[*MT evaluation: basis for future directions*. Proceedings of a workshop...2-3 November 1992, San Diego]

INTERNATIONAL COORDINATION: EC Evaluation Activities

Loll Rolling
Commission of the European Communities

Evaluation criteria must correspond to the wishes of the user community and not to those of academic designers and industrial producers of MT systems.

My contribution today aims at describing the evolution of evaluation philosophy in the last 30 years and the part taken by the European Community. By way of general background, it can be said that the methodology of MT evaluation has gone through three major phases:

(1) Between 1960 and 1980 machine translation wasn't really operational and not at all comparable to human translation. The question was always "To what extent can machine translation be compared to human translation?" The very first criteria that came up were intelligibility, correctness, and style, which are essentially being used to this day. Cost, rapidity, and user satisfaction were not even mentioned during that period.

(2) Then came a new period, 1980 to 1990, when machine translation produced results that could be compared with human translation. The overall usefulness of MT was measured in terms of:

- The cost of raw translation, and not the cost of revised machine translation, which would have included pre- or postediting or the interaction of a human linguist.
- The rapidity of machine translation, which did not mean seconds of computer time but rather total turnaround time, including revision.
- Revisability by translators or subject specialists. Initially, everybody asked translators to revise, but the translators were perhaps not completely objective because they feared for their future employment, so subject specialists were gradually brought in to assess the value of machine translations not only at the Commission but also by other MT developers and users.
- Finally, to measure overall usefulness, the one criterion that we consider now to be the most reliable and efficient—namely, what I call *revision rate*. This is a global measure of quality, cost, and rapidity, and it is used in the following way: We take a sample of 100 translated words and we count the number of corrections that have to be introduced. If there are 12 corrections, we consider that the quality is 88%, corresponding to a revision rate of 12%. Of course, we can improve the revision rate criterion by weighting the types of errors to be corrected and the ease of correcting these errors so that they will not occur again in machine translation.

(3) Today, in the third phase, we are less interested in measuring the quality of machine translation output; we want to assess the progress of the MT system. *Improvability* is the keyword. We can measure the improvement in turnaround by modifying the infrastructure. We can assess user satisfaction and improve user-friendliness through integration of the MT system into an informatics infrastructure that can cope with it more efficiently than at present. We can improve the reliability of a machine translation system by extending its lexical resources, and we can measure improved reliability by counting the number of new items introduced into the lexicons, etc. And we can measure improved quality through periodic benchmarking, which is what we're presently planning to do at the European Commission. We consider that the most effective way to contribute to the usability of machine translation systems is

through disambiguation routines. I doubt that anyone any longer considers evaluating machine translation against a criterion that reflects its conformity with linguistic theory. This, for me at least, is out.

We now use both macro- and micro-evaluation. Macro-evaluation measures the global usefulness of a system for the user. Micro-evaluation is meant to measure the progress of one specific MT system. Again, revision rates comes out as the most useful criterion. For benchmark testing, the text samples must be representative of the texts to be translated; there must be a minimum of 10,000 words and preferably around 50,000; the benchmark corpus must be modular so that it can be adapted to the evolution of subject area coverage; it has to be confidential; and the text must be easily handled by the MT system's infrastructure.

In the European Community a tremendous communication problem arose when the countries got together in the 1950s and decided to create the Coal and Steel Communities and proceeded to hire over 1,000 translators. Clearly there was a need to find alternative ways of overcoming the language barrier. As one might have expected, it was not the translators who took the initiative, but the information specialists, who had developed hundreds of data banks that could be consulted by a minority of potential users only.

Two MT systems were being considered by the Commission in the early 1970s. One was TITUS, developed by the French Textile Institute, and the other was SYSTRAN, developed by Peter Toma in La Jolla. TITUS was a restricted language translation system that required expensive pre-editing (or controlled writing) but supplied excellent translations, while SYSTRAN accepted any subject and any text type, but its translations required cumbersome postediting at the time.

Thus the Commission's involvement in MT started with an unavoidable comparative assessment of TITUS and SYSTRAN, paralleled by a worldwide technology watch covering all multilingual tools, including lexical resources, term banks, and translation systems. All this started back in 1977. Two of the Commission's major contributions were a workshop convened in 1978 for the purpose of selecting the most efficient evaluation methods and criteria (Van Slype 1979), and the SYSTRAN evaluation, which proved the usefulness of these criteria. The workshop brought together 35 experts, including such names as Yorick Wilks, Margaret King, Georges Van Slype, Margaret Masterman, Bernard Vauquois, Wallace Sinaiko (the author of PIMA), Juan Sager, Herbert Bruderer, Frank Knowles, André Petit, Bozena Dostert, and Z.L. Pankowicz.

We started our multilingual action plan by evaluating SYSTRAN and TITUS. This exercise led to our choice of SYSTRAN for use by the European Commission. We purchased a license for the English/French system and immediately proceeded to have an evaluation done by Bureau Marcel Van Dijk in Brussels. A second evaluation was done in 1980, and from 1981 onward we made a regular corpus-based assessment of day-to-day improvements. Ultimately, within the framework of its Multilingual Action Plan for 1977-1990, the Commission was to develop 16 language pairs for SYSTRAN using a large corpus of text for periodic benchmark testing.

In 1984 we did a comparative evaluation of SYSTRAN and LOGOS for language combinations that included German. Although the results favored LOGOS, SYSTRAN offered us better conditions for the improvement of the system than LOGOS did, and thus we decided to develop SYSTRAN for German also.

In 1986 Professor Nagao helped us to make a comparative assessment of Japanese/English MT systems that might be used by the European Commission, and it came out that ATLAS, at that time, was the best system for our application. We now translate about 12,000 abstracts a year from Japanese to English for potential European users.

SYSTRAN generated around 100,000 pages of translation in 1992, and we think this figure will go up to around 300,000 pages in 1993—that is, 30% of the total translation volume of the Commission.

In 1991 the Commission asked an expert group, chaired by Brian Oakley, to make a global assessment of the methods used and the results attained so far. The group approved the methodology we had been using for SYSTRAN but strongly recommended that the Commission extend its evaluation activity to include periodic benchmark testing based on well-defined principles.

Here are the three recommendations of the Oakley Group:

- The Commission should devote a proportion of its investment in machine translation to developing benchmarks and propagating their use. In view of the importance of this topic, say, 5% of its investment seems appropriate.
- There is a role for the Commission to catalyze the creation of a body drawn from the major MT users and system developers of Europe devoted to the development and implementation of a set of benchmarks.
- Because of the mutual interest in the subject, the Commission should extend the proposed work in benchmarking to other continents—in particular, a benchmarking program with, say, DARPA in the United States and MITI in Japan seems appropriate.

The Commission intends to define a permanent corpus, which can be expanded by adding modules, and we are currently deciding on the size of that corpus, its modularity, the tagging mechanisms, the types of text that it should include, and the types of subjects and the languages (or language pairs) that it should cover. The corpus is to be used not only for periodic determination of the revision rate as a measure for improved cost, quality, and rapidity, but also for comparison with other mainframe systems, mainly those that are currently available in Europe (like LOGOS and METAL) and also with PC-based and interactive systems and with human translation. The benchmarking should be incorporated into the development environment, which includes not only the Commission itself but also the contractors who are responsible for development and the production environment. One of the objectives of the benchmark is to estimate the staff, computer time, and overall cost involved in these activities.

Reference

Van Slype, Georges. 1979. *Critical Study of Methods for Evaluating the Quality of Machine Translation: Final Report.* Brussels: Bureau Marcel van Dijk. Copies available from the office of Loll Rolling at the European Commission.