PANEL: THE ROLE OF MT EVALUATION

Evaluation of MT is an old topic, yet never ceases to stir up controversy. In light of the ongoing ARPA MT Evaluation series, a panel involving some of the most experienced voices in MT evaluation, some of the most eager evaluators, and some real-world users of MT, is a perfect forum in which to discuss how (if at all) MT evaluation contributes to the ongoing practice of the field. And here a central burning question is: how can one ensure that MT evaluation is of practical benefit and does not simply become an empty exercise in itself?

Moderator: Maghi King, 1SSCO University of Geneva
Panelists: Scott Bennett, Logos Corp.
George Doddington, ARPA
Mary Flanagan, CompuServe
Laurie Gerber, SYSTRAN
Marjorie Leon, Pan American Health Organization
John White, PRC Inc.

The Role of Evaluation in Machine Translation
Winfield Scott Bennett, Logos Corporation, Mt. Arlington, NJ

While machine translation is hardly a new technology, those of us in the discipline have never truly come to grips with the issue of evaluating our systems. This is equally true for commercial systems as for those under development in laboratories. In recent years the increasing interest in MT has underscored the lack of evaluation criteria or any sort of standardized approach to assessing our systems. Internal evaluation of a particular system is generally on an idiosyncratic basis, valuable to the particular group, but not useful to those outside the circle of developers with that system. External evaluation, especially between systems, has likewise tended to be specific for particular purposes (e.g., the evaluation of the ARPA-sponsored efforts by PRC).

The difficulty, of course, is that there are no real criteria for evaluating human translation. This situation has meant that we in machine translation cannot fall back on translation as a whole to give us guidance in the matter. Further complicating the issue is the reality that machine translation is a unique kind of translation, often requiring a change in method of operation for the user.

From a commercial standpoint an MT system has to be evaluated on the basis of its ability to "get the job done", normally measured in terms of just how much it contributes to reducing the corporations time-to-market with a localized product. Turning such economic criteria into an evaluation metric is not a trivial task. The needs of individual (corporate) users may vary considerably. For example, a user may accept raw output which is below the standards of the development group, if the results may be post-edited quickly and easily.

The result of this is that we in commercial machine translation must ask our users and potential users to aid us in creating the sorts of evaluation criteria which truly reflect their needs. Evaluation

done on the basis of development criteria is ultimately pointless from a commercial standpoint.

Questions for Evaluation

Mary Flanagan, CompuServe, Cambridge, MA

There is no single, correct method of MT evaluation. To be of practical benefit, MT evaluations must be customized to reflect the needs of users. Many evaluations of MT focus on output quality. Yet to successfully integrate an MT system into a translation environment, many other questions must be answered:

- Who will see the translations? What level of quality is acceptable?
- Who will use and maintain the system? What skills will they need?
- How much postediting will be done? Who will do it?
- What hardware platforms can we support?
- What speed of translation is needed?
- What are the costs for the system, training, upgrades, maintenance and staffing?
- What level of support can the vendor provide for questions, training, upgrades?
- Has the system been used in a similar environment to ours? What problems were encountered?

Public interest in MT is growing. The MT community should work together to develop a set of evaluation tools that is adaptable to different user environments, and that addresses each of the above issues.

The Rather-Larger-than-Expected Utility of Black-Box MT Evaluation

John White, PRC Inc., McLean, VA

Until the ARPA MT Evaluation took shape in the last couple of years, I would have considered any attempt to evaluate machine translation to be purely driven by the design of the system or by its end-application. I assumed that this must be the way that developers could evaluate their progress. A semo-syntactic transfer system evaluation, for example, would track the proper firing of rules that govern the source and target, and ensure that the characterization of language-pair contrasts actually cover the phenomena presented to it. And what are the phenomena presented to it? Well, I assumed that it was mostly grammatical patterns that were known to be problems in translation (whether in analysis, transfer, or generation), with enough "real" domain text thrown in to make the sponsor feel like the developers are paying attention to their needs.

So my presumption was that evaluation looked deeply inside the box, and that it focused on diagnostic patterns to determine progress or regression. So far, so good: I imagine that any one system will employ some of these techniques on a day-to-day basis. But the presumptions underlying the approach ultimately poison our minds. We interpret language phenomena in terms of a particular MT approach with which we are familiar. We confuse description and explanation. We

ask how other people's systems handle, say, Subject Raising without even realizing how deeply myopic we have become in the presumptions of particular linguistic theories, and in the presumptions of linguistics itself.

The profoundly black-box evaluation approach in the ARPA initiative is necessitated by the well known differences among the systems being evaluated. Again, so far, so good: we can maintain the myopic perspective for internal evaluation, and the black-box for cross-system comparisons. But I believe that I am coming around to the even stronger position that black-box should become more of the evaluation picture, and glass-box less, even in internal evaluations. I have come to believe that the functional perspective driven by black-box sets a clearer path for improvement in terms of the ultimate use of MT systems. Even beyond this, though, the tests derived from black-box approaches invariably reveal weaknesses in phenomena that are simply not thought of when developing glass-box diagnostics.

My contribution to the panel, then, is the assertion that the best way to prevent MT Evaluation from becoming Empty Evaluation is to continue to abstract the particular theories, approaches, and designs from the process and strive for the blackest of possible boxes.