[From: IFTT’89, April 26-28, Oiso Prince Hotel, Japan]

THE TOOLS FOR THE JOB:
PREREQUISITES FOR EFFECTIVE MT UTILISATION

Veronica Lawson

Consultant
United Kingdom

Last month, at the Georgetown University Round Table on Linguistics, we celebrated 35 years of machine translation on electronic computers. The Georgetown/IBM experiment took place in 1954, when the computer itself was only a few years old. It was an exciting new tool, incredibly fast by the standards of that time. Yet to us today it was crude and very slow, and it is hard for us now to imagine the early days of machine translation (MT): the lack of knowledge, the lack of technology. People did not know what MT could do, and they did not have the tools to utilise it to the full. Development, the actual translating, input/output, editing, updates - all were far slower and more tedious then.

This paper about practical MT will summarise how the technology has improved in four of the areas which affect MT utilisation:

input;
text preparation;
postediting;
dictionary development.

It will then outline areas for effective MT utilisation, given the tools for the job.

Input

Input, in those distant days, was on punched cards; boxes full of cards had to be carried over to the computer centre. Punching the cards was expensive - so expensive in the USA that the cards were prepared in Germany instead, and then flown over the Atlantic. There was no question of a rapid turnaround. But over the years it became possible to key text directly into the computer, or to feed it in via various media, first in house and then via telecommunications. In recent years office automation has dramatically cut the effort, time and cost of input. Many documents are prepared with wordprocessing or can be fed to an optical scanner. So MT services often accept only documents which are already in machine-readable form.

An example of what the progress in input technology means to a customer was given by the Nuclear Research Centre at Karlsruhe, in Germany. When input was manual, a 500-page translation took four weeks to input. When the Centre bought an optical scanner, input time dropped to only three days. The machine translation by Systran takes half an hour to run and is then ready: the Centre's scientists use raw output, and so there is no delay for editing. Thus the introduction of optical scanning cut turnaround time from four weeks to three days.

Text preparation

There have also been advances in text preparation: the pre-processing of text specially for MT, either to remove errors or to make the text simpler.

Typographical or grammatical errors need to be removed because they can render text nonsensical and impossible to parse. Ten years ago, when I began to work in MT, I could see that we sorely needed to clean up source documents, and I suggested using the spelling checkers and text critiquing software which had by then been developed for monolingual environments. Now software of this type is commonplace.

As for making text simpler, there are two approaches. One way is to take text originally written in natural language, and to pre-edit it to remove ambiguities, etc. This is not a trivial operation: it has been compared to half translating the document in one's head, and it takes time. One international organisation found that, as a general principle, human intervention of any kind in the translation process added a minimum of three days to turnaround time. Also, purely human pre-editing is unreliable, because our world knowledge clarifies for us much that remains ambiguous to a machine. Postediting may therefore still be necessary.

Another way of making text simpler is actually to originate the text in controlled language, with limited syntax and/or a limited vocabulary. Both pre-editing and controlled language tend to be uneconomic unless text is to be translated into more than two target languages. Much more is known about them than before, however, and we shall hear a paper on controlled language shortly. Software such as the Smart Editor (developed for use upstream of the Smart Translator MT system) provides valuable computer assistance. It is noteworthy that some users of controlled language have drastically reduced the number and complexity of their rules over the last ten years.

Postediting

One of the greatest changes in technology for MT utilisation has been the invention of wordprocessing. It helps with input and pre-editing (including the correction of errors made by optical scanners), but most of all with postediting. Even a few years ago most posteditors were still correcting MT printout by hand. The work was very tedious, very time-consuming. The result was often a remarkable mess - what we call a dog's dinner - which the typist or even the customer then had to decipher. Editing on screen is much easier, quicker and more reliable, and retyping is eliminated. Many standard wordprocessing functions are extremely useful; for example, global search and replace functions can be used to replace not-found words or other defective vocabulary. In addition, however, the editor can write macros: mini-routines for performing editing functions which, though not needed in an ordinary office environment, are irritatingly common in the editing of machine translation. Examples include the rearrangement of words, the deletion of articles, and many more.

Another development has been the partial automation of postediting. The US Air Force, which machine-translates a very large volume and variety of technical text with Systran for expert analysts, has been using semi-automated postediting for some years. The routine looks for certain known hazards, such as inversion of word order. Any passage containing such a hazard is brought up automatically onto the posteditor's screen. The editor then examines the passage and makes any necessary amendments .

A further change, again well established for some years, is remote postediting. Raw MT is sent to editors at remote sites by telecommunications or on diskettes, to edit on their own personal computers. This is now common practice among many translation users, whether corporations, MT services or international organisations .

Dictionary development

Easy, rapid dictionary updating is vital to efficiency and motivation. Here too there have been major changes, for updating too has been partly automated. It used to be difficult and slow: coders needed lengthy training before they could fill in the complicated coding sheets; updating was performed in large batches, and delays of two months were common. Ten years ago, however, Weidner encouraged customers to perform a dictionary search and update the dictionary before running a translation. They therefore offered screen-based updating, with prompts and defaults. Other vendors followed suit. Logos customers can even input linguistic rules - a facility some value highly.

Effective MT utilisation

Given all these improved tools, how can MT be used effectively? MT's strengths are speed (or volume), ease of integration into automated document handling and publishing systems, consistency of terminology. It is therefore best suited to large volumes of fairly predictable material, usually technical, preferably in machine-readable form and perhaps intended for circulation or further processing in that form. Many of the texts translated in the industrialised countries constitute such material. MT is being used in practice for a wide variety of purposes. Its chief uses are for information (patents, abstracts and other technical documents of many types, minutes of meetings and other working documents); for scanning (to make abstracts, to identify passages which need postediting or human translation, or even to use as a basis for dictating a human translation); and even for publication (manuals, weather bulletins, etc.). But not for Shakespeare, or only for fun. Not for any literary or personal texts, in fact, since their style varies widely, and they rely heavily on nuance. And not for speech, or at least not yet.