[*PC Magazine* 13 July 2004]

Scaling the Language Barrier

By Sebastian Rupley

In the annals of computer comedy, one of the most famous anecdotes is about asking a speech recognition engine, “Recognize speech?” The translation comes back: “Wreck a nice beach.”

Getting machines to understand both spoken and written language has been an elusive goal for the tech industry for many years. Now, thanks to a wave of government funding and technical breakthroughs, machine translation (and understanding) of written language is getting unfunnier by the minute.

Part of the reason accuracy rates for machine translation have crept along at a snail’s pace until now is that translation software needs a very large database of text for training and for comparing similar phrases to extract meaning. Huge increases in storage capacity have greatly boosted the size of such databases.

“We’re making use of terabytes of data now,” says Steve Klein, chairman and CEO of Meaningful Machines, a developer of translation technologies. “That boosts accuracy enormously and is why we like to say that our technology is for machine understanding, not just for machine translation.”

Klein and his start-up company have kept the software they will release later this year hush-hush, but Klein bills it as revolutionary. He says it could have a broad impact on everything from the U.S. government's machine translation of foreign documents to Internet search engines that recognize voice.

The one clue Meaningful Machines has given about its software is that it will use new methods of statistically ranking the likelihood of what entire phrases mean, rather than just translating one word at a time. That allows it to discern whether the word baseball in a given phrase refers to a ball or a game.

Phrase-based, statistically ranked translation is also the secret sauce in Language Weaver’s machine translation software. The company recently got a gust of financing from In-Q-Tel, the venture capital arm of the CIA, which would like to use the software not just for translating documents from numerous languages but for automating machine understanding of text to flag statements that might be dangerous or revealing.

Machine translation software could also make today's Internet search engines seem like relics from the distant past. “We’re only a few years away from Internet search engines that can return high-quality results translated from nearly every language around the globe,” says Daniel Marcu, founder of Language Weaver. Eventually, software will be able not only to understand spoken language but also to act upon it.

“DARPA [the Defense Advanced Research Projects Agency] is funding numerous research efforts doing cross-language information retrieval,” adds Marcu. At the recent DARPAtech conference, the organization showered praise on a handheld machine translation gadget called the VoxTec Phraselator P2. Speak into the Phraselator and it instantly translates phrases into spoken Arabic.

Carnegie Mellon University, the University of Southern California, and Microsoft Research operate some of the largest programs for developing machine translation software. Microsoft is primarily focused on extracting meaning from documents in English.

“We don’t have to reach perfect accuracy rates to get to very useful applications,” Marcu points out. In machine translation, as in horseshoes and darts, “close enough” counts for valuable points.