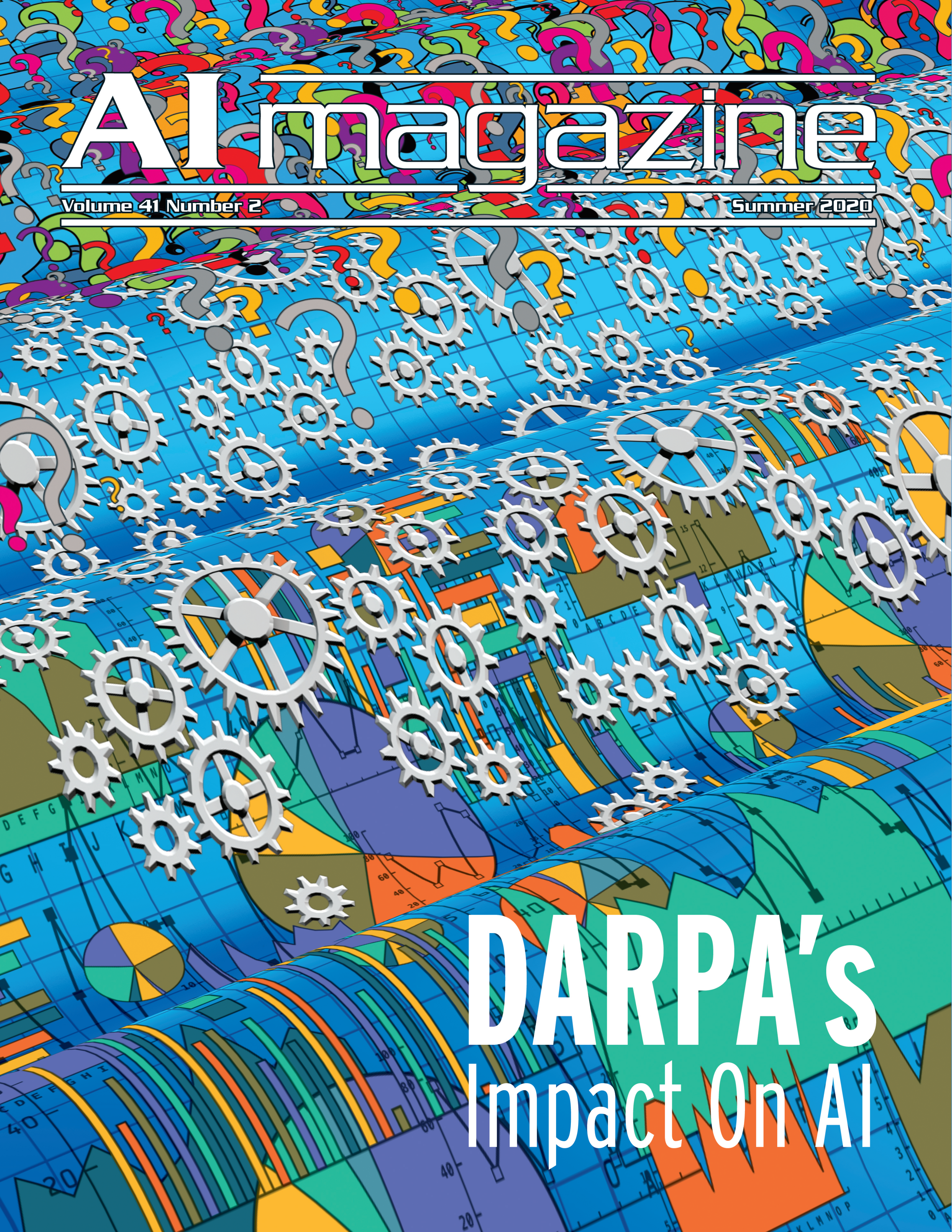
AI magazine

Volume 41 Number 2

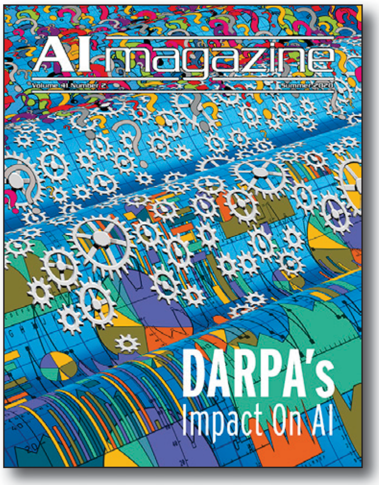
Summer 2020



DARPA's Impact On AI

AI magazine

VOLUME 41, NUMBER 2 SUMMER 2020 ISSN 2371-9621 (online) ISSN 0738-4602 (print)



***Cover:** DARPA's Three Waves of AI Research by James Gary, Brooklyn, New York*

DARPA'S IMPACT ON AI

Human Language Technology

Mark Liberman, Charles Wayne

■ *Human language technology encompasses a wide array of speech and text processing capabilities. The Defense Advanced Research Projects Agency's pioneering research on automatic transcription, translation, and content analysis were major artificial intelligence success stories that changed science fiction into social fact. During a 40-year period, 10 seminal DARPA programs produced breakthrough capabilities that were further improved and widely deployed in popular consumer products, as well as in many commercial, industrial, and governmental applications. The Defense Advanced Research Projects Agency produced the core enabling technologies by setting crisp, aggressive, and quantitative technical objectives; by providing strong multiyear funding; and by using the Defense Advanced Research Projects Agency's Common Task Method, which was powerful, efficient, and easy to administer. To achieve these breakthroughs, multidisciplinary academic and industrial research teams working in parallel took advantage of increasingly large and diverse sets of linguistic data and rapidly increasing computational power to develop and use increasingly sophisticated forms of machine learning. This article describes the progression of technical advances underlying key successes and the seminal programs that produced them.*

For more than a century, science fiction authors and screenwriters have imagined machines able to converse naturally with humans. In 1960, J.C.R. Licklider predicted powerful man–machine symbiosis, perhaps including conversational interaction by voice. Today, successful human language technology (HLT) underlies a growing array of increasingly popular consumer products and services, such as Apple's Siri, Amazon's Alexa, Microsoft's Cortana, Google's Google Assistant, and other voice-controlled digital helpers; Google's Google Translate; Nuance's Dragon Naturally Speaking; complex commercial systems, such as IBM's Watson; and a wide variety of defense, intelligence, and industrial applications. All of those capabilities resulted from robust efforts with academia and industry funded by the Defense Advanced Research Projects Agency (DARPA), followed by subsequent improvements and product development by industry.

DARPA's role in producing the core enabling technologies is an interesting, multipart story. To keep it manageable, this article concentrates on just three major thrusts — automatic transcription, translation, and content analysis — and 10 seminal programs from 1971 through 2011, highlighted in figure 1.

DARPA's Common Task Method — a virtuous cycle involving shared objectives, data, and evaluations — powered the successes. The cycle begins with ambitious technical challenges and quantitative performance targets established by DARPA and continues with data acquisition and

Copyright © 2020, Association for the Advancement of Artificial Intelligence. All rights reserved. ISSN 0738-4602

HLT: First Steps to Success

The Advanced Research Projects Agency (ARPA, later renamed DARPA) supported AI research from the beginning. In 1963, under the leadership of J.C.R. Licklider, ARPA provided funding for AI research at Carnegie Mellon University, the Massachusetts Institute of Technology, and Stanford. I was a graduate student at Stanford in 1963 and was the beneficiary of that funding of what turned out to be the beginning of ARPA/DARPA-funded research in spoken language technologies.

In 1971, at DARPA's request, a committee chaired by Allen Newell produced a report that recommended a five-year research effort toward a demonstration of a large vocabulary connected speech understanding system. For the first time in ARPA history, there were specific performance goals: the resulting system should accept connected speech from many speakers, use at least a 1,000-word vocabulary within a task-specific environment, and perform with less than 10 percent semantic error. It was to run in real-time on a 300-million-instructions-per-second computer (projected power of multiprocessor-computer that may be available by 1980).

The report was prescient in many ways. Tools and techniques that were developed at that time continue to be used. Knowledge representation using hidden Markov models (HMMs) and beam search turned out to be enduring techniques, still used after four decades. But, we missed the relevance of many other topics: importance of large data sets; automated learning and discovery of knowledge sources; and development of machine learning (ML) techniques, such as statistical ML.

In retrospect, it was to be the first Grand Challenge task from DARPA. It was ambitious and successful. It would take the community another 40 years to demonstrate systems that could recognize unrehearsed spontaneous speech from an open population, a task many of us believed was impossible in our lifetime. To its credit, DARPA continued to support HLT research throughout the period.

– Raj Reddy

annotation, parallel research efforts, and objective evaluations. Workshops are held to discuss evaluation results, data, technical approaches, and future directions; program managers make adjustments; and the cycle repeats until the program attains its goals or exhausts its funding.

Objective performance evaluations were essential. From the mid-1980s onward, the National Institute of Technology administered most of DARPA's official HLT performance evaluations. Working closely with DARPA and the research community, NIST defined evaluations to benchmark progress, selected test sets, created scoring software, administered evaluations, analyzed results, and organized workshops.

Vast quantities of data were also critical. From the early 1990s onward, the Linguistic Data Consortium (LDC; created via a DARPA grant) acquired, annotated, and distributed most of the speech and text data used in DARPA's HLT research and evaluations. The LDC's catalog currently contains terabytes of data spanning approximately 900 datasets. The 277 datasets associated with the Global Autonomous Language Exploitation (GALE) program span Arabic, Chinese, and English; include recordings and transcripts of broadcast news, broadcast conversations, and talk shows; and encompass newswire and magazine text, newsgroups and blogs, treebanks, lexicons, and more. In addition to meeting DARPA program needs, DARPA-funded data have fueled a great deal of other important academic and industrial research.

The Common Task Method was an enduring, energizing feature of all but one of the seminal programs described below. When DARPA opened HLT evaluations and workshops to interested research groups around the world, major contributions often came from researchers not funded by DARPA. This method has spread widely within artificial intelligence (AI)-related fields, leading to literally hundreds of open technical challenges. But, in many areas (for example, clinical, educational, legal applications) where such methods could be helpful, they remain rare or absent.

Technical Advances

This section describes key technical advances within each of the three thrusts. The Automatic Transcription thrust started first and provided valuable lessons for the other two.

Automatic Transcription

Automatic transcription — otherwise known as automatic speech recognition (SR) or speech-to-text (STT) — converts speech to its text equivalent. In the simplest form of this technology, the input is a stream of digital audio and the output is a stream of digital text. Complexities abound as systems must deal with different speakers, languages, styles, and contexts. Researchers explored many different approaches, and four lessons emerged: Learning is better than programming; Global optimization of gradient local decisions is crucial; Top-down and bottom-up knowledge must be combined; and Metrics on shared benchmarks matter.

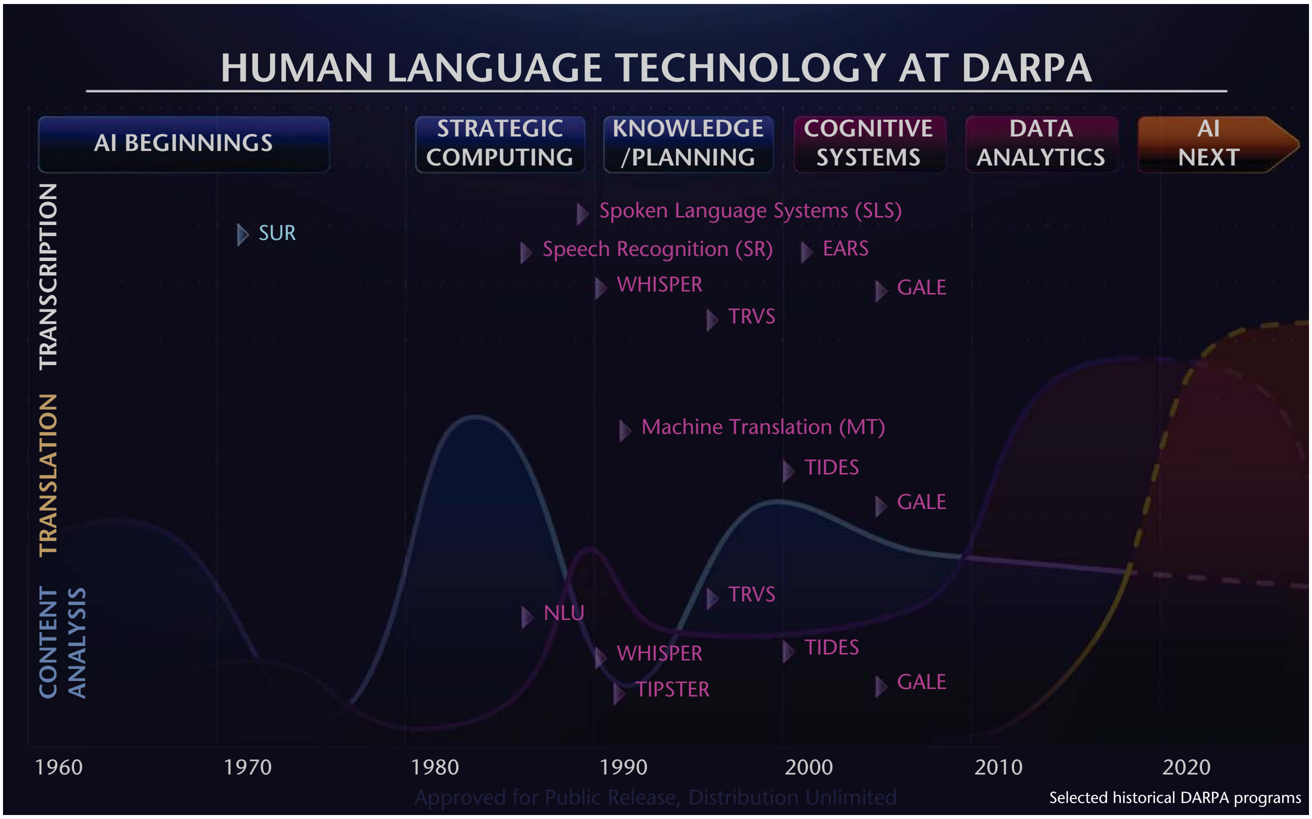


Figure 1. HLT at DARPA.

Ten seminal DARPA HLT programs from 1971 through 2011 that are highlighted in this article. *Figure courtesy of DARPA*.

An effective STT system must cope with complexity on many interacting levels, from words and word sequences to speaker characteristics, audio environments, and recording systems. Rather than trying to address this complexity via hand-crafted rules, successful systems use ML to distill large bodies of training data into millions of parameters representing actionable knowledge about these levels and their relationships.

Like many other AI problems, automatic transcription requires global optimization. An STT system considers thousands of possible words in recognizing each single word, and correspondingly makes an astronomically large number of interconnected decisions in recognizing connected speech. A successful system must combine those decisions to produce a globally optimal result, meaning that local decisions should be gradient rather than categorical — defined by probabilities or other soft scores rather than specific and final choices. Correspondingly, the methods for combining local decisions should also have gradient outputs, up to the point where a system must commit itself to a determinate final answer.

Successful systems need top-down knowledge about the text streams that are their putative outputs as well as bottom-up knowledge about the audio streams that are actual inputs. They also need to find practical and effective ways to integrate those knowledge sources into the recognition process and to adapt them to new circumstances. All of these forms of knowledge should be gradient and should be learned from training data.

The substantial overall improvement in STT performance over several decades was the cumulative result of thousands of experiments, many of which were unsuccessful. Small successes built upon each other through the Common Task Method pioneered by DARPA.

Versions of these same four lessons have played a central role in the other areas of HLT described below, and across AI research in general. But it was in DARPA's automatic transcription programs that these lessons were first learned, before spreading to other areas.

Statistical Models

The first systems to embody the four lessons given above were inspired by Claude Shannon's noisy channel model, which represents information transmission as a stream of symbols encoded by a source, transmitted over an imperfect channel, and then decoded by a receiver. In the application to automatic transcription, the symbol stream at both input and output is text; the encoder is a human speaker who turns the text into speech, the transmission adds noise, and the decoder attempts to recover the original input.

The first glimmer of a successful solution emerged in the early 1970s, as part of DARPA's first speech program. The idea was to consider that the human encoder (that is, the speaker) first encodes a string of words into dictionary pronunciations described by phonetic symbols, termed *phones*. In the second step, the phones are encoded as sounds represented as a probability distribution over sequences of spectral vectors, allowing the system to assign a probability about the time-linked correspondence between a word string and an audio clip. The independence assumptions built into the statistical model made it feasible, in principle, to perform a parallel search for the initial string of words and find a globally optimal solution.

Systems of this type typically use HMMs, in which the underlying word sequence is treated as a Markov chain and we can observe only a probabilistic function of that sequence, namely the sounds. In general, an HMM is represented by two tables: a table of transition probabilities among the hidden states, and another table specifying each state's distribution over observables.

In the case of SR, the hidden states are words, and the transition table is known as a language model. The words are broken down into phonetic segments analogous to those in a dictionary's pronunciation fields (phones) and these are related to snippets of an audio recording via an output table known as an acoustic model. If we have a sufficiently large volume of transcribed and phonetically segmented training material, we can estimate the values in these tables by simple observation (with the usual caveats about statistical estimation).

HMM parameters can be estimated without hand-segmented training data, and the forward–backward algorithm makes it possible to train systems on hundreds or thousands of hours of material, because the only requirements are audio recordings and corresponding transcripts.

HMMs became the dominant paradigm in the mid-1980s and opened a vast algorithmic space for exploration. For example, because phonetic segments are not in general uniform, and vary in length, one can try expanding each segment as a number of variably-connected substates. Because phonetic segments are strongly affected by their context, phones might be replaced by phones in context, such as triphones. Such replacements make statistical estimation a serious challenge; instead of 40 or so phones, in principle there could be 403 triphones with five substates per phone, each of which needs to specify a multivariate distribution over sounds.

Neural Models

We expressed the four key lessons in a somewhat abstract form above, using terms such as *learning* and *gradient decisions*, in contrast to writing more specifically about statistics and probability as we might have 10 or 15 years ago. At that earlier time, SR algorithms were explicitly framed as the application of complex statistical models. More recently, we have seen a resurgence of deep learning and neural net algorithms.

Since the 1940s, researchers have explored the idea of connecting inputs to outputs through complex networks of weighted sums with interspersed thresholds and other simple nonlinearities. These networks can be viewed as radically simplified models of the neural networks in animal brains, and in principle can be programmed to compute any computable finite function. A suitably designed network can learn the weights needed to solve a given problem. As computers became exponentially more powerful, and increasing amounts of training material became available, improved algorithms were developed for configuring and training such neural systems. Producing greater accuracy than HMMs for STT, they have become an increasingly important part of ML and AI, as discussed in other articles in this issue.

Systems of this kind are perfectly adapted to implement our four lessons, and therefore they easily fit into AI research programs, in HLT as in other areas, supplementing or replacing the statistical models that led to the earlier generation of successes.

Automated Evaluations

Simple, automated evaluations played a crucial role in DARPA's HLT successes. For instance, to evaluate automatic transcription performance, DARPA adopted word error rate (WER) as its standard metric. WER matches a system-generated transcript to a human reference transcript, adds up substitutions, insertions, and deletions (that is, errors), and divides by the number of words in the reference transcript.

There are obvious problems with the WER metric — not all errors cause equally important changes in meaning, and many errors can simply be ignored. A better measure would accurately reflect the suitability of the system's output for a specific task. However, such measures are often impossible to automate. The fact that WER is easily computed and task-independent, enabled an extraordinary record of success in improving STT technologies during decades of DARPA-sponsored HLT research.

Automatic Translation

Automatic translation is the conversion of speech or text between languages. Modern automatic translation systems have been based on the same noisy channel model used to frame the STT transcription problem. Applying this metaphor to automatic translation, we assume that someone means to say something in English, but as a result of encoding and channel noise, it comes out in some other language, say Arabic or Chinese, and the receiver then uses a statistical (or neural) model to decode the sender's English intent.

In the simplest form of the decoder for translating a foreign language to English, each input word is associated with a probability distribution over a set of English words. Each possible arrangement of these English words is then scored as a combination of its intrinsic estimated probability as an English-language sequence (its language model score) and the probability of its constituent words as translations of the foreign-language input. Translation from English to another language is the same process in reverse.

That original form, investigated almost 30 years ago, has been superseded by more sophisticated models, but all versions fit the first three of our four lessons. Such systems need parallel text in the two languages for training to learn probabilistic associations between words and monolingual text in the target language to train the output language model. Improving the model requires an easy-to-compute measure for benchmarking.

The challenge in automatic translation, as opposed to automatic transcription, is that there are many different valid translations, involving many possible choices of words and word orders. The BiLingual Evaluation Understudy (BLEU) measure allows for that variation, but in a way that strikes many people as implausible. BLEU evaluates a candidate translation against a small set of alternative human translations by asking what proportion of its words and word sequences can be found, in any order, in any of the corresponding human-translation sentences. The scores are then averaged across all the sentences in the test set.

Despite its simplicity, BLEU correlates reasonably well with human judgments, allowing researchers to hill-climb in the space of possible variations on the simple noisy channel translation model, just as they did using WER in the area of automatic transcription. The result was the same: steady progress over several decades, to the point that automatic translation systems are now used effectively by many millions of people every day. The BLEU experience shows that even a crude evaluation metric can serve to foster significant technical progress in programs that also adhere to the first three lessons.

Automatic Content Analysis

The goal of automatic content analysis is to turn language into information. This can take many forms and must cope with many challenges.

For example, one simple content-analysis goal could be to detect entities (for example, people, places, organizations, etc.) mentioned in a text. But a person might be referenced by first and last name, by title and last name, by a pronoun, or by a description. A company might be referenced by a full formal name, by a shortened name, or by initials. The automatic content analysis goal might simply be to flag this information, but it might also be to combine results from multiple sources, summarize facts, or track a topic or event.

As in automatic transcription and automatic translation, approaches to automatic content analysis began solely with handcrafted rules, which were expensive to create, limited in both performance and scope of application, and brittle when applied to real-world problems. Again, solid and sustained progress began when researchers, in the context of DARPA programs, replaced most if not all of the handcrafting with ML. Researchers designed systems that combined gradient local evidence into globally optimal decisions, found ways to integrate learned knowledge at different levels and from different sources, and evolved their algorithms through evaluation on shared tasks with well-defined automatic measures.

Through a continuing series of DARPA programs, described in the next section, content-analysis research developed a rich, and continually expanding, set of tasks, datasets, measures, and algorithms.

Seminal Programs

This section describes the seminal programs within the three thrusts. The first program listed under each thrust was essentially a warmup; more ambitious, better informed programs followed it, sometimes years apart. Text, Radio, Video, Speech (TRVS), Translingual Information Detection, Extraction and Summarization (TIDES), and GALE appear under multiple thrusts.

Automatic Transcription

Over a span of 40 years, DARPA drove significant transcription technology advances, making transcription more effective for a wider variety of speech, across forms, languages, and genres. Table 1 highlights the key programs and the types of speech each attacked.

Speech Understanding Research addressed a simple type of speech: short information-seeking queries from a 1,000-word vocabulary with a highly constrained syntax. DARPA funded several research groups, challenging them to construct end-to-end systems able to transcribe and understand the queries well enough to produce a correct response. That was the outer limit of what experts thought could be possible, and no one knew whether it could be achieved.

The groups took advantage of the considerable knowledge scientists and engineers had regarding speech production, perception, and analysis to design complex systems incorporating mixtures of handcrafted rules and statistical models that integrated multiple sources of knowledge. In addition to producing a system that exceeded DARPA's goals, these efforts laid useful groundwork for future advances. But despite Speech Understanding Research's success, DARPA chose not to extend the work primarily because 1970s computational power could not support real-time applications.

After a 10-year hiatus, DARPA resumed work on automatic transcription by launching a series of

SUR	Speech Understanding Research *Speaker-dependent read speech with 1,000-word vocabulary*	1971–1976
SR	Speech Recognition *Speaker-dependent read speech with 1,000-word vocabulary and highly restricted grammar* *Speaker-independent read speech from* Wall Street Journal *sentences with 5K, 20K, and 64K vocabularies*	1986–1994
SLS	Spoken Language Systems *Speaker-independent, goal-directed spontaneous speech with unlimited vocabulary*	1989–1994
—	WHISPER *Speaker-independent conversational telephone speech*	1990–1993
TRVS	Text, Radio, Video, Speech *Broadcast news in Chinese and English*	1996–2000
EARS	Effective, Affordable, Reusable STT *Broadcast news and telephone conversations in Arabic, Chinese, and English*	2001–2004
GALE	Global Autonomous Language Exploitation *Broadcast news and talk shows in Arabic, Chinese, and English*	2005–2011

Table 1. DARPA Programs that Advanced Automatic Transcription.

increasingly ambitious research programs attacking various styles of speech in multiple languages. Taking advantage of increasing computational capabilities and data availability, these programs produced ever more powerful transcription algorithms.

The evolving transcription systems incorporated increasingly sophisticated acoustic models that captured the variability of speech sounds (within and across words) and language models that captured how words are strung together. The models were trained on large quantities of spoken and written data.

From 1986 through 2004, DARPA focused on driving down WER — a simple, objective measure of accuracy introduced previously — computed by comparing a system's output to an official transcript and counting substitutions, insertions, and deletions as errors. Figure 2 shows the speech styles attacked and the WER reductions obtained by the best speaker-independent transcription systems from 1988 through 2004. The various lines do not always decline monotonically as one would expect, because NIST chose new test sets for each evaluation, some unintentionally harder or easier than others.

DARPA-funded researchers had to participate in periodic open evaluations and workshops; outside groups were permitted to participate as well, with DARPA providing linguistic data but no funding. This mutually beneficial arrangement increased competition, introduced new ideas, and accelerated progress while simultaneously conferring the credibility and access sought by the volunteers. All participants had to share their data and describe their algorithms, thereby accelerating progress. And, because NIST provided automated software for calculating WER, researchers could perform algorithmic hill-climbing on their own test sets and conduct in-house evaluations on set-aside data. The result was strong, steady progress. Year-to-year improvements were rarely dramatic, but the cumulative effects were. The resulting technology now empowers countless applications.

In a series of steps, SR significantly improved transcription accuracy on increasingly challenging types of read speech. Early SR efforts included both rule-based and statistical approaches, but DARPA quickly abandoned an expensive handcrafted, rule-based system when the 1988 evaluation demonstrated the superiority of the less-expensive, more-accurate, and more-flexible statistical approaches.

Researchers first attacked small-vocabulary, narrow-domain read speech using a corpus of military-style sentences from a 1,000-word vocabulary with a bigram (that is, two-word) language model. After three years of research on this highly-artificial task, speaker-dependent systems (for which test speakers had laboriously trained acoustic models to their voices) achieved a WER of 1.8 percent; speaker-independent systems achieved a WER of 3.6 percent.

Researchers then moved on to speaker-independent, broad-domain read speech, specifically sentences from the *Wall Street Journal* with vocabulary sizes increasing from 5,000 words to 64,000 words. The black lines in figure 2 depict the WER reductions obtained with acoustic models trained on the voices of multiple speakers, none of whom was a test speaker. Subsequent programs dealt only with speaker-independent, unlimited vocabulary speech.

From DARPA Research to Commercial Applications

Several decades of DARPA spoken-language programs have helped industry advance AI and natural interaction in profound ways. Many new speech and language services from companies like Amazon, Apple, Google, IBM, Microsoft, and Nuance directly benefited from DARPA's pioneering speech and language research.

I participated in DARPA's spoken language research at Carnegie Mellon University before I joined Microsoft in 1993 to lead the newly formed Microsoft Speech Research and Development Group. Microsoft then licensed Carnegie Mellon University's Sphinx-II speech technology, which I had helped to create, and which was mostly funded by DARPA. Several prominent Carnegie Mellon University speech and language researchers also joined my Microsoft team, and we built on DARPA HLT foundations to significantly advance the technology. In 2016, Microsoft became the first company to reach human parity in transcribing conversational speech on the SWITCHBOARD corpus.

As a result of capabilities first developed in DARPA HLT programs, Microsoft has been able to partner with others in industry and academia to deliver speech and language products and services that help to remove language barriers and ease human–computer interaction. SR, machine translation (MT), and digital assistants have become widely available in many of Microsoft's products, including the Xbox entertainment system, productivity applications such as Office Dictation and PowerPoint automatic captioning and transcription, personal assistants like Cortana and Microsoft Translator, and Azure Cognitive Services for developers. Other companies have also benefited directly from DARPA HLT research — for example, Apple's Siri was originally developed as a spin-off from the DARPA Cognitive Assistant that Learns and Organizes program.

– Xuedong Huang

SLS developed technology to automatically transcribe goal-directed spontaneous speech; specifically, requests for information related to air travel planning that could be answered by a hypothetical Air Travel Information System, a forerunner to Apple's Siri.

Participating sites collected training and test data via Wizard-of-Oz simulations. When a naïve user posed a question, unseen assistants would quickly transcribe the speech, post the transcript on the user's screen (as if a computer had produced it), query an appropriate database, and then display the response. The simulation was quite effective, fooling even some technical managers. As spontaneous speech transcription technology improved, hidden human transcribers were no longer needed; and the automatically transcribed speech allowed data to be collected more rapidly and encouraged users to speak more briefly. The green line in figure 2 reflects the rapid reduction of WER achieved on the resulting test data sets.

WHISPER was designed to develop speaker- and topic-spotting technology for conversational telephone speech. To fuel that research, Texas Instruments created the first version of the SWITCHBOARD corpus containing thousands of telephone conversations between strangers labeled by speaker and topic.

WHISPER unexpectedly inaugurated the government's work on conversational speech transcription — all because one site decided to attack topic spotting by creating automatic transcription technology for conversational speech. Although the resulting transcripts were full of errors (WER near 100%), they proved more effective than acoustic template-based word spotting for identifying topics. Inspired by this discovery, the National Security Agency started sponsoring research on conversational speech transcription technology. DARPA returned to the challenge eight years later in the Effective, Affordable, Reusable STT (EARS) program discussed below. The red lines in figure 2 depict WER reductions on various conversational speech corpora.

TRVS included DARPA's first attempt to automatically transcribe broadcast news. The solid blue line in figure 2 depicts WER declines for English; the dashed blue line, for Chinese.

EARS was a strong, systematic attempt to create fast, accurate transcription technology for broadcast news and telephone conversations in three major languages: Arabic, Chinese, and English. Researchers developed increasingly sophisticated probabilistic acoustic and language models, estimating parameters from what would become an order-of-magnitude-more training data. For broadcast news, the solid blue line in figure 1 depicts WER declines for English sources; the dashed and dotted blue lines, for Arabic and Chinese sources. For telephone conversations, the solid red lines depict WER declines on various English corpora; the dashed and dotted red lines, for Arabic and Chinese.

In a span of just three years, EARS researchers slashed WERs in half for English — cutting them from 18.0 percent to 8.6 percent for broadcast news and from 27.8 percent to 12.4 percent for conversational telephone speech. These numbers are from

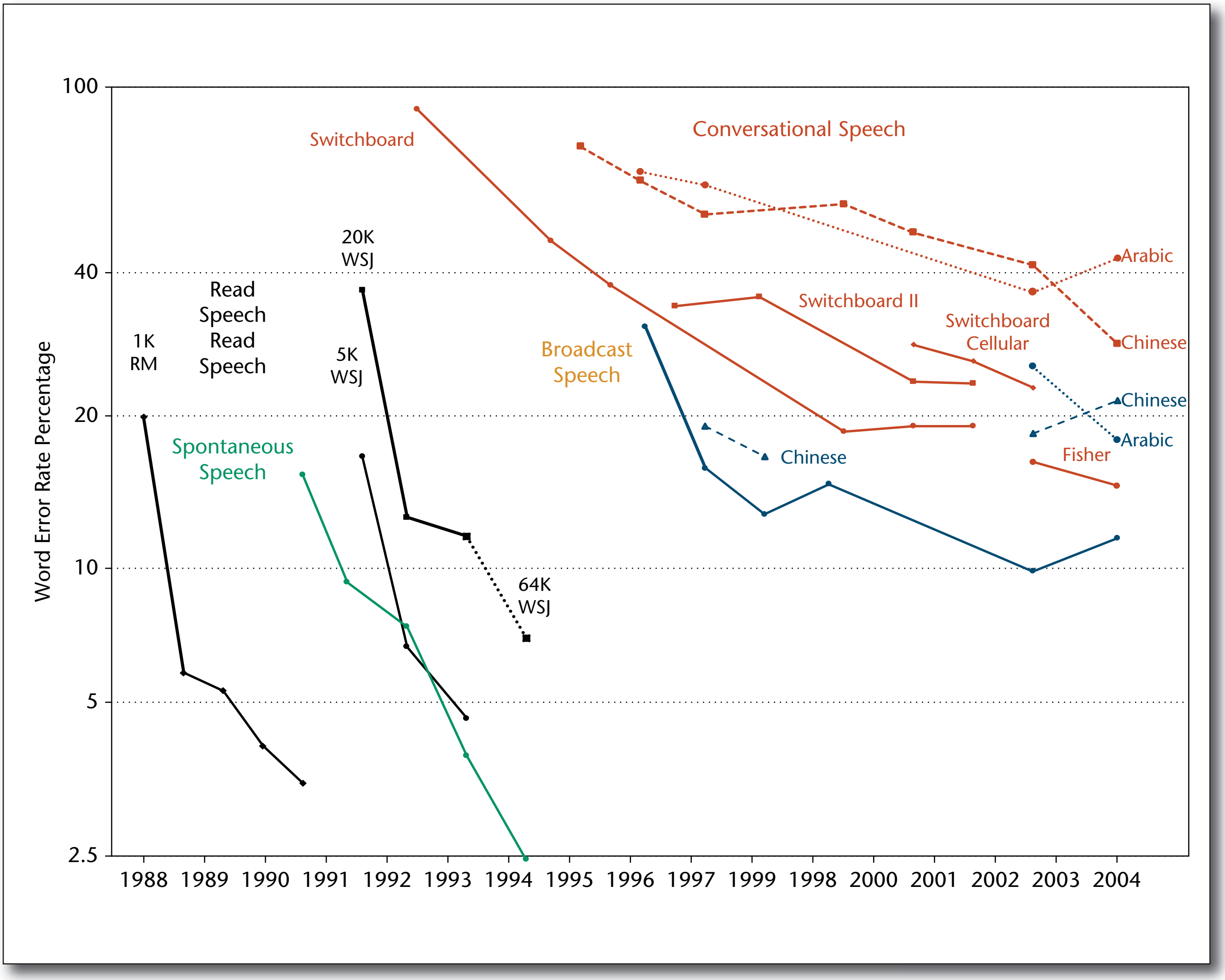


Figure 2. Advances in Speaker-Independent Transcription Accuracy.

Figure courtesy National Institutes of Standards and Technology.

Go/No-Go evaluations on stable, set-aside test sets, not the everchanging test sets used for figure 2.

To produce more rapidly readable transcripts, EARS also worked on metadata extraction techniques — to detect disfluencies, sentence boundaries, and who spoke when in multiparty speech — and measured impact via readability speed experiments.

Although EARS was meeting increasingly challenging Go/No-Go criteria for speed and accuracy and was scheduled to run for a total of five years, DARPA terminated it at the three-year mark to launch the GALE program described below. STT research on broadcast news continued within GALE, but research on telephone conversations ceased.

GALE created transcription engines to transcribe broadcast news and talk shows in Arabic, Chinese, and English. Researchers tailored those engines to optimize the output of downstream translation and distillation engines; they did not focus on, nor officially measure, WER reductions produced by the transcription engines.

Automatic Translation

During a span of 20 years, DARPA programs thoroughly revolutionized automatic translation. The first two programs highlighted in Table 2 addressed text-to-text translation; the third added STT translation.

MT was DARPA's first foray into translating naturally occurring text. DARPA funded three research groups, each pursuing a different technical approach: manually encoded transfer rules, statistical techniques informed by automatic transcription, and a combined linguistic and statistical approach. Deviating from the Common Task Method, DARPA permitted each group to focus

Source	2002	2004	Human
مصر للطيران قد تعاود غدا الاربعاء رحلاتها الى ليبيا	insistent Wednesday may recurred her trips to Libya tomorrow for flying	EgyptAir Has Tomorrow, Wednesday to Resume Its Flights to Libya	Egypt Air May Resume its Flights to Libya Tomorrow
القاهرة) 4-6اف ب - (اعلن مسؤول في شركة الخطوط المصرية للطيران اليوم الثلاثاء ان شركة "مصر للطيران "قد تستانف اعتبارا من يوم غد الاربعاء رحلاتها الى ليبيا اثر قرار مجلس الامن الدولي تعليق الحظر المفروض على ليبيا .	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 EgyptAir may resume as of tomorrow, Wednesday flights to Libya following the decision of the Security Council to suspend the embargo imposed on Libya.	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.

Figure 3. Advances in Translation Accuracy during TIDES.

MT	Machine Translation *Journalistic text from Spanish, French, and Japanese to English*	1991–1993
TIDES	Translingual Information Detection, Extraction and Summarization *Newswire from Arabic and Chinese to English*	2000–2004
GALE	Global Autonomous Language Exploitation *Newswire and newsgroups from Arabic and Chinese to English*	2005–2011

Table 2. DARPA Programs that Advanced Automatic Translation.

on a different source language. While the results were somewhat promising, it was not possible to compare the effectiveness of the different approaches. Furthermore, slow, labor-intensive, manual evaluations of translation accuracy — using techniques developed for grading human translators — proved inappropriate and greatly impeded progress.

Eight years later, TIDES launched a strong, sustained effort to create technology for automatically translating Arabic and Chinese text to passable English. To fuel the research, the LDC amassed and annotated substantial corpora of Arabic, Chinese, and English newswires, including parallel (Arabic-English and Chinese-English) data. TIDES researchers extended and substantially improved upon the statistical translation approaches that had shown promise during the MT program. DARPA adopted a crude automated method (BLEU) for evaluating translation accuracy. In the same way that WER helped transcription technology improve, BLEU helped translation technology advance — by allowing researchers to use error-based ML to rapidly improve their translation algorithms.

Because the meaning of a BLEU score (that is, the geometric average of *n*-word matches between a candidate translation and several good human reference translations) was difficult for laymen to comprehend, NIST also computed Percent-of-Human figures based on the ratio of a machine-translation BLEU score to a human-translation BLEU score.

Figure 3 is an example of how well TIDES turned Arabic text into increasingly comprehensible English text—figuratively, darkness into light. The 2004 translation is imperfect, but gives English readers some understanding of the Arabic.

GALE pushed automatic translation further forward. Its text inputs were newswires and newsgroups; its speech inputs were broadcast news and talk shows; its source languages were Arabic and Chinese. Researchers jointly optimized GALE's transcription and translation engines to maximize end-to-end (speech input to translated text output) accuracy.

Researchers continued to use BLEU internally to guide algorithmic improvements, but DARPA

NLU	Natural Language Understanding *Extraction of facts from Navy messages about equipment failures*	1986–1989
—	WHISPER *Spotting of speakers and topics in conversational telephone speech*	1990–1993
—	TIPSTER *Retrieval of desired documents from large, diverse collections. Extraction of facts about entities, relations, and events from news stories* *Summarization of news stories*	1991–1998
TRVS	Text, Radio, Video, Speech *Detection and tracking of unforeseen topics (events) from newswire and automatically transcribed broadcast news in Chinese and English*	1996–2000
TIDES	Translingual Information Detection, Extraction and Summarization *Question answering and cross language retrieval plus TDT from newswire and automatically transcribed broadcast news in Arabic, Chinese, and English* *Extraction of facts about names, entities, and relationships from Arabic, Chinese, and English newswires* *Summarization of one or more news articles*	2000–2004
GALE	Global Autonomous Language Exploitation *Improved detection and extraction capabilities for formal and informal text (newswires and news groups) plus automatically transcribed speech (broadcast news and talk shows) in Arabic, Chinese, and English* *Distillation engine to deliver (in English) precise information requested by English speakers (without redundancy) from those Arabic, Chinese, and English sources*	2005–2011

Table 3. DARPA Programs that Advanced Automatic Content Analysis.

adopted a more natural measure of translation accuracy for GALE's Go/No-Go evaluations. This involved having English-speaking editors revise system outputs as little as possible to make them contain the same information as gold-standard reference translations. The fewer distinct edits needed, the better. Figure 3 shows how rapidly translation errors declined on various types of Arabic data during GALE's five phases (P1 … P5). Progress on Chinese was similar, albeit slower.

Automatic Content Analysis

During a 25-year period, DARPA's goals and funding revolutionized automatic content analysis, producing technology for detecting desired information in speech or text, extracting information to populate knowledge bases, and summarizing information in readable forms. Table 3 highlights the seminal programs and their corresponding focuses.

Natural Language Understanding began DARPA's work on extracting facts from documents to populate databases. The research used brief Navy messages about shipboard equipment failures (that is, casualty reports, or CASREPs). To spur progress, the Naval Ocean Systems Center organized two Message Understanding Conferences: MUC-1 in 1987, and MUC-2 in 1989.

MUC-1 was an exploratory free-for-all, in which each participating group decided independently how to express the content of a set of sample messages, and there was no formal evaluation of the success of their efforts. For MUC-2, the organizers specified a data model in the form of a template with 10 slots; provided training data with correctly filled templates; and tested participants' analyses of unseen test examples using recall and precision as quantitative performance measures.

WHISPER developed techniques for spotting specified speakers and topics in conversational telephone speech. It used the SWITCHBOARD corpus wherein no one spoke to the same person or discussed the same topic more than once. When WHISPER ended, SWITCHBOARD contained 2,438 two-way conversations among 543 speakers on 70 topics. WHISPER was remarkably successful, correctly detecting specified speakers and topics at least 80 percent of the time with no more than three-percent false alarms for speakers and 10 percent for topics.

TIPSTER developed detection, extraction, and summarization capabilities for text. Unlike earlier information retrieval research that used small sets of documents from a few specialized sources, TIPSTER's detection research assembled and distributed large, diverse corpora of naturally occurring text — 750,000 documents in 1992 alone. This was shortly after the appearance of the Web in 1991, but before the first generally available web browser in 1993, and well before the founding of Google in 1998; it was a time that digital documents generally had to be converted from proprietary typographical-engine formats. When TIPSTER ended in 1998, it had provided researchers 4,550,453 documents from 38 sources.

TIPSTER addressed five types of detection — English routing/filtering, ad hoc English retrieval, high-precision retrieval, Chinese retrieval, and Spanish retrieval. To evaluate ad hoc retrieval, NIST provided a large set of documents and a set of carefully crafted queries. For each query, systems automatically scored all of the documents for relevance and submitted the results to NIST; analysts reviewed the highest ranked documents and made binary judgments about the relevance of each; and NIST then used those judgments to produce precision-recall graphs and to calculate mean average precision metrics.

To exchange information about technical approaches and evaluation results more broadly, NIST founded the annual Text Retrieval Evaluation Conference series. Now in its 28th year, the Text Retrieval Evaluation Conference has explored more than a hundred fundamental capabilities and leading-edge applications. Capabilities span detection, extraction, and summarization, and are integrated with SR, MT, and the analysis of images and videos. Research inspired by these tasks and discussions has played a key role in nearly every aspect of modern research in AI.

TIPSTER's extraction research focused on techniques for filling databases with structured information extracted from text. Systems had to analyze documents to tag textual mentions of named entities (for example, people, places, organizations, dates, times, etc.); find information about the entities, relations, and events; and enter that information in multislot templates. TIPSTER continued the MUCs, broadening the range of subject areas to include news stories about terrorist incidents (MUC-3 and MUC-4), corporate joint ventures and microelectronic production (MUC-5), labor disputes and corporate management changes (MUC-6), and rocket launches and airplane crashes (MUC-7).

TIPSTER's summarization work sought to reduce the amount of text a person would have to read to understand a document. NIST started the Summarization Analysis Conference series to discuss technical approaches and evaluation results.

TIPSTER transferred its most promising technologies to various government agencies and received a Hammer Award from Vice President Gore for significant contributions toward reinventing government.

TRVS broke new ground by launching research on topic detection and tracking (TDT). Unlike traditional retrieval applications that seek specified information, TDT tackled a new problem — detecting unforeseen (unspecified) events described in continuously arriving streams of speech and text, and tracking stories about them. The purpose was to alert analysts to the occurrence of new events and to group stories about them for further review.

TDT research, conducted on English and Chinese data, included story segmentation (automatically partitioning streams of text and audio into stories), topic detection (identifying events described in stories), and topic tracking (identifying other stories about the same event in the same or another language). To support the research and evaluation, the LDC assembled and annotated large, diverse data sets: 84,896 stories in English (from two news services plus five broadcast news sources), and 14,267 stories in Chinese (from two news services and one broadcast news source).

In TDT evaluations, NIST used miss-false alarm in lieu of precision-recall to emphasize the importance of minimizing errors and to avoid the confounding effects of target richness in different corpora. At each possible decision point, NIST calculated a normalized cost from 0.000 (perfect) to 1.000 based on the miss-and-false-alarm probability at that decision point, and hypothesized costs for misses (10) and false alarms (1), plus the a priori probability for the target condition (for example, a story being on topic). Because the number of on-topic stories varied widely and topic difficulty was a major source of variability, NIST reported topic-weighted results (wherein each topic contributes equally to the overall averages) to improve the reliability of the performance measures. Figure 5 shows a detection error tradeoff graph.

TIDES built upon the successes of TIPSTER and TRVS by developing technology to help English speakers find and interpret needed information quickly and effectively regardless of language or medium. The inputs were a variety of news sources (newswires plus automatically transcribed radio and television news broadcasts) in three important and distinctly different languages (Arabic, Chinese, and English).

TIDES detection research produced operationally useful capabilities for simple question answering, cross language retrieval, audio retrieval, topic tracking, and topic clustering; these were fielded in various demonstration systems and delivered to military customers. TIDES also began groundbreaking research on high-accuracy retrieval. Extraction research moved from ad hoc, domain-specific tasks to domain-neutral tasks; produced highly effective name recognition technologies for all three languages; and expanded the number of relation types it could extract. Summarization research sought to substantially reduce the number of words an analyst would have to read to understand the content of news articles (single articles or groups of articles), in which summaries could be fluent text or bullets.

Combining the above capabilities with automatic transcription technology, TIDES built real-time systems that enabled English-speaking operators to access and interpret information from various Arabic speech and text sources. TIDES used two operational prototypes in the 2004 Strong Angel exercise — a proxy for humanitarian operations across the civil-military boundary in Iraq, Afghanistan, and future

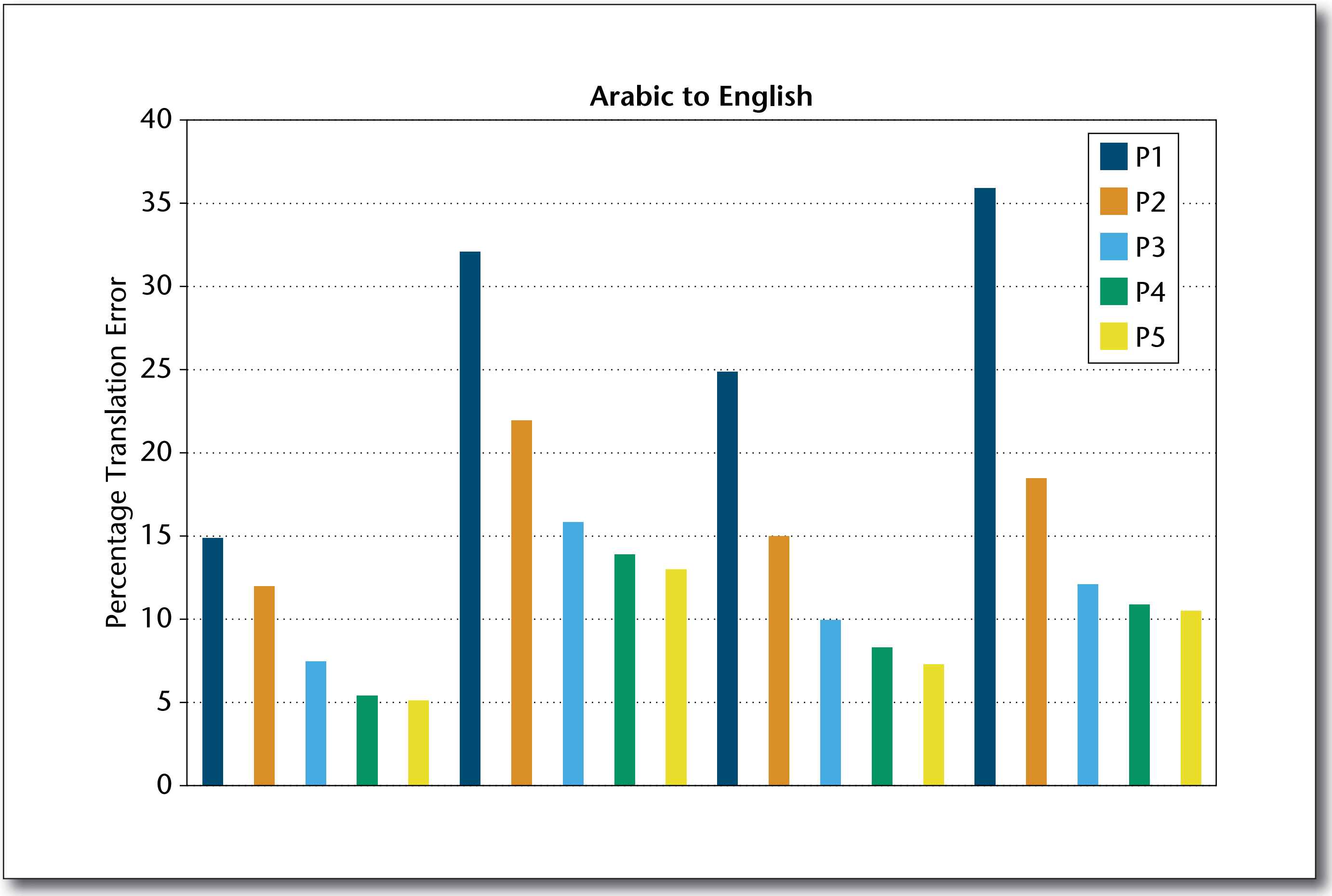


Figure 4. Advances in Translation Accuracy during GALE.

Figure courtesy of DARPA.

conflict zones — and transferred other interactive systems to the military. TIDES also conducted two surprise language experiments to determine how quickly HLT technologies could be ported to languages that might suddenly become operationally important. Cebuano and Hindi were the two trial languages, very different linguistically, and in terms of data availability. Those experiments showed that strong multisite collaboration could produce somewhat useful capabilities in less than a month.

GALE sought to give users the precise information they requested, with citations and without redundancy. Researchers pushed automatic content analysis further forward, significantly improving upon TIDES capabilities for entity and relation tagging, topic modeling, event detection, and cross-document entity linking. Researchers integrated those technologies in distillation engines that operated on the outputs of GALE's transcription and translation engines.

When users expressed their information needs via English language template queries — such as, "Describe attacks in [location] giving location (as specific as possible), date, and number of dead and injured" — systems were expected to provide comprehensive responses containing all relevant information without redundancy, note corroborating information and contradictions, and include all appropriate citations.

To evaluate distillation performance, British Aerospace (BAE) Systems compared system outputs to those produced by time-limited humans, by looking at information nuggets relevant to the template queries. Systems found many more facts than humans but also produced a great number of false alarms. Combining miss and false-alarm information, BAE found that system scores exceeded human scores 50 percent of the time.

This is a convenient place to conclude our discussion of the seminal programs that made HLT a reality and moved it far forward. The GALE program extended and integrated all three thrusts — producing systems with transcription, translation, and distillation engines connected in series — transcription to convert speech to text, translation to convert Arabic and Chinese to English, and distillation — to provide the precise information requested by English-speaking analysts and decision makers, regardless of

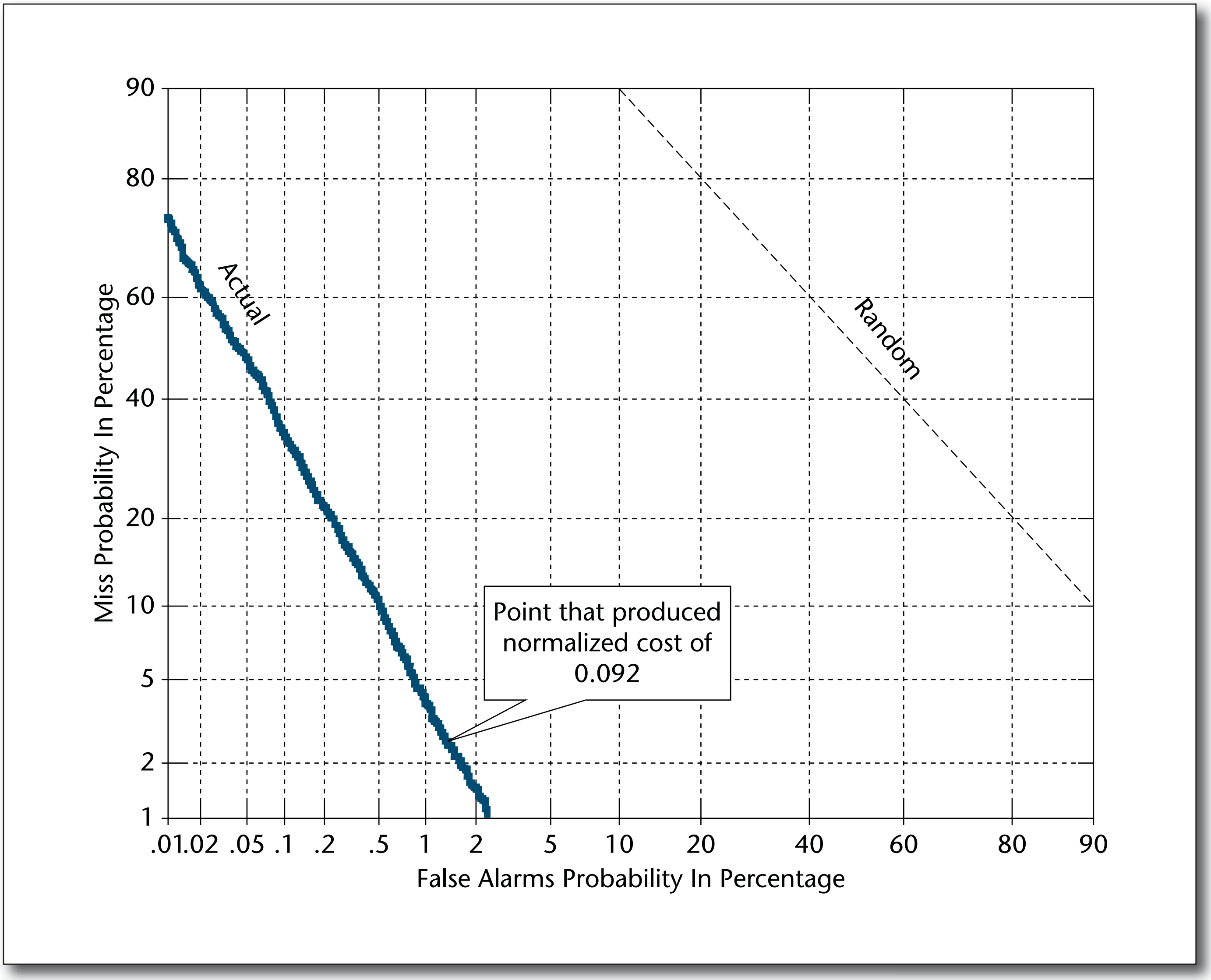


Figure 5. Tracking Results for Best System.

Figure courtesy of the National Institutes of Standards and Technology.

source language (Arabic, Chinese, English), medium (speech or text), and genre (formal or informal). Unlike previous speech programs that focused on cutting WER or translation programs that raised BLEU scores, GALE focused on overall utility as judged by humans. Its Go/No-Go criteria were stated in terms of translation and distillation accuracy.

Additional Programs

In addition to the 10 seminal programs described above, other programs advanced HLT capabilities in various ways, including understanding spoken commands in extreme noise, for example, inside tanks (SPINE); performing optical character recognition on handwritten and printed text and translating them into English text (MADCAT); performing speech activity detection, language identification, speaker identification, and word spotting in noisy and degraded signal environments (RATS); performing automatic translation of and information retrieval from, informal text (BOLT); performing two-way speech-to-speech translation using human-machine dialog for error correction and ambiguity resolution (TRANSTAC); creating knowledge bases from multilingual text by consolidating information regarding entities, events, relations, and sentiments (DEFT); and creating capabilities for extracting entities and events from low resource languages (LORELEI)

Two new programs are now focusing on creating a semantic engine to automatically generate multiple alternative interpretations of events, situations, or trends, based on a variety of unstructured multimedia, multilingual sources that may be noisy, conflicting, or deceptive (AIDA), and developing a semiautomated system that identifies and links temporally sequenced complex events described in multimedia input, identifying participants, subsidiary elements, and event types (KAIROS).

Other programs adapted and demonstrated HLT capabilities in various military applications.

Open Problems

Another generation's worth of research remains to be done, and current DARPA programs are getting it started.

Technology

In technology, automatic transcription, translation, and content analysis must become more robust, able to perform well on ever more diverse and difficult types of speech and text. Systems must learn to set decision thresholds automatically and output trustable confidence figures. Methods must be devised to port HLT capabilities to new languages, problems, and domains, at low cost and with smaller amounts of manually annotated training data. New types of automatic content analysis must be defined and developed, including discourse and conversation analysis in both speech and text. Effective automated scoring methods are needed for all types of automatic content analysis. Better integration of HLT and image understanding technology with large-scale knowledge bases and systems that can actively explore the physical world must be achieved.

True language understanding does not yet exist. It is a Holy Grail that should be pursued.

Applications

Important new HLT applications are being developed using current state-of-the-art tools. As researchers make progress on the challenges described above, we can expect to see widespread use of improved versions of applications that now exist in limited forms, as well as many new applications not yet imagined. Here are just two of these many opportunities: First, we expect many educational applications. These include areas where speech and language are the thing being taught, such as foreign-language learning and the many levels of writing instruction. More broadly, as conversational and user-modeling abilities improve, we will see intelligent tutoring systems applied pervasively to teach (and test) subjects from accounting to zoology. Second, there are enormous opportunities for automatic analysis and monitoring of the linguistic correlates of clinical categories such as Alzheimer's disease, mood disorders, and schizophrenia. For hundreds of years, physicians have diagnosed neurocognitive health from the way that people talk. For many decades, these subjective evaluations have been reinforced by hand-calculated quantitative scores on neurocognitive tasks with verbal responses. Over the past few years, researchers have begun to use speech technology and ML to infer additional diagnostic information from recordings of such tasks and also from recordings of interviews, picture descriptions, and other interactions that previously were analyzed only subjectively.

Summary

During several decades of research, DARPA HLT programs created enormously valuable core capabilities for automatic transcription, translation, and content analysis — changing science fiction to social fact.

DARPA created these game-changing capabilities by setting crisp, aggressive, quantitative technical objectives; soliciting innovative ideas for solving them; selecting strong multidisciplinary research teams; providing strong multiyear funding; exploiting large quantities of linguistic data; conducting objective performance evaluations; and making course corrections based on those results, and iterating multiple times.

Out of this, our four key technical lessons emerged: learning is better than programming; global optimization of gradient local decisions is crucial; top-down and bottom-up knowledge must be combined; and metrics on shared benchmarks matter.

The Common Task Method (multiple parties sharing resources, competing, and collaborating to achieve a stated objective) was extremely powerful, efficient, and easy to administer.

Acknowledgments

Many people contributed valuable information and advice. The following were particularly helpful in the indicated areas: Raj Reddy, CMU, (Speech Understanding Research program), Allen Sears, CNRI, (many programs), Joe Olive JP Olive S&T Management, (figure 3, many programs, overall critique), Jon Fiscus at NIST (figure 1), and external reviewers (corrections, editorial suggestions). This project was supported by DARPA.

Figure 2 was produced by Jon Fiscus (NIST) at our request. Figure 3 was produced by the authors. Figure 4 was provided by Joe Olive; it appeared in an overview of the GALE program that DARPA approved for public release. Figure 5 appeared in "Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation," a paper that Charles Wayne presented at the Second International Conference on Language Resources and Evaluation, 2000; it went through National Security Agency prepublication review.

Mark Liberman began his career at AT&T Bell Labs in 1975, and moved in 1990 to the University of Pennsylvania, where he is a professor in the departments of Linguistics and Computer and Information Science. He has participated in DARPA's HLT programs since the mid-1980s. His involvement in the open data movement began with the Association for Computer Linguistics' Data Collection Initiative in the 1980s, and continued with the provision of shared data for DARPA and other HLT programs, and the founding of the LDC in 1992. His current research activities include methods for easier development of resources for languages that lack them, and the application of data-intensive linguistic analysis to clinical, educational, and legal issues.

Charles Wayne played central roles in government-sponsored efforts to develop effective HLT during a long career in the defense and intelligence communities. He served as a program manager at DARPA from 1988 to 1992 and again from 2001 to 2005.