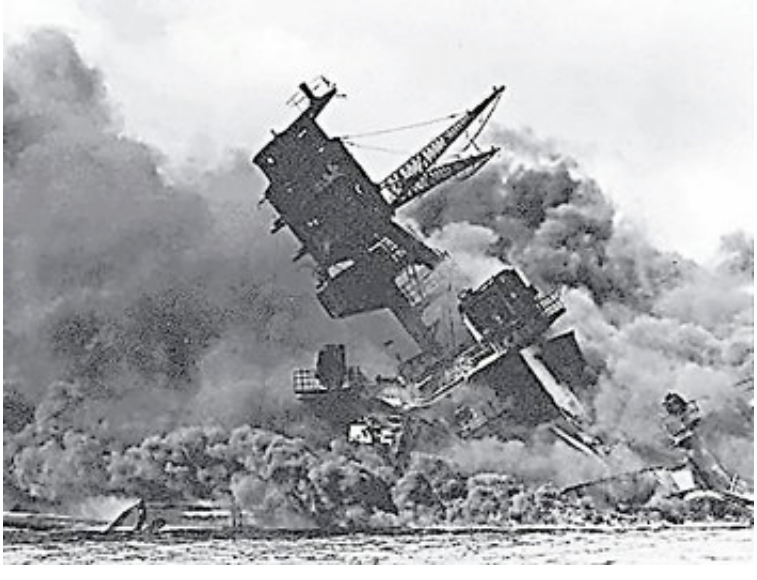
Holy grail of data-sifting proves elusive

Goal is system that will prevent attacks on U.S.

By Peter Rowe, UNION-TRIBUNE STAFF WRITER

Monday, December 7, 2009 at 1:25 a.m.



/ AP file

The battleship Arizona toppled into the sea after being hit during Japan's attack on Pearl Harbor on Dec. 7, 1941.

Call it the Pearl Harbor Paradox. On this day 68 years ago, the United States was thrust into World War II when Japan devastated the Pacific Fleet's base in Hawaii, sinking four battleships, destroying 188 aircraft and killing more than 2,400 military personnel. Historians still debate the events surrounding the attack, but we know that U.S. intelligence had cracked the Japanese codes.

That information, though, came too little, too late.

In today's era of cell phones, blogs, Facebook and instant messages, the stockpile of information has done little to ease the fears of those trying to prevent the next Pearl Harbor or Sept. 11. But those fears have changed. Instead of too little, too late, the new worry is too much, too late.

"There's a lot of free data floating around in the ether," said Jean Mark Gawron, a San Diego State University professor of linguistics and a former Pentagon consultant to the Defense Advanced Research Projects Agency, or DARPA. "The only problem is there is a huge amount of it."

If that's the problem, isn't the solution obvious? Most Americans regularly use automated devices to sift mountains of data for nuggets of fact. They're called computers. Long ago, scientists began searching for the intelligence community's holy grail: an automated system capable of collecting data from all available sources, accurately translating the information into English and then connecting the seemingly unrelated, but relevant, dots.

Oh, and it should do all of this in mere moments.

"A system that takes a week to discover a bombing that will occur in a day isn't very useful," said Andy Kehler, a professor of linguistics at the University of California San Diego and, like Gawron at SDSU, a former DARPA consultant. "They're throwing a lot of money at this problem, but I don't know how much progress is being made."

Problems remain, conceded Beth Sundheim, who studied these systems for the Navy until her retirement in 2006. There are technical issues, though researchers believe they will yield to technical solutions. The bigger problem is that no computer program has been able to bridge the greatest obstacle — the inability of a machine to think like a human.

Mucking up the MUC

Among the missions of the U.S. armed forces is the daunting task of preventing attacks on the nation or its interests. This is a battle fought by tanks, jets and ships, but also by computational linguists. These scientists use statistics and the rules of "natural language" — the words people speak and write — to help design computer systems.

This is surprisingly tricky task. Computers may be able to outcalculate the smartest mathematician, but even the most powerful machine can be stymied by the simplest wordplay. During a 1995 conference sponsored by DARPA, computers were foiled by a particularly devious code: Wall Street Journalese.

The Message Understanding Conferences — known by the unlovely acronym MUC — were command performances by federally funded researchers who were designing automated systems to sift information. If you were taking money from the National Security Agency or DARPA and wanted to continue to receive funding, you participated in MUC.

"These were the bake-offs of information extraction," Kehler joked.

In the 1995 conference, the task was to search 100 Wall Street Journal articles for news about changes in management. Who was being hired? Fired? Promoted? Reassigned?

The researchers built systems, ran the exercise and submitted the results for independent grading.

The winning team analyzed the articles with 70 percent accuracy. In other words, it earned a C-.

What happened?

English happened. The Journal's writers had used the language to its fullest, raiding its rich storehouse of slang, euphemism, shorthand and ambiguity.

"An article might say that someone 'turns over the reins' to someone else," said Sundheim, who helped organize MUCs while working for the Naval Ocean Systems Center and its successor, the Space and Naval Warfare Systems Center. "The system is not sure what this means."

Linguistic nuances

MUC was succeeded by DARPA's Global Autonomous Language Exploitation program. DARPA noted that earlier experiments had been conducted in almost ideal conditions: All of the data were printed in one language and professionally edited. In 2007, GALE was assigned higher accuracy targets for "structured data" such as the Journal's news than for "unstructured data" such as conversations.

Other variables can lead to misunderstandings. Suppose your system is dealing with audio recordings and written material. Or that the sound or print quality has degraded. Or that two or more languages are being used. Or that some sentences are spiced with sarcasm.

While computers are notably adept at technical tasks — removing layers of unwanted sound from a taped conversation, for example — linguistic nuances are generally beyond them.

Take a site devoted to film reviews. A critic writes, "If you're looking for one of the best war movies of the last two decades, you'll have to keep looking."

The computer, seeing "best" and "last two decades," classified it as a positive review.

"Language is a human object," Kehler said. "Humans know a lot of information about the world, and they bring that to bear when they use language. Computers don't."

With spoken language, this gap tends to widen. Computer systems have to make logic-defying leaps to stay abreast of the typical conversation. People speak in sentence fragments, punctuated with "ums" and other verbal tics. They may pick up a topic they had abandoned minutes earlier. They may refer to five or 10 people yet mention only "he" or "she."

The computer has to discern the meaning and simultaneously monitor millions of other messages to find links.

"You have one e-mail that says a guy is traveling to Dallas," Kehler said. "Another e-mail says the president is traveling to Dallas at the same time. A third says this guy is the brother of someone on the watch list. Any one of these e-mails is not interesting by itself."

Together, though, it may be a pattern worthy of the Secret Service's attention.

Keystrokes and calls

GALE's goal is to provide automated, near-real-time translation of Mandarin Chinese and Arabic with 90 percent to 95 percent accuracy by 2011. Will this ambitious experiment meet its deadline? In 2006, the project announced it had reached 75 percent accuracy with written Mandarin and Arabic, 69 percent with spoken Arabic and 67 percent with spoken Mandarin.

Kehler, a veteran of the GALE project, believes progress is being made, but slowly. Despite Washington's excitement about the system's potential, this remains a daunting scientific challenge. "You cannot legislate breakthroughs, Kehler said.

Still, a system like GALE might have prevented the Nov. 5 shootings at Fort Hood, Texas. FBI agents in several offices, including San Diego, failed to compare e-mail messages between the suspect, Maj. Nidal Hasan, and a radical imam in Yemen. In theory, GALE would have processed all these messages in nanoseconds, noticing links that human agents had missed. In fact? Mistakes can be made.

"The more complicated the data set you are trying to generate, the more likely you'll have things there that shouldn't be there, and have things missing that should have been there," Sundheim said.

But GALE is designed to sift intelligence, an often slow task, at light speed. This would hasten the delivery of data to analysts — who, being only human, might still make mistakes.

Even outside this cloak-and-dagger arena, technology is moving us toward a sort of global info-grid. IBM's current TV commercial says we are fast approaching "1 trillion connected devices in the world. Can you hear them? Food is talking to store shelves. Cargo containers are talking to supply chains. Power grids are talking to the grid."

"Now that's smart," the ad maintains.

For anyone who cares about privacy, it's also frightening. Is the government listening as our devices speak to the world? Even if this surveillance prevents another Sept. 11 or Pearl Harbor, you may wonder who's tracking your keystrokes and calls, your movements and messages.

Gawron doesn't wonder. He knows. "If you care about whether your conversations are private," he said, "you should never send e-mail messages unencrypted."

And our cell phone conversations?

"Same thing."

Peter Rowe: (619) 293-1227; peter.rowe@uniontrib.com
