13.3 Evaluation of Machine Translation and Translation Tools

John Hutchins
University of East Anglia, Norfolk, UK

While there is general agreement about the basic features of machine translation (MT) evaluation (as reflected in general introductory texts (Lehrberger & Bourbeau, 1988; Hutchins & Somers, 1992; Arnold et al., 1994), there are no universally accepted and reliable methods and measures, and evaluation methodology has been the subject of much discussion in recent years (e.g., Arnold et al, 1993; Falkedal, 1994; AMTA, 1992).

As in other areas of NLP, three types of evaluation are recognised: adequacy evaluation to determine the fitness of MT systems within a specified operational context; diagnostic evaluation to identify limitations, errors and deficiencies, which may be corrected or improved (by the research team or by the developers); and performance evaluation to assess stages of system development or different technical implementations. Adequacy evaluation is typically performed by potential users and/or purchasers of systems (individuals, companies, or agencies); diagnostic evaluation is the concern mainly of researchers and developers; and performance evaluation may be undertaken by either researchers/developers or by potential users. In the case of production systems there are also assessments of marketability undertaken by or for MT system vendors.

MT evaluations typically include features not present in evaluations of other NLP systems: the quality of the raw (unedited) translations, e.g., intelligibility, accuracy, fidelity, appropriateness of style/register; the usability of facilities for creating and updating dictionaries, for post-editing texts, for controlling input language, for customisation of documents, etc.; the extendibility to new language pairs and/or new subject domains; and cost-benefit comparisons with human translation performance. Adequacy evaluations by potential purchasers usually include the testing of systems with sets of typical documents. But these are necessarily restricted to specific domains, and for diagnostic and performance evaluation there is a need for more generally applicable and objective test suites; these are now under development (King & Falkedal, 1990; Balkan et al., 1994).

Initially, MT evaluation was seen primarily in terms of comparisons of unedited MT output quality and human translations, e.g., the ALPAC evaluations (ALPAC, 1966) and those of the original Logos system (Sinaiko & Klare, 1972; Sinaiko & Klare, 1973). Later, systems were assessed for quality of output and usefulness in operational contexts, e.g., the influential evaluations of Systran by the European Commission (Van Slype, 1982). Subsequently, many potential purchasers have conducted their own comparative
evaluations of systems, often unpublished, and often without the benefit of previous evaluations. Valuable contributions to MT evaluation methodology have been made by Rinsche (1993) in her study for the European Commission, and by the JEIDA committee (Nomura & Isahara, 1992), which proposed evaluation tools for both system developers and potential users---described in more detail in section 13.5. The evaluation exercise by ARPA (White et al., 1994) compared the unedited output of the three ARPA-supported experimental systems (Pangloss, Candide, Lingstat) with the output from 13 production systems from Globalink, PC-Translator, Microtac, Pivot, PAHO, Metal, Socatra XLT, Systran, and Winger. The initial intention to measure the productivity of systems for potential users was abandoned because it introduced too many variables. Evaluation, therefore, has concentrated on the performance of the core MT engines of systems, in comparison with human translations, using measures of adequacy (how well a text fragment conveys the information of the source), fluency (whether the output reads like good English, irrespective of accuracy), and comprehension or informativeness (using SAT-like multiple choice tests covering the whole text).

13.3.1 Future Directions

With the rapid growth in sales of MT software and the increasing availability of MT services over networks there is an urgent need for MT researchers, developers and vendors to agree and implement objective, reliable and publicly acceptable benchmarks, standards and evaluation metrics.

References


