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

More advanced machine translation?

Yorick Wilks (yorick@nmsu.edu)
Computing Research Laboratory
New Mexico State University
Box 30001, Las Cruces, NM. 88003, USA.

A distinguished colleague said to me recently that parsing was now solved and we should all direct our attention to language generation. I realised I could only understand, or accept, this remark if I took it to be about the current concerns of English-speaking linguists rather than about the processing of substantial quantities of text of the kind required for general machine translation. I can still see no evidence at all that new grammar formalisms and claims about extensive coverage of linguistic phenomena have actually resulted in large volumes of MT using those techniques, as we are entitled to expect if the claims are true. This regrettable fact may be due only to the shortage of time, but I remember that that argument has now been used by all leading linguistic theories for over twenty years. It is, I believe, because such developments remain as fragile as all their predecessors, while the real advances in MT lie elsewhere and are hard to discern: they are, like much in computer science, social, organizational and engineering matters rather than strictly intellectual ones. The Eurotra system has, in its ten years, recapitulated much of the history of modern linguistics (as ontogeny is said to recapitulate phylogeny in biology) and that has been the source of many of its problems. Again we have, at bottom, a social and organizational, possibly even a management, issue.

This is not a new phenomenon in the history of MT: the American ALPAC report itself was a devastating socio-political instrument, one that halted MT research in the West for a generation, but which said something very different from what many now believe: it certainly did not say MT was impossible.

I take it for granted that advance in MT will come from "phenomena of scale": the use of very large dictionaries in particular, and the extraction from them and from large text samples of collocational, semantic and pragmatic information, as well as new techniques for combining these sources in differing circumstances. It will also require, as it has with historical MT, an understanding of, as well as techniques for, maintaining and adapting very large programs whose original structure has become obscure. I doubt very much that what now passes for syntactic research on English will ever yield the robustness that real MT requires. But I know I am in a minority about this. I am also one who accepts the negative side of the connectionist movement, in that its criticisms of language processing would be almost exactly those above, without having any real faith in its remedies.

It is not that I am a cynic about the notion of theory or formal system, on the contrary, but rather that the history of MT shows, to me at least, the truth of two (barely compatible) principles that could be put crudely as Virtually any theory, no matter how silly, can be the basis of some effective MT and Successful MT systems rarely work with the theory they claim to. Systran is a fine example of both: its real techniques owe a great deal to good software engineering, good software support, and bizarre atheoretic devices that are closer to certain artificial intelligence programs than linguistics. It could be fun assessing existing MT systems against those principles: for commercial systems it is easier to see they obey the first than the second, since they rarely make theoretical claims.

The last refuge of the empiricist is that "Time will tell", though it is harder to know just when it will tell us and (given the second principle above) what conclusions we will be able to draw from substantial, unquestionable, large-scale success in MT when it comes. It is also far from clear that radically differing languages will yield to the same theoretical devices. Another ground for empirical optimism, however, is that the evaluation of MT systems is almost certainly more developed than MT itself.