TEACHING COMPUTERS TO TRANSLATE JAPANESE

In 1991 Jon Yamron thought he "would look around and see what was out there." With a 1987 PhD in theoretical physics from the University of California, Berkeley, and postdocs at the Institute for Advanced Study in Princeton, New Jersey, and the State University of New York, Stony Brook, under his belt, as well as a wife with a career of her own to consider, Yamron wanted to settle down but found himself face-to-face with the rather bleak job market for physicists.

Unsure of finding an academic slot, he broadened his search to include the commercial sector. A chance connection via his mentor, Warren Siegel, led to an interview at Dragon Systems Inc, a developer of computer-based speech-recognition systems in Newton, Massachusetts. Yamron recalls being rather amused when told he could give a seminar to the Dragon staff on any subject, even string theory, at the interview. His talk on that subject in fact elicited "a pretty good debate" from an enthusiastic audience, which turned out to include a high proportion of physicists.

During the interview Yamron also saw a demonstration of Dragon's computer-dictation system, which turns the spoken word into text in a word-processing document. The system includes a speech board, software and a headset and has a 30 000-word vocabulary; it runs on an IBM-compatible PC with an Intel 80386 processor. "I was completely blown away," Yamron says. "I couldn't believe what they were doing was even possible, and I thought I was in touch with what was going on with computers."

The demonstration changed his career path. "When looking at opportunities outside academic physics, I had thought of computers, but I didn't want to just program all day. It's hard to see what could be more interesting than speech recognition and similar artificial intelligence problems."

Once on board at Dragon, Yamron joined a new project that was being funded by the Defense Department's Advanced Research Projects Agency to develop a computer-aided, interactive Japanese-text translation system. The project goal is not to replace human translators with a computer but to enhance their capability. ARPA is funding similar work by the IBM Research Division and by a consortium of three universities: Carnegie–Mellon University, the University of Southern California and New Mexico State University.

At Dragon, the text-translation system, known as Lingstat, relies heavily on statistical algorithms applied to large bodies of text. The idea is to build a system that can "learn" to translate, rather than to "teach" it by, for example, incorporating definitions and rules. To give the flavor of the statistical approach, Yamron discusses the construction of a bilingual dictionary: "Given only a large amount of parallel text [Japanese sentences and their English translations] and making no assumptions about how words in the two languages are paired, it is possible to construct a statistical model that can learn the translation of individual words. For a particular word, the basic idea is to find all of its occurrences in the English half of the parallel text, then determine the Japanese word that most consistently occurs in the matching sentences of the Japanese half."

The actual procedures used by Lingstat are based on the estimation-maximization algorithm. In this algorithm, the statistical analysis proceeds iteratively. In one model, the initial set of probabilities is the same for all words, so any word in one language can translate into any word in the other language with equal probability. At the end of one pass, the algorithm begins to refine the probabilities as it determines which words are found in the same sentences; it uses these probabilities as the input for the next pass. Eventually, the most probable translation remaining is the correct one.

This example demonstrates how to use parallel texts to discover translations of words, but more sophisticated models can extract other kinds of information from large texts, such as how to translate phrases, recognize idioms and even relate grammatical constructions between two languages.

Jon Yamron

The Lingstat translation system is not yet complete. In its current form, the system presents sentences broken into words, word pronunciations and best-guess translations of each word to the user, who works with a word processor. In tests on human translators having various levels of expertise, the system has increased the average speed of mid-level translators by about 30%, and it has enabled those with a year or so of Japanese-language study to make rough translations of texts that are comparable in complexity to newspaper articles.

Computer translation of Japanese is a long way from theoretical physics. "I had an interest in physics as far back as I can remember," Yamron recalls. "When I was about 13, I had to write an essay on what I wanted to be, and I picked physicist. That was before I even had my first physics course." Significantly, he also developed an interest in computers when

his sixth-grade classroom was equipped with one. "I always had a little computer project on the side."

During his undergraduate studies at Princeton University, Yamron was drawn to general relativity, and he wrote junior and senior papers on the subject. Upon graduating in 1980, he headed west to Berkeley for a change of venue and to continue his physics studies. There his eyes were opened to the fascinating role that symmetry plays in elementary particle theory.

At Berkeley, Yamron met Siegel, who was an expert in supersymmetry. When Siegel later moved to the University of Maryland, the association continued at a distance and the subject of study shifted to string theory. Strings, of course, have been proposed as the most fundamental of all entities and are the current best hope for unifying the forces of nature into one formalism—a so-called theory of everything. Siegel was investigating how to formulate strings as a field theory, an alternative to the usual treatment, which has more in common with particle mechanics. Yamron's topic was the Ramond string, the fermionic half of a "real" supersymmetric string.

During his postdoctoral stint at the Institute for Advanced Study, Yamron married Janna Leonoff, who was developing a career as a registered dietitian and had a private practice that relied heavily on personal referrals. Each move required her to rebuild her practice anew. "Our nomadic existence, moving every two years, made it difficult," Yamron recalls. Looking at the prospect of a third move in four years, he found the highly interesting work, the chance for some stability and the opportunity to return to his home state all combined to make a position at Dragon Systems an offer too good to refuse.

One and a half years after making the switch from academe to the corporate world, Yamron has found that life in his new environs offers one considerable advantage that string theory lacked. "One of the nice things at Dragon is that I can actually do an experiment. I can make a computer model, test it and see if the model is any good."

On those occasional days when things are not too busy, Yamron's thoughts may stray back to former times, but for just an instant. "There are still lots of interesting problems in physics, but I'm working with physics and mathematics PhDs who are valued for their way of thinking, their way of addressing problems and their creativity. I see doing what I'm doing now for many years to come."

—Arthur L. Robinson ■