Evaluation of Machine Translation Systems*

A report by Monique C. Cormier

The International Working Group on Evaluation of Machine Translation Systems, chaired by *Margaret King* of the Dalle Molle Institute of semantic and cognitive studies (University of Geneva) and *Gudrun Magnusdottir* of the University of Gothenburg, brought together some 40 researchers, developers and users, at a forum organised by *Kirsten Falkedal* in Switzerland, from April 21 to 24, 1991.

When judging whether or not a machine translation system is worthwhile, what must be evaluated: the quality of the translations it produces, its performance speed, its cost, the time it requires for post-editing, its user-friendliness or even its potential for development? This first question leads to others: what to use when testing systems: real corpora or fabricated corpora? Would it be useful to elaborate a general evaluation methodology?

Adequacy and potential of the systems

In order to evaluate the quality of the product delivered by the computer, inspiration has been largely drawn, up to now, from the evaluation methods of the celebrated ALPAC report.[1] Intelligibility, fidelity, accuracy of the text, are the criteria generally given to a group of readers as those to use as a basis for evaluating a machine translation; the result is uneven and sometimes irreconcilable assessments, probably due to the subjective and very general character of the chosen criteria.

Now, although not denying the value of the above criteria, there is an increasing movement towards evaluation methods which focus on a limited and specific number of linguistic problems integrated in fabricated or artificial corpora.

Artificial corpora, known as test suites, grouping sentences with a syntactic construction which is representative of a customer's texts, are very useful to the developer who can, thanks to them, check the real strength of his system. How many lexical units must a test suite contain to be considered representative? In answer to this question, many figures have been put forward, from 4,000 to 60,000. It is agreed, however, that the evaluation of a general system requires a larger suite than a system needed to fulfil limited needs. But it is not only a matter of figures and the quality of the suites must never be neglected. If, in all objectivity, the system reaches the end of the tests having solved all the problems it had been presented with, then it is good.

In cases of failure, all is not lost as the developer can try and enhance his system with new rules. Its capacity or its incapacity to accept them will give the developer valuable indications as to the development potential and, therefore, the real strength of his system.

A general evaluation methodology

Ideally, there ought to be a general methodology for evaluating machine translation systems, which could benefit everyone and replace individual evaluations carried out according to one's individual criteria.

In practice, this is not possible for several reasons. First of all, the requirements of potential buyers of systems vary considerably from one to the other, as does the capacity to meet specific requirements. Then, one must bear in mind that an evaluation is very costly for a firm who cannot be certain of obtaining satisfaction.

The general evaluation methodology does not exist any more than the perfect multi-use system. Customers' requirements and constraints vary too much. The true evaluation must therefore concentrate on comparing the results of different systems.

We know now that machine translation is most useful for specific applications, i.e. for carefully defined subdomains where the volume of translation is very large. It is therefore increasingly considered not as an end, but as a means serving the translator. According to this reasoning, what the translators require from the researchers is a machine which is adaptable to them and not vice versa.

The days of comparison of the speed of the machine with the speed of the human being seem well and truly over.

1. American finance organisations who, for ten years, financed research on machine translation and devoted some twenty million dollars to it, decided to review the situation. In 1964, the United States National Academy of Science set up a committee, the Automatic Language Processing Advisory Committee or ALPAC, whose mandate was to assess the results obtained in machine translation and, according to these results, to formulate recommendations as to the follow up of the financial backing of research in this field.

One of the ALPAC conclusions was the following: *(...)we do not have useful machine translation. Further, there is no immediate or predictable prospect of useful machine translation* (ALPAC (1966): *Languages and Machines. Computers in Translation and Linguistics*, Publication 1416, Washington (DC), National Academy of Sciences, p.32). The committee then recommended that research funds be preferably re-oriented towards fundamental research on computational linguistics.

Monique Cormier is a professor in the translation section of the Department of Linguistics and Philology, University of Montreal.

* Translated from the French