[MT Summit IV, July 20-22, 1993, Kobe, Japan]

Panel Discussion: "Evaluation Method of Machine Translation"

Muriel Vasconcellos
AMTA, 655 Fifteenth Street, N.W., Suite 310, Washington, D.C. 20005
Fax: (202) 667-8808

1. Experience in MT Evaluation

When it comes to credentials in MT evaluation, I have earned my stripes mainly as a frustrated observer of the process. I have watched MT evaluations from ALPAC to DARPA. As a hands-on user of MT for more than 13 years, I have seen and thought about the many forces and factors that come together to make MT effective. And I have learned how difficult they are to measure, especially as they combine in countless different ways. It has always worried me to see hard-and-fast conclusions, sometimes sharply at odds with day-to-day experience, being drawn from isolated fragments of the picture, much as the apocryphal blind men felt different parts of the camel and made guesses about the whole animal that were widely off the mark.

I am also a certified evaluee. During my watch, the MT project at the Pan American Health Organization was subjected to six major studies. In the early years, three separate progress evaluations were done: by Wilks in 1978,by Zarechnak in 1981, and by Macdonald, also in 1981. In 1987, after seven years of practical operation, I was responsible for designing and implementing a controlled 11-month study that focused on cost-effectiveness and user satisfaction (Vasconcellos 1989). In 1989, the English-Spanish system, Engspan, was benchmarked against four others in a massive test conducted by McGraw-Hill (Benton 1989). And most recently, PAHO's Spanish-English system, Spanam, was one of the five to be tested in the DARPA exercises of 1992 (White et al. 1992) and is again participating in 1993.

2. The Cook and Her Batterie de Cuisine

Certainly one can agree that it is "unreasonable to look for common, or simple, evaluation techniques," There should be as many different approaches to evaluation as there are reasons for undertaking it. Each evaluation should be tailored to the purpose for which it is being conducted. Moreover, it must be tied to the purpose for which the particular system was developed.

It must then take into account (in varying degrees) the dynamic environment in which the system operates. A one-time performance tells very little. Myriad factors can affect the final product, and the roles they play must be fully understood. Even a diagnostic evaluation or an assessment of linguistic progress must go beyond the bounds of the glass box and take into account elements of the environment such as the purpose of the project and the translation being produced, the origins of the text being translated, the human team that is involved with the system, the physical platform, and the flow of funding and other resources available to the project.

Given the range and complexity of the factors to be considered, it is of course essential that the "cook" be thoroughly trained. The question that remains is whether or not there is yet, or will be in the near future, a definable "batterie de cuisine" that she can use for the task. For the diagnosis of breakdowns and the evaluation of progress in a system's development, the tools tend to be quantifiable and may in fact be largely known, if not fully developed. However, for the evaluation of adequacy, the situation may be too complex to allow an evaluator to simply pick out a combination of approaches from Column A and Column B. Moreover, the current state of the art leaves a lot to be desired. The existing instruments, which focus on static snapshots of system output, yield results which may have questionable meaning in terms of the life of a system. They may even be misleading - either by calling undue attention to an easily fixable problem, or by giving a false sense of security when important problems are never tested (as, for example, with a small prototype that has a minimally developed lexicon and requires relatively little branching in the decision-tree).

3. Categories

The three-category breakdown is very useful because it should keep the evaluation in perspective and help to specify the most appropriate tools. The first two categories are fairly discrete and manageable. The third, however, is complex and slippery.

Valuable as they are, it is important not to see these categories as hermetic. It's never easy to slice off a particular task and limit the evaluation to this or that. There needs to be an understanding of the larger picture. No evaluation can be completely informative without a look inside the glass box, and, on the other hand, the view inside the glass box is never the whole story. Evaluations for adequacy tend to be black box: they may take one-time snapshots, look at how well the system as a whole has met its intended purpose, or, better yet, do both. But they often fail to take the important additional step of looking inside and assessing the relative importance of the ways in which the system failed. This knowledge is crucial to knowing whether or not it can be relied on to produce consistent results, whether it can be scaled up to handle more vocabulary and a broader range of text types, and whether it can be extended to other domains.

4. Shared Resources

From the overall perspective of progress in machine translation, some kinds of sharing would seem to be more productive than others. Two areas that should have very high priority are lexicons and test suites for parsers. Lexicons should be shared to the greatest extent possible in order to reduce duplication of effort and free up scarce resources for more creative activities. Since centralization is obviously the key, the most effective approach would be through the continued pooling of existing resources and the steady incorporation of new initiatives. As far as test suites for parsers are concerned, sharing them may open the way to eventually setting standards that systems could voluntarily try to meet. On the other hand, large corpora of matched translations are difficult to obtain and may have limited application. They are essential, of course, for the development of corpus-based MT. However, progress will always be trained within a narrow niche if research is focused in the areas for which such corpora are provided. It is unlikely that the corpora available will match up exactly with the areas in which MT development is truly needed. Too much focus on shared corpora could result in research getting stale.

5. Free Topic

All too often a black box evaluation unfairly penalizes a system for a minor linguistic problem that could easily be fixed or, contrariwise, overlooks a major limitation. Each little "failure" tells something about the system and provides a valuable clue to the evaluator. Regardless of its purpose, an evaluation should be able to get to the cause of linguistic shortcomings and weigh their seriousness in terms of overall system function before drawing hasty conclusions. Mistakes can be caused by a glitch in the front end interface, poorly formatted input, ill-formed input, input that is inappropriate for the particular system, incorrect coding of the lexicon, protective overcoding or undercoding, too many rules, not enough rules, the wrong kind of rules, and so on. The architecture may allow for growth an improvement, or it may have severe inherent limitations. Some mistakes may not be worth fixing at all, depending on the purpose of the translation project. Only a trained evaluator can make these judgments.

References

Benton, Peter M. 1989. A practical test of machine assisted translation systems and agencies. In: Proc 30th Annual Conference American Translators Association, ed. Deanna L. Hammond. Medford (NJ): Learned Information.

Vasconcellos, Muriel. 1989. Long-term data for an MT policy. Literary and Linguistic Computing 4(3):203-213.

White, John S., Theresa O'Connell, and Lynn Carlson. 1992. Report of the DARPA MT Program system evaluation. Washington, DC: DARPA Software and Intelligence Systems Technology Office.