten, and best of ten is contrasted with performance given the reference translations. In general, more difficult questions are harder to answer correctly, although this trend is not exhibited perfectly for every system. Results for Arabic-to-English MT systems, shown on the left of Figure 10, are far better than for the Chinese-to-English MT systems, shown on the right. In the Arabic system, the best MT system has performance that approaches the performance on the reference translations. Our comprehension results are consistent with the widespread observation that Chinese MT lags significantly behind English MT for automatic word-matching evaluations as well.

Orthogonality

Although the same general trends can be observed for both human comprehension tests and automatic word-matching scores, each technique measures translation quality in a somewhat orthogonal way. Figure 11 shows a plot for items in the GALE Phase 1 DLPT-STAR test against the Phase 1 Human Translation Error Rate (HTER) scores. Each point is the average comprehension score for one

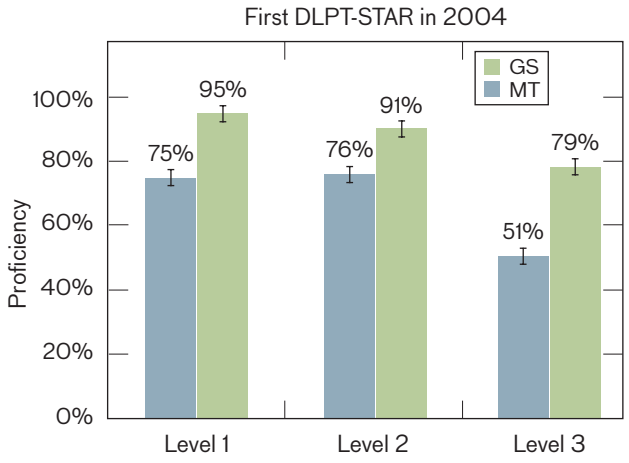


FIGURE 8. Initial results of a comparison between a gold-standard (GS) reference translation and MT show the expected reduction in proficiency when MT is used. The reduction in proficiency through the levels is also apparent.

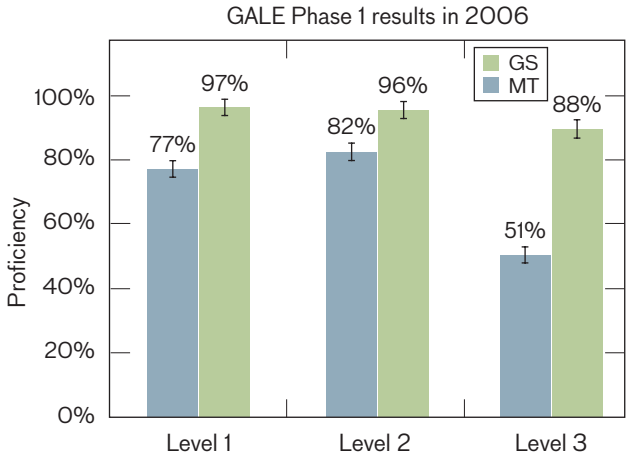


FIGURE 9. In this study, MT had to analyze audio files, meaning that the MT had to recognize the Arabic words before translating them. Thus, this test is harder than the test shown in Figure 8, but the similarity of the results suggests the improvements gained by using the newer technology.

test question, plotted against the translation error rate for the text associated with that question. Test subjects lose about 12% in comprehension for every 10% of translation error [6]. The R^2 value for this linear regression is 33%.

Fragile Translation Example

The test item in Figure 12 makes the implications of the large degree of scatter in the correlation plots more concrete. The test question has been constructed very

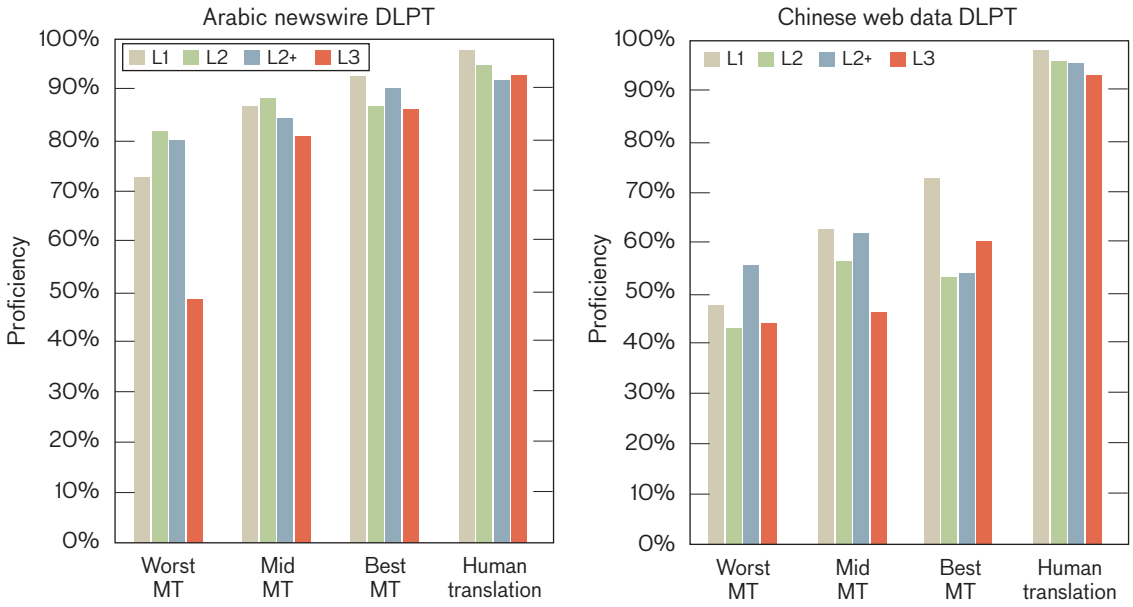


FIGURE 10. Two additional tests, developed in 2007, compared ten separate MT test conditions to reference translations. Performance for the best MT output of the ten systems, the worst of ten, and the system ranked fifth of ten are contrasted with the human translation. More difficult questions do follow the trend of achieving lower proficiencies for most of the tests, as expected, but there are some variations in the results. By comparing the results for the Arabic work on the left to the Chinese work on the right, it is clear that there is much more work to be done on the Chinese MT side.

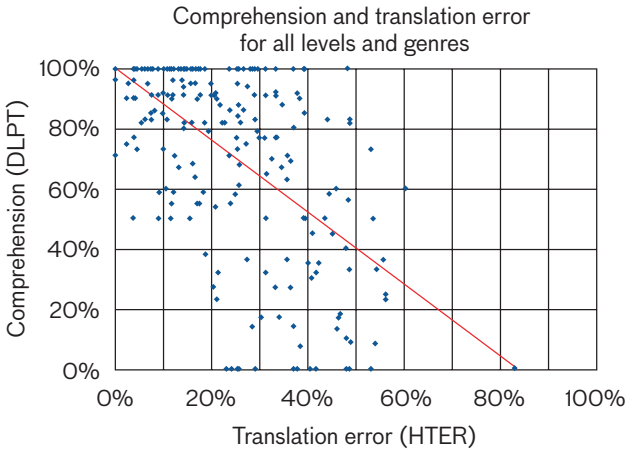


FIGURE 11. The graph shows the comparison between translation error and comprehension. Clearly, an error in translation equates to a reduction in comprehension, as determined by a series of questions at each of the three comprehension levels. Here, DLPT is the Defense Language Proficiency STAR Test and HTER is the human translation error rate.

carefully: "What is said about a warning before this incident?" The reader should not be able to guess the answer from the way the question is worded, and it is clear from the reference translation that there was no warning before the attack. Everyone who read the passage in the reference condition answered correctly; everyone who read the MT output got it wrong. The negative portion of the text that said "not receive any telephone threats" was garbled, leading the reader to infer that there was a warning. It is clear that not every word is of equal value in the translation.

Generalized Measures of Effectiveness

We began to view the DLPT-STAR results as a special case of broader questions: how job performance is currently being measured and how those tests could be adapted as measures of effectiveness for technology. In other words, how do particular soldiers prepare for deployment? For example, if military linguists need to process signals intelligence, or another intelligence specialty, they first need to pass a soldier qualification test. Likewise, if they are using foreign-language skills, they would have passed a DLPT, an Oral Proficiency Interview (OPI), and any service-specific measures of foreign-language skills. It became clear that one could adapt a wide variety of standardized military tests involving foreign-language expertise to fit the concept shown earlier in Figure 6.

Army Intelligence Study

In 2007, we conducted a series of experiments for the Department of the Army, Army Intelligence, designed to measure the usefulness of MT for specific army intelligence jobs. Working with the U. S. Army Intelligence Center at Fort Huachuca, Arizona, we identified soldier qualification tests for human intelligence and signals intelligence tasks. As part of one of the soldier qualification tests, there is a screening worksheet that a soldier has to complete, for example, to find out if a detained person should be selected to answer further questions. On the basis of an interview, the soldier fills in the name, address, and languages spoken of a local national. For the reference condition, the examinee reads an error-free translation. For MT, there is some degree of garbling. The translation system in the experiment was not designed for this domain, so the experiment was to see if such a device could recognize names, dates, and other information needed in the forms.

DLPT-STAR versus BLEU

Sample test item

Q: What is said about a warning before this incident?
A: There was no warning (before the attack on the USS *Cole*).

Machine translation

... He added the military spokesman said the ship to receive a telephone threats or and Eid may be to link to the incident.

Reference translation

... The military spokesman added that the ship did not receive any telephone threats or any threat that might be linked to the incident,...

FIGURE 12. Unbiased questions such as this should not direct the reader to interpret a positive or negative result. The corresponding machine and reference translations show that a garbling of the translation by the MT gave opposite results to the reference translation and the true facts.

Screening Exercise Results

The test was administered using three conditions: (1) the English reference translation; (2) speech-to-speech MT output; and (3) text-to-text MT output, the last of which was based on error-free transcripts of the audio. Results are shown in Figure 13. The performance is reasonably good in the reference condition and quite poor in the speech-to-text condition: the soldiers could fill the forms, by using the MT output, with about 30% accuracy. The test administrators typically use a 70% passing threshold; basically, the MT failed. But notice the improvement in the third condition, for which we provided the MT system error-free Arabic reference transcripts of the conversations in order to eliminate the additional errors incurred by recognizing spoken language. Here we see that, in some cases, the MT would have been good enough to be useful. That result tells us to attack the audio problem strongly since that is the biggest gap. These types of insights can help guide technology development toward specific aims.

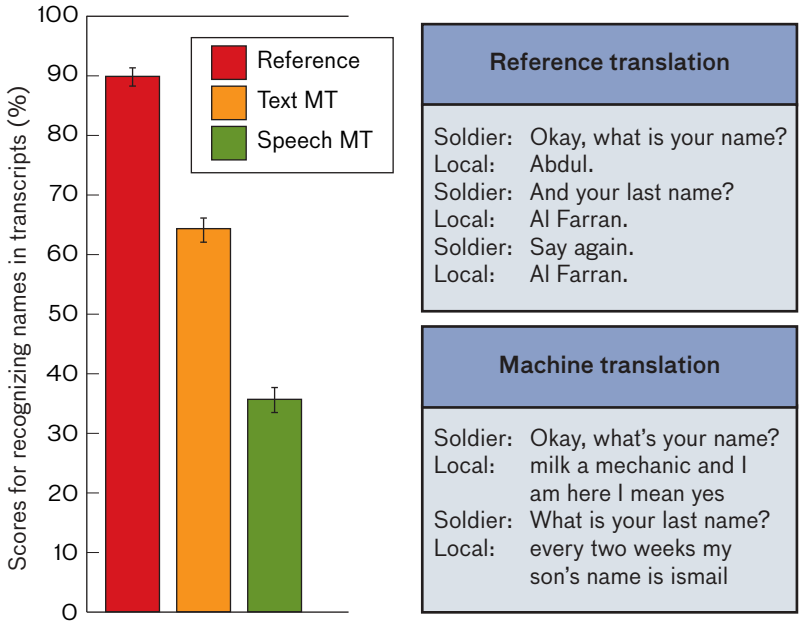


FIGURE 13. The results of tests on the recognition of names in transcripts show the inability of MT to translate properly when the information is out of its contextual range. The database for individuals' names, addresses, and other personal information is typically sparse.

Future Directions

We view MT technology as part of a larger information workflow. The entry point is the wide variety of foreign-language data for which there are processing requirements. This data are consumed by several stages of multilingual language processing, identifying the specific language, annotating it with metadata, and routing it appropriately in a larger information processing architecture. In our future work, we would like to investigate what would happen to the overall process if the translation stage were essentially solved. How would that success affect workflow downstream? An essential aspect of our future work is to identify testing procedures that are used for training and evaluating personnel and to adapt them for measuring the contributions of various types of multilingual information processing applications, as we have done for MT systems.

Transition DLPT-STAR

The DLPT-STAR type evaluation methodology is by now very stable. We are transitioning the capability back to the Defense Language Institute, in order to provide the capability of evaluating MT quality in addition to testing foreign-language learners. DLPT-STAR is designed to evaluate materials that are administered in written form, and the protocol has been very useful for measuring the quality of noninteractive MT capabilities. We have observed an increase in the level of comprehension each year since administering our first test in 2004. For a variety of reasons, we used different materials for each test, so the comparison across years is not precise. The 2008 test materials have been sequestered for reuse in future years and will enhance our ability to make a longitudinal comparison. An overall comment is that the text-to-text MT systems appear to have the most natural fit at level 2 on the ILR criteria.

Speech-to-Speech MT Evaluation

Interactive MT requires a different testing procedure. The closest analog in the government foreign-language testing field would be the OPI, which has been used for half a century to test speaking according to the ILR scale in many agencies. One future task is to adapt portions of that testing methodology for interactive spoken MT evaluation. In our preliminary work in this area, we believe that the

most natural fit for speech-to-speech MT may be level 1 on the ILR scale. As we have done with text-to-text MT, substantial and rigorous work will be needed to measure the capabilities of MT against the ILR criteria.

To date, the work on evaluation of MT systems has been done in Modern Standard Arabic (MSA) and, to a lesser extent, Mandarin Chinese. Areas to be explored include spoken Arabic in venues where MSA and a dialect are combined. This is a common phenomenon in popular gatherings and media, and it could be of considerable interest in the intelligence field. Another area requiring more attention is languages like Pashto that exist in several varieties, separated geographically and often influenced by still other languages. In addition, the distinctions between written and spoken language that are carefully learned by foreign-language professionals now appear to break down in new modes of expression such as blogs. Areas of such volatility could impact translation systems and our need to evaluate their output.

Acknowledgments

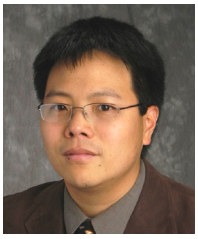
Many people have contributed to this work in ways too numerous to mention; without at least the following people, in chronological order, it would not have happened: David Savignac, Ray Clifford, Doug Reynolds, Cliff Weinstein, Marc Zissman, Ted Gibson, Neil Granoien, Dan Scott, Joseph Olive, Hussny Ibrahim, Mike Emonts, Dan Ding, Tim Anderson, Jurgen Sottung, Ed Cerutti, Irene Zehmisch, and Rodney Githens. Osaila Al-Araby, Neween Al-Wahab, Sabine Atwell, and Fauzia Farooqui also deserve special thanks. ■

REFERENCES

1. J.-P. Séris, "Language and Machine in the Philosophy of Descartes," *Essays on the Philosophy and Science of René Descartes*, Stephen Voss, editor, Oxford, England: Oxford University Press, 1993.
2. http://www.wired.com/wired/archive/8.05/timeline_pr.html.
3. K.A. Papineni, S. Roukos, R.T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," *Proc. 40th Annual Conf. Assoc. for Comp. Ling.* (ACL 02), 2002, pp. 311–318.
4. J.S. White and T.A. O'Connell, "Evaluation in the ARPA Machine Translation Program: 1993 Methodology," *Proc. Wkshp. on Human Lang. Tech.*, 1994, pp. 135–140.
5. C.R. Voss and C.R. Tate, "Task-based Evaluation of Machine Translation (MT) Engines Measuring How Well People Extract Who, When, Where-Type Elements in MT Output," *Proc. Eur. Assoc. Mach. Trans. (EAMT)*, 2006.
6. 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 Int. Conf. Intell. Anal.*, 2005.

ABOUT THE AUTHORS

Douglas Jones is a technical staff member in the Information Systems Technology Group. He joined Lincoln Laboratory in 2001 and worked to establish a new research program on machine translation with an emphasis on adapting standardized tests for technology evaluation. His civil-service job at the Department of Defense Natural Language Research Branch set the stage for his current focus. He earned bachelor's and master's degrees in linguistics from Stanford University and a doctoral degree in linguistics from MIT, supervised by Noam Chomsky.



Wade Shen is a member of the Information Systems Technology Group. His current areas of research involve machine translation and machine translation evaluation; speech, speaker, and language recognition for small-scale and embedded applications, named-entity extraction, and prosodic modeling. He received his bachelor's degree in electrical engineering and computer science from the University of California, Berkeley, and his master's degree in computer science from the University of Maryland, College Park. Prior to joining Lincoln Laboratory, Shen helped found and served as Chief Technology Officer for Vocentric Corporation, a company specializing in speech technologies for small devices.

Martha Herzog retired from her position as Vice Chancellor of Evaluation and Standards at the Defense Language Institute Foreign Language Center in Monterey, Calif., after 31 years of service, in June 2005. Since then, she has worked as a consultant, with development of methods to apply the ILR language-proficiency scale to evaluation of MT output being one of her primary activities. She received her bachelor's and doctoral degrees from the University of Texas at Austin. She now lives in Pacific Grove, Calif.