Essays of an Information Scientist, Vol:4, p.574-578, 1979-80 Current Contents, #33, p.5-9, August 18, 1980

Current Comments

Can Machines Be Scientific Translators?

Number 33 August 18, 1980

Within the last few years portable "electronic dictionaries" have been introduced.[1] These gadgets were originally intended to help foreign travelers communicate with the natives. The devices, which cost from $200-$300, are initially a lot of fun. But often they prove to be less useful than a good bilingual dictionary or phrasebook because they are so limited in vocabulary. The models I tested[1] are limited to about 2,000 words and phrases. Texas Instruments and Sharp Electronics have each unveiled speaking translators.[2] But before such machines can compete successfully with printed dictionaries, they'll have to include from 5,000 to 10,000 items. Still, about 200,000 non-speaking electronic dictionaries were sold last year in the US, and 200,000 more worldwide.[3]

These machines work on a word-to-word basis. However, scientists have been working on a more sophisticated means of machine translation (MT) for about 30 years. When MT began, researchers hoped that machines could be made to translate with little or no human editing. So far this hasn't been accomplished. But many MT researchers believe that human editing of machine output is a viable approach to MT. This process is called machine-aided translation (MAT). The idea behind MAT is that machines can free human translators from the repetitious parts of the translation task. Critics of MT and MAT believe that terminology banks, as opposed to machines designed to carry out translation tasks, are a more effective means of aiding the flesh-and-blood translator.

In 1949, Warren Weaver of the Rockefeller Foundation was the first to suggest MT seriously in a "memorandum" distributed to 200 colleagues.[4] Weaver proposed that "electronic computers of great capacity, flexibility, and speed" be applied to translation. He believed this would promote greater cultural exchange and international understanding. Such an MT system would not merely translate one word to another, but translate whole sentences. He advanced this analogy to describe how MT might be accomplished:

> Think...of individuals living in a series of tall closed towers, all erected over a common foundation. When they try to communicate with one another, they shout back and forth, each from his own closed tower. It is difficult to make the sound penetrate even the nearest towers, and communication proceeds very poorly indeed. But, when an individual goes down his tower, he finds himself in a great open basement, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers. Thus may it be true that the way to translate from Chinese to Arabic or from Russian to Portuguese, is not to attempt the direct route, shouting from tower to tower. Perhaps the way is to descend, from each language, down to the common

> base of human communication—the real but as yet undiscovered universal language—and then re-emerge by whatever particular route is convenient.[4]

Weaver's proposal sparked much interest in MT. Indeed, some of MT's proponents were far too optimistic. Hopes were high that fully automated, high quality MT was imminent. By 1966, 17 US universities and other institutions, as well as the Department of Defense and the CIA, had spent about $20 million trying to crack the problem. Lack of progress towards total automation of translation resulted in a backlash of disappointment.[5] In 1966, the National Academy of Sciences released the report of its Automatic Language Processing Advisory Committee (ALPAC).[6] ALPAC, chaired by J.R. Pierce (then with Bell Labs), concluded that MT was not worth pursuing. However, the committee did recommend that research on the theory and practice of translation be funded.

For its negative stance on MT, ALPAC was criticized for short-sightedness. But Weaver and many others had perhaps been overly optimistic about MT. I'm confident he was not unaware of the possible limitations of *any* highly developed MT system. The use of second generation computers would be very limiting, but fundamental linguistic problems also needed a lot more attention. But if the MT field did not develop as rapidly as was hoped, it was partly because too few of the people in the field were pragmatists. No one, apparently, anticipated the relevant economic question of getting text into machine-readable form. The cost of keying in texts in Russian is still not trivial. Optical character readers eventually may reduce these costs.[7]

The goal of MT is to translate words and sentences from a "source language" (SL) to a "target language" (TL). If one is to accomplish more than "mere" word-for-word translation, one must be able to identify each grammatical unit of the SL sentence. This is sometimes called parsing. Putting sentences into grammatically correct form is only half the battle, however. A sentence can make grammatical sense, but can be semantically incorrect or meaningless.[8]

Even trained human beings can't always translate perfectly. But they can perform fairly well, because they know enough to resolve the ambiguities and metaphors of natural language. Speakers of a given language also share common assumptions and knowledge about the world in general, or a given specialty. The presence or absence of these assumptions varies from person to person and from culture to culture. But they are required to make appropriate inferences. Computers are as yet poor at making such inferences.

The inability of machines to infer led to some amusing errors in the early days of MT. In 1962, President Eisenhower's interpreter, Colonel Vernon Walters, told a *Harper's* author about a group of engineers who built a translating machine.

> [They claimed] that it would translate instantly without the risk of human error. In the first test they asked it to translate the simple [English] phrase: "Out of sight, out of mind".... The machine typed out in Russian: "Invisible idiot." On the theory that the machine would make a better showing with a less epigrammatic passage, they fed it the scriptural saying: "The spirit is willing, but the flesh is weak." The machine instantly translated it, and came up with "The liquor is holding out all right, but the meat has spoiled."[9]

However, such errors are no longer a problem if one assumes unlimited computer memory and a thorough dictionary of such common sayings.

Sentence-to-sentence translation has been attempted many different ways.

Some MT systems used one bilingual dictionary; others used separate dictionaries for SL analysis and TL synthesis. Some dictionaries contain all forms of a word; others try to match base forms with endings (such as *-ed* or *-ing*).[10] The methods by which sentences are scanned also vary. Some systems scan sentences from right to left. Others break sentences down into small phrases and scan them separately. Others try to break those phrases down into still smaller phrases.

Most MT systems since the 1960s have tried to develop a grammar, on a general set of rules, to handle translations. But researchers often needed to develop rules on an *ad hoc* basis, as exceptions were always cropping up. W.P. Lehmann and colleagues at the University of Texas Linguistic Research Center warn that a too-theoretical approach to building MT grammars doesn't work well: "MT grammars thus tend to be eclectic, applying features and concepts from a number of linguistic theories to the problem of parsing one language and generating another." The Texas team says its German-English system is still in the research stage. They tested it on 50 pages of German telecommunications text and said that 84% of the sentences in that sample were translated correctly.[10]

One of the earliest MT systems to attract wide attention was developed by Leon Dostert, Michael Zarechnak, and colleagues at Georgetown University. By 1960 they had four groups experimenting with different Russian-English systems. (I almost went to work with them in 1954. Instead I went to Philadelphia to consult "temporarily" for Smith, Kline & French. This was how ISI® got started.[11] The proximity of MT and so-called mechanical or automatic indexing was so obvious to me that I studied structural linguistics under Zellig Harris at the University of Pennsylvania. This came about in 1955 when my friend Casimir Borkowski introduced me to Harris. Cas had worked with Dostert but was taking his doctorate at Penn. He is now professor of computer science at the University of Pittsburgh.[12]) A survey of 58 users of the Georgetown systems at EURATOM and Oak Ridge National Laboratory reported that 92% rated unedited translations "good" or "acceptable"; 87% of the sentences were deemed correctly translated; and 76% of the technical terms were judged intelligible.[5]

Another well-known MT system is SYSTRAN, developed in 1968 by Peter Toma of LATSEC, Inc., La Jolla, California. Toma tells us that SYSTRAN contains over 100,000 computer instructions for SL analysis and TL synthesis. The US Air Force has used SYSTRAN since 1970. Toma claims, "The quality of these translations is so high that 90% of them are used without post-editing." The National Aeronautics and Space Administration used Russian-English and English-Russian versions of the system during the 1975 Apollo-Soyuz project.[13] Toma is quite optimistic about MT, but says it is cost-effective only when huge amounts of material must be translated.

Some MT systems take an approach similar to that suggested by Warren Weaver. That is, they use artificial languages to mediate between SLs and TLs. The idea behind intermediary languages is to write a program that can deal with any language pair, rather than a separate program for each pair.

The LOGOS system, a product of the LOGOS Development Corporation of Middletown, New York, employs an intermediary language, according to company spokesman Jack Kelly.[14] A Canadian MT project called TAUM (Traduction Automatique de l'Université de Montréal) is developing a system that takes the intermediary language approach. The University of Montreal team, headed by Marcel Paré, is building

a system called AVIATION, which is designed for English-French translation of aviation manuals. The project should be operational later this year.[15] Since 1962, TAUM has been operating a system for English-French translation of weather reports. The system, called MÉTÉO, does not use the intermediary language approach. Instead, it relies on a stereotyped format used to write the weather forecasts. The system translates about 2,000 reports a day, with an error rate of about 15%. MÉTÉO translations have been broadcast to the general public since 1976.[5]

Since all the MT systems developed to date require human intervention, some researchers are developing systems of human-aided interactive machine translation. One such system is being developed by linguist Alan Melby and co-workers at Brigham Young University in Provo, Utah. In this experimental system, humans help the machine resolve syntactic and semantic problems. The computer can easily translate a sentence like, "He washed the dishes with a rag." But it could have a problem with: "He washed the dishes with Susan." Human aid lets the computer know that the man and Susan are washing the dishes together.[16] Weidner Communications, Inc., La Jolla, California, has also developed a MT system with an interactive component. It allows the human translator or editor to correct words or phrases at a video-display terminal.[17]

The brief *New Scientist* report[16] on Melby's work suggests that MT will replace human translators. However, nobody has yet developed a system that works without human aid. And there is an acute shortage of translators. In the US, according to one estimate, there are at most 600 translators in government and business.[18] It is apparent that computer technology is the translators' ally, not their competitor.

However, MT is not necessarily the only way computers can benefit the translator. Another means is computerized dictionaries, or terminology banks. It has been estimated that translators spend about 60% of their time looking up new or unfamiliar words.[19] On-line dictionaries can speed this process. Another advantage of on-line dictionaries is that they can be constantly up-dated. In contrast, printed English-French dictionaries are typically issued every 2.4 years.[19] Also, printed dictionaries are limited in size; for example, our *Transliterated Dictionary of the Russian Language* contains about 20,000 items.[20] By putting our files on-line we could store millions of words, phrases, idioms, etc. The limitation is not in the technology of storage but rather the frequency of use to justify the cost of storage. I'll be saying more about ISI's on-line dictionary activities in the near future.

On-line dictionaries have been developed in about 20 countries since Siemens AG in the Federal Republic of Germany began operating one about 15 years ago. In 1975, Carnegie-Mellon University developed the US's first foreign language terminology database. In their evaluation of CMU's system, A. Andreyewsky and D. McCracken are very critical of MT. They view it as a misguided search for a solution to the US's decline in foreign language skills; they believe it costs too much to edit MT and that it's too slow. As an alternative to MT, they propose an international network for terminology exchange. In this scheme, terminology banks would be constantly updated to be accessed by human translators. The US, they propose, should establish a government institution to organize in-

ternational terminology exchange. They believe that automation of translation need not include MT, but that an adequate MT system could be worked into this network.[18]

To many researchers in and out of MT, the 1966 ALPAC Report seemed like the end of MT. But clearly MT is not dead. In contrast to the overzealous statements one heard in the 50s, when MT was said to be just around the corner, there is today a more restrained kind of optimism. Researchers still disagree over whether fully-automated, high quality MT is possible, but technology has come a long way. Whether in the form of pocket-sized electronic dictionaries with at least 20,000 words of memory or microcomputers with almost unlimited disc capacity, practical MAT is clearly within our grasp. Whether these developments have a significant impact on breaking the so-called language barrier remains to be seen. But individual scientists will be able to create their own translations much more easily if they are willing to make the effort.

* * * * *

My thanks to Linda Cooper and Tom Marcinko for their help in the preparation of this essay.

REFERENCES

1. **Garfield E.** Reflections on learning foreign languages. Part II: electronic translators. *Current Contents* (38):5-9, 10 September 1979.
2. **Berger M.** The future is now—and it talks. *NY Times* 20 December 1979, p. C3.
3. **Quirt J.** The solid state linguist makes its debut. *Fortune* 100(7):170-4, 8 October 1979.
4. **Weaver W.** Translation. (Locke W N & Booth A D, eds) *Machine translation of languages.* New York: Wiley, 1955. p. 15-23.
5. **Hutchins W J.** Machine translation and machine-aided translation. *J. Doc.* 34:119-59, 1978.
6. **National Research Council Automatic Language Processing Advisory Committee.** *Language and machines: computers in translation and linguistics.* Washington, DC: National Academy of Sciences, 1966. 124 p.
7. **Garfield E.** Has OCR finally arrived? Or is it a technology with a lot more problems than meet the eye? *Current Contents* (19):5-13, 7 May 1979.
8. **Zachary W W.** A survey of approaches and issues in machine-aided translation systems. *Comput. Hum.* 13:17-19, 1979.
9. **Kouwenhoven J A.** The trouble with translation. *Harper's* 225(1347):38-44, August 1962.
10. **Lehmann W P, Bennett W S & Slocum J.** *Machine translation: 1980.* (Unpublished paper). 43 p.
11. **Garfield E.** What do you do for a living? *Current Contents* (6):5-7, 5 February 1979.
12. ----------. Information science education—an ivory tower of Babel? *Current Contents* (22):5-13, 2 June 1980.
13. **Toma P.** Personal communication. 15 May 1980.
14. **Kelly J.** Personal communication. 19 June 1980.
15. **Paré M.** Telephone communication. 19 June 1980.
16. Machines break down language barriers. *New Sci.* 85:836, 1980.
17. **Weidner Communications Inc.** Brochure, 1980. 6 p.
18. **Andreyewsky A & McCracken D.** *Carnegie-Mellon University Center for the Study of Translation.* 24 April 1980. (Unpublished report). 31 p.
19. **Greenfield C C & Serain D.** *Machine-aided translation: from terminology banks to interactive translation systems.* Pittsburgh, PA: Carnegie-Mellon Univ., 1977. 84 p. ED 168 542.
20. **Garfield E,** ed. *Transliterated dictionary of the Russian language.* Philadelphia: ISI Press, 1979. 382 p.