Lost in the machine translation

You can go out right now and buy a machine translation system for anything between £100 and £100,000. But how do you know if it's going to be any good? The big problem with MT systems is that they don't actually translate: they merely help translators to translate. Yes, if you get something like Metal (very expensive) or GTS (quite cheap) to work on your latest brochure, they will churn out something in French or whatever, but it will be pretty laughable stuff.

All machine-translated texts have to be extensively post-edited (and often pre-edited) by experienced translators. To offer a useful saving, the machine must make the time the translator spends significantly less than he or she would have taken by hand.

Inevitably, the MT manufacturers' glossies talk blithely of 'a 100 per cent increase in throughput', but scepticism remains. Potential users want to make their own evaluation, and that can tie up key members of the corporate language centre for months.

A few weeks ago, translators, system developers, academics, and others from Europe, the US, Canada, China, and Japan met for the first time in a Swiss hotel to mull over MT matters. A surprisingly large number of European governmental and corporate organizations are conducting expensive and elaborate evaluations of MT, but they may not produce 'buy or don't buy' results.

Take error analysis, a fancy name for counting the various types of errors the MT system produces. You might spend five months working out a suitable scoring scheme – is correct gender agreement more important than correct number? – and totting up figures for a suitably large sample of text, but what do those figures mean? If one system produces vastly more errors than another, it is obviously inferior. But suppose they produce different types of error in the same overall numbers: which type of error is worse? Some errors are bound to cost translators more effort to correct, but it requires a lot more work to find out which.

It isn't just users who have trouble with evaluation. Elliott Macklovitch, of Canada, described an evaluation of a large commercial MT system, in which he analysed the error performance of a series of software updates only to find – as the system's suspicious development manager had feared – that not only had there been no significant improvement, but the latest release was worse.

And bugs are still common. Using a 'test suite' of sentences designed to see linguistic weaknesses, researches in Stuttgart found that although one large system could cope happily with various complex verb-translation problems in a relative clause, it fell apart when trying to do exactly the same thing in a main clause. Developers are looking for bigger, better test suites to help to keep such bugs under control.

Good human translators produce good translations; all MT systems produce bad translations. But just what is a good translation? One traditional assessment technique involves a bunch of people scoring translations on various scales for intelligibility ('Does this translation into English make sense as a piece of English?'); accuracy ('Does this piece of English give the same information as the French original?'); style, and so on. However, such assessment is expensive, and designing the scales is something of a black art.

Properly designed and integrated MT systems really ought to enhance the translator's life, but few take this on trust. Of course, they do things differently in Japan. While Europeans are dabbling their toes and most Americans deal only in English, the Japanese have gone in at the deep end. The Tokyo area already sports two or three independent MT training schools where, as the eminent Professor Nagao casually noted in his presentation, activities are functioning with the efficiency of the Toyota production line. We're lucky they're only doing it in Japanese.