[from: *Survey of the state of the art in human language technology*, ed. Giovanni Battista Varile & Antonio Zampolli (Pisa: Giardini, 1997); pp. 418-419]

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

ALPAC (1966). *Languages and machines: computers in translation and linguistics*, National Academy of Sciences, Washington, D.C. Appendices 9-15.

AMTA (1992). *MT evaluation: basis for future directions*, Washington, D.C. Association for Machine Translation in the Americas.

Arnold D. et al. (1993). Special issue on evaluation of MT systems. *Machine Translation*, 8(1-2): 1-126.

Arnold D. et al. (1994). *Machine translation: an introductory guide*. NCC/Blackwell, Manchester, Oxford.

Balkan, L. et al. (1994). Test suites for natural language processing. *Translating and the Computer*, 16: 51-58.

Falkedal, K. editor (1994). *Proceedings of the Evaluators' Forum, 1991*, Les Rasses, Vaud, Switzerland. ISSCO, Geneva.

Hutchins, W.J. and Somers, H.L. (1992). *An introduction to machine translation*. Academic Press, London.

King, M. and Falkedal, K. (1990). Using test suites in evaluation of MT systems. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 211-216, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

Lehrberger, J. and Bourbeau, L. (1988). *Machine translation: linguistic characteristics of MT systems and general methodology of evaluation*. John Benjamins, Amsterdam, Philadelphia.

Nomura, H. and Isahara, H. (1992). JEIDA's criteria on machine translation evaluation. In *Proceedings of the International Symposium on Natural Language Understanding and AI*, Kyushu Institute of Technology, Iizuka, Japan.

Rinsche, A. (1993). Evaluationsverfahren für maschinelle Übersetzungssysteme: zur Methodik und experimentellen Praxis. Technical report, Kommission der Europäischen Gemeinschaften, Bericht EUR 14766 DE.

Sinaiko, H.W. and Klare, G.R. (1972). Further experiments in language translation: readability of computer translations. *ITL*, 15:1-29.

Sinaiko, H.W. and Klare, G.R. (1973). Further experiments in language translation: a second evaluation of the readability of computer translations. *ITL*, 19: 29-52.

Van Slype, G. (1982). Conception d'une méthodologie générale d'évaluation de la traduction automatique. *Multilingua*, 1(4): 221-237.

White, J.S., et al. (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Technology partnerships for crossing the language barrier: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, pages 193-205, Washington, D.C. Association for Machine Translation in the Americas.