A Workshop on Evaluation: Background Paper

Margaret King
ISSCO and ETI
University of Geneva
Switzerland.

The title of this session is meant to suggest that it will be concerned not with a formal presentation of any kind but with an attempt to launch a communal reflection on what is and what should be involved in evaluating a machine translation system.

The session will be organised around a number of questions, some of which are briefly outlined here. No serious attempt is made to pre-empt the discussion by providing detailed replies, although some starting points are suggested in order to provoke reactions and further thought. It is hoped that participants in the session will contribute further questions, and will prepare by thinking about possible answers.

Before turning to the more specific questions, it is perhaps worth putting forward a preliminary answer to the logically a priori question of why we, the designers and constructors of machine translation systems, should worry about how systems should be evaluated. Except on those few occasions when we are called in to evaluate one another's work, we are, after all, more likely to be the judged than the judges.

The problem is that there is no established methodology for evaluation: the majority of evaluations are done under contract and often under a confidentiality agreement, so that relatively little constructive criticism of the techniques employed is in the public domain. Even where an evaluation report does include a study of previous evaluations, it is usually in order to conclude that the methodology adopted or techniques used are inadequate or inappropriate to the current context. Thus, in essence, every new evaluator invents his own methodology, and those to whom he reports have no means of comparing his standards or checking his judgements against any common standard.

Even worse, many evaluations are done by people who have no expertise in machine translation techniques, for example by the translation service within which the system is to be installed or by the marketing department which is looking for a hot product. No matter how thorough or how honest an evaluator with such a background may be, he necessarily remains a lay-man with respect to machine translation technology, lacking a solid knowledge of what is possible and what is far-fetched, unable to

estimate the potential of a system with any accuracy and lacking the intimate knowledge of the relation between linguistic description and computing requirements which will allow him to estimate a system's potential efficiency. He may well be able to assess a system's current performance (which is typically rather poor when the system is first applied for some particular purpose) but lack the competence to assess more important criteria, such as the ease with which the system can be modified to eliminate errors or its extensibility.

All this comes down to saying that if we do not think about evaluation methodologies, no-one else will, at least not in public, and we remain vulnerable to partial or inaccurate evaluation of our work. Of course, there is a fine flavour of paranoia about this, but we need only think of one or two notorious evaluations (ALPAC and the TAUM-Aviation evaluations are the best known) to decide that perhaps we really are in some danger of being persecuted.

On the more positive side, public discussion of evaluation methodologies by informed parties could do much to keep us honest: if the public becomes more knowledgeable, it will be less easy to get away with demonstrating twenty six carefully chosen sentences, or a vocabulary of five hundred or so words, and claiming that one has the basis of a large-scale multi-purpose system. Public discussion might do much, too, to ensure that the expectations of the lay world become realistic. We are all familiar with the outsiders who cannot convince themselves that there is anything difficult or mysterious about language, since even young children show a remarkable ability in language use, and therefore believe that simultaneous interpretation by telephone is just round the corner, just as we are familiar with the other extreme, often, sadly, represented by professional translators, who believe that because automatic translation of Shakespeare is not feasible, we might as well give up and go and do something else. On a more basely material note, it is worth reminding ourselves, too, that Alpac was partly the fault of the professional machine translation community, who had actively encouraged their sponsors to have unrealistic expectations.

With the case for discussion made, I hope, let me describe briefly some of the issues to be discussed.

The basic question can be stated thus:

Is it possible to define a general evaluation methodology which will be *fair, reliable* and

- *applicable* to any kind of system
- *informative* for all evaluation purposes
  (i.e. for satisfying research sponsors, convincing potential purchasers, for demonstrating practical benefits in actual use ...)

One's first instinct is to reply no. Let us look at why.

First, there is the question of generality. For example, it is tempting to claim that there is a fundamental difference between evaluating a research prototype, intended to show the feasibility of some particular approach, and evaluating a commercial product, intended to answer some specific need in some specific context. We should try to work out whether this is really true, and if it is, in what the difference consists. A potential customer for a commercial product is likely, if he is prudent, to start by asking himself what his particular needs are and what constraints are imposed by the context in which the system will be installed. Is this really very different from a research sponsor agreeing to fund a project aimed at a particular type of system running in a particular computing environment, and, at the end of the project period asking himself if he has really got what he agreed to pay for?

It is sometimes claimed that computing time and efficiency is much more critical in the case of a commercial product. But anybody who has worked with a research prototype which takes two hours to translate a sentence may well think that efficiency is at least relevant there too.

Another plausible difference, from the intellectual side, might be the lexicon to be expected. Intuitively, at least, one would expect a research prototype to have lexical material covering a wide range of linguistic phenomena but not necessarily a great volume of essentially similar lexical items, whilst a commercial system would have to deal with volume. This remains plausible as long as one is thinking of showing the syntactic coverage of the source language, but once one thinks of the variety of translational problems posed by lexical material, it is hard to set a very clear borderline between what could be expected of a research prototype and what of a commercial product. Another way of stating this would be to ask how it can be shown that a prototype will not fall apart when scaled up to realistic size, except by actually doing the scaling up.

Perhaps the only difference which is clearly justifiable is that a commercial product must be demonstrably time or money saving compared to what was done before the system was installed, whilst such a consideration simply does not apply in a pure research context - although it well might if the research is undertaken with the eventual aim of producing an operational system.

To recapitulate the question: what is at issue here is: can there be a general methodology applicable in all contexts?

A second rather general issue to be tackled is the difference between different kinds of evaluation, for example between what in the AI community is called glass box evaluation and black box evaluation. In the former it is assumed that the evaluator has access to all the inner workings of the system and can inspect

intermediate results. In the latter he has only input/output(s) pairs to work with. Both kinds of evaluation present their own problems, but black-box evaluation is obviously more uncertain, especially when the system's potential for extension is being considered. Although this is a problem for all natural language processing systems, in the case of a machine translation system it is rendered even more delicate first by the subjectivity of judgements of the output, and secondly by the variety of different system components which can give rise, independently or in interaction, to an unsatisfactory output. The particular question here is whether any techniques can be found to make black-box evaluation less opaque, or, alternatively, to identify evaluation contexts where one type of evaluation rather than the other is more appropriate. It might also be worth asking whether other types of evaluation can be identified, together with contexts in which they would be appropriate.

Mention of the subjectivity involved in judging the acceptability of a translation brings us fairly naturally to thinking about what the criteria are on which a system should be evaluated. One tradition concentrates on the quality of the translations produced, relying almost entirely on subjective judgement, and talking about characteristics like fidelity, intelligibility, style. The dangers here are obvious and have often been commented on.

A more recent school, influenced by work elsewhere in computational linguistics, has tried to set up test suites of critical inputs to serve as a set of benchmark tests. Doing this in order to test source language coverage, whilst lengthy and delicate, may turn out to be feasible. When one thinks of constructing a test suite to test translational behaviour, the job sounds much more daunting. And, in either case, the size of the test suite is likely to become a major hindrance to applying it. (In the extreme case, one could finish up with a test suite which took longer to administer than the system had taken to create).

Yet another school has suggested that the only feasible criterion *is* the total length of time taken to produce an acceptable translation, usually measured by submitting comparatively large quantities of "real" text to the system, as opposed to constructing artificial test materials, the argument being that real text will, in the end, contain all the problems that the constructor of test material might think of, and then some more. An obvious objection to proceeding in this way is that the system is only tested relative to a particular set of texts, and that its behaviour might be quite different if a different set of texts were chosen. This can be an advantage if the customer's/sponsor's needs are paramount. One way round it, if generality is required, might be to set up a canonical text corpus containing a variety of different kinds of texts. But then the risk of unwieldy bulk surfaces again.

A different kind of problem with criteria based on throughput time is that the time taken is heavily influenced by the experience and attitude of those interacting with the system (pre-editors, post-editors or inter-actors), and by the level of perfection they are aiming at. For example, the amount of work invested is likely to be different if the translation is intended to go straight to those who want to read it (and may vary according to their imagined purpose in reading it), to a revisor or to a system developer.

So far, no mention has been made of criteria and techniques based on counting errors or on classifying errors in running text. The basic weakness of any judgement based on error analysis can be reduced to the simple question of defining precisely what is to count as an error and how it should be classified. Nonetheless, the kinds of errors produced can be critical in deciding whether a system is reparable. Is there some classification, even if not a very refined one, which can be clearly defined and used?

A number of people, including the present author, have tried to work out a methodology combining the positive aspects of many different techniques mentioned above. The problem here is that the testing designed to provide the data on which the assessment of the system will be based becomes very lengthy and impossibly expensive to administer. This becomes even worse if measures are included to try to diminish bias in those administering the tests. Can we find some way of producing realistic tests that will retain the intellectual virtues of analytic tests such as test suites or error analysis and the practical virtues of testing based on real text?

One problem with all these testing techniques is that they relate to the current performance of the system. They have to be supplemented at least by an up-date and re-test cycle if any assessment of reparability or extensibility is required. A further question then is whether one kind of testing is more informative than another about the system's potential?

Many other issues have not even been touched on here. They will, one hopes, be brought out in discussion. Even if we cannot hope to solve the problems, just spelling them out will be *a* useful consciousness raising exercise.

Acknowledgement.

My thanks should go, as always, to Kirsten Falkedal for much fruitful discussion of evaluation and its problems. Her careful reading of a first draft of this paper did much to improve it; the remaining faults are of course my own.