"The Evaluation of Machine Translation"

Talk given by Ms. Siety Meijer of the MT Evaluation Group, Dept. of Language and Linguistics, University of Essex, on 20 May 1993 at King's College, London.

Ms. Meijer began by explaining that the MT Evaluation Group had been set up by the Computational Linguistics/Machine Translation (CL/MT) Group within the Department of Language and Linguistics, to satisfy a perceived need for a UK centre of excellence for MT evaluation. They are also prepared to evaluate systems either for developers or potential users.

The Group have carried out a number of evaluations, including some for the EUROTRA system and one of the Globalink (GTS) system, and have presented papers at a number of MT conferences. They are also currently preparing a book about MT.

Ms. Meijer identified eight types of evaluation of MT, namely, End User, Developer, Comparative, Single, Snapshot, Cyclic, General, and Application Specific.

Drawing an analogy with football teams she thought the evaluation of the "best" was liable to be difficult and subjective, especially since so many different factors need to be taken into account.

Some of these factors are,

- Hardware required
- Software - Word processor facilities
- Speed of translation
- Quality
- Vocabulary
- Robustness

But this was not enough. These factors needed to be related to user specific questions like, "How does the MT system perform on user specific text types?", "How serious are the errors in the output text, and how much work is required to correct them?", "What vocabulary is covered by the system?" and "How complicated is it to update the dictionary?". The user will also want to keep in mind the organizational changes that would be required in his company, and the solvency of the MT vendor.

The most difficult aspect was obtaining an objective evaluation of the quality of translation. Five of the most common evaluation methods were then described, most of which tried to establish a quality measure in some way or other. In the ALPAC review 1964-1966, machine translations, and some human translations, were scored for Intelligibility, Fidelity and Accuracy, and Style.

For this type of method, clearly defined scales have to be established for evaluators. The following very simplified example scale for intelligibility consists of four points,

1 clear and intelligible
2 easily understood
3 general idea understood
4 unintelligible

The results of this type of evaluation are to some extent subjective and difficult to relate to the cost-effectiveness of an MT system.

Error analysis is a more objective way of assessing translation quality. It involves an error count, extended by a weighting of each error. This weighting is normally related to the number of alterations and the time required for post-editing, which are, of course, of prime interest to potential users. The method is labour intensive.

Test Suites of specially designed sentences are sometimes used to check how well systems handle particular language phenomena. This method is favoured by NLP system developers who deal with rule-based systems which include linguistic knowledge. It allows testing of the system's rules, individually and in combination with others. The University of Essex is starting up a Linguistic Research and Engineering program (LRE).

The best method for potential users is probably a full operational evaluation, which would allow testing the cost-effectiveness of a system within the end user environment. Evaluating a system according to this method will take several months at least, especially since staff have to be properly trained and the system tuned to user specific text types and vocabulary.

Ms. Meijer then concluded by referring to a rather new method which involved automatically measuring the 'distance between translations' which required the machine translation to be compared with a number of human translations of the same text to see how much it differed from them on average.

Editor's note: The book mentioned in Siety's talk has now been published and is listed in the Books List. She also kindly made available a copy of the well known 31 page "HP-NL Test Suite" published by the Hewlett-Packard Laboratories in 1987. I can supply copies at cost of copying and postage for £2.50 within the UK.