[*MT Summit III*, Washington, DC, USA, July 1991]

Evaluation of MT Systems
Panel Discussion

Chair: Margaret King, University of Geneva, Switzerland
Yorick Wilks, New Mexico State University, USA
Sture Allen, University of Gothenburg, Sweden
Ulrich Heid, University of Stuttgart, Germany
Doris Albisser, Union Bank of Switzerland

1 Introduction, Margaret King

Let us start from the assumption that just as there can be no single general purpose machine translation system, so there can be no single general purpose evaluation methodology. The factors which will determine what counts as a 'good' system are many and various, and their interactions complex. In what follows, I have tried to disentangle some of these factors and explain them briefly. Subsequently, the panelists will be asked to suggest what set of factors are relevant to a particular kind of customer for evaluation, and to outline what choices they would make if they were that customer.

1.1 Customers for evaluation

Let us take a fairly arbitrary number and identify five typical customers for evaluation: people, or groups of people, who might want to carry out an evaluation themselves or to commission an evaluation to be done for them. In what might be called a loose chronological order these are:

- Research workers, looking for new ways to design and implement machine translation systems. Typically, they will want to test their ideas through the construction of a research prototype, and will be interested in evaluating either their progress, or, if the prototype is nearing completion, how well it has in practice validated the ideas on which it was based.
- Research sponsors, investing in research on machine translation in the pre-competitive stage, but presumably in the hope that the research will eventually provide the basis for an operational system. (It is sadly hard to find funding for research in the pursuit of pure knowledge these days). Their mam aim in carrying out an evaluation will be to discover whether they should continue funding a project, or perhaps to decide between candidate projects competing for continued funding.
- Commercial enterprises who are trying to decide whether it is worthwhile acquiring an embryonic machine translation system, perhaps a research prototype, perhaps a system that some other enterprise is ready to give up on, develop it into a commercial system and market it. Assuming that they are already convinced that the market exists, their primary concern will be with the potential for development of the embryonic system, and with what further work will be needed to make it commercially attractive.
- System developers, who are already committed to developing a particular system to the point where it becomes commercially viable. Their primary concern in evaluation will be to monitor their own progress towards commercial viability, and perhaps, to assess what additional features must be added to increase the product's chances of success.
- Potential customers of commercial systems, who may range from individual translators or translation agencies to firms, large or small, confronted with a particular translation problem or to governmental agencies of one type or another. Their primary interest in evaluation will be in order to assess whether a particular system being evaluated meets their needs, whether it does so better than its competitors, and, perhaps, whether it is at all worthwhile to introduce a machine translation system.

It is intuitively fairly clear that these different typical customers for evaluation have different needs and different concerns. A first question for our panelists is to ask them to sketch a little more in detail what these different needs might be.

It should also be clear that already at this stage matters have been considerably over-simplified to take account of time constraints and in order not to clutter too much the general picture. Other kinds of customers for evaluation might be imagined, for example someone faced with a multi-lingual documentation problem or someone wanting to teach a course which included hands-on experience of one or more working systems. And nothing has been said about the potential relevance of different kinds of systems: a commercial enterprise wanting to develop from an embryonic system might well, for example, be interested in the embryo's potential as an interactive system, or as part of an authoring system as opposed to a batch system.

1.2 Constraints on evaluations

But if we can, perhaps, at least for the moment and for the sake of concentrating on broad issues, neglect this set of questions, we cannot neglect the fact that the circumstances in which the evaluation is to be done can impose quite severe constraints on the kind of evaluation which can be envisaged. First, whether or not the evaluator can inspect the inner workings of the system or whether he is limited to inspection of input/output pairs will have strong consequences. In the latter case, for example, it is impossible to found a judgement of the system's extensibility on an assessment of the linguistic or computational techniques employed. More generally, the degree to which the manufacturer, constructor or designer of the system is prepared to collaborate with the evaluation can severely affect the design of the evaluation; it is no use, for example, including a test of how long it takes to code fifty new verb entries if the manufacturer will not allow users to enter verbs. At a more banal level, the amount of time and money which can be invested in the evaluation is of great importance. (The present author was once guilty of coming up with an evaluation scheme which would have cost more to administer than the system had cost to develop).

Once these and other constraints (again, matters have been considerably over-simplified here) have been determined, the evaluator can proceed to ask himself what it is he wants to evaluate. This question really breaks down into two sub-questions.

1.3 What is being evaluated?

First he needs to know in general terms what type of evaluation is required. Is the system output, for example, to be assessed relative to various dimensions of translation quality, like fidelity to the original, intelligibility, or stylistic facility? Or is it to be assessed relative to some pre-defined set of desirable criteria, like a description of certain linguistic phenomena of the source language which must be treated adequately or its ability to deal with a specific sub-language in which, for example, imperative verbs in an English technical manual should be translated as infinitives in French? Or is it to be assessed in terms of global coverage of a particular language pair or set of language pairs? Or, more prosaically, is the only important question one of a cost/benefit analysis in a particular context?

Secondly, he needs to determine the evolutionary stage the system is assumed to be at, i.e. whether it is to be considered as a finished product, with no further modification envisaged, or whether he is trying to assess its value as a starting point for modification to fit it to a particular use, or whether it should be considered as a prototype, intended only as a demonstration of the feasibility of the techniques employed, with all the real development work yet to come.

The answers to these questions will determine the broad framework of the evaluation. (With, as always, the caveat about over-simplification). Within the broad framework, the evaluator will have to decide which out of quite a wide range of data collection techniques will provide him with the appropriate data on which to base a judgement.

1.4 Techniques for data collection

Some of these techniques could be called classic; they have been used in a number of different evaluations, and their strengths and weaknesses have been discussed in the literature. Of these, the best known is probably the use of scales; test subjects, drawn from some suitable population, are asked to rate outputs on a scale according to such factors as their intelligibility, fidelity or so on.

The greatest weakness of such tests is their subjectivity; different subjects can vary widely in their ratings, and even the organization of the scales themselves can affect the results. In an attempt to counteract this, there has been considerable recent discussion of the use of test suites, pre-defined lists of linguistic and translational phenomena to be used as bench-mark tests against which individual systems can be measured. The problem here is that construction of such test suites is a complex and very time consuming job, especially when translational phenomena are to be taken into account, and their administration loo can be very time-consuming. Use of a general purpose test suite may also prove to be disappointingly unilluminating when a system's suitability in a particular work context is in question: tailoring a test suite to fit the context or constructing one specially can again be unacceptably labour-intensive.

An alternative is to use “real text”: test corpora either drawn from a body of text representing the specific needs in a particular context or more general corpora drawn from the text collections which are becoming more widely available. The difficulties here tend to be ensuring that enough representative text in machine readable form is available to provide valid results, and the amount of time required to assess the results. There is also, of course, the question of how the results are to be assessed. Favorite techniques include measuring post-editing time or total throughput time. The amount of time and money that can be committed to the evaluation again becomes relevant at this point.

Another classic technique is to try to analyze and classify the errors found in unrevised output. Essentially, two dimensions of classification are possible. The first is to classify according to some system independent scheme, identifying, for example, morphological errors, syntax errors and semantic errors. Alternatively, the classification scheme can be based on what is known of the system, either in terms of the relative difficulty of repairing the mistake (typically, here, errors due to dictionary coding are considered less serious than other errors), or by trying to identify what component of the system is responsible for the fault.

In both cases, the major difficulty comes first in trying to decide what counts as an error and subsequently in disentangling the interactions of the various system components in order to decide what kind of error is in question. To illustrate this, the would-be French sentence *Je l'ai vues* is certainly an incorrect translation of "I saw her," since the French past participle should agree in number and gender with a pre-posed pronominal direct object. But the error could be due to incorrect dictionary coding of the pronoun, either in English or French, to poor functioning of the French morphology component, which has resulted in the plural 's' being incorrectly added, to a fault in a grammar rule which has failed to insist on the correct agreement of object and participle, or even to a wrong treatment of plurality at a semantic level.

In certain circumstances, exhaustive testing of a system's internal behavior may be envisaged, for example by identifying individual grammar rules and the possible interactions between them. Clearly this is only possible when the evaluator has access to the internal workings of the system. Furthermore, systematically testing all possible interactions between rules may lead to a combinatorial explosion of tests, and the testing itself only aims at the present state of the system, saying little about its capacity for extension.

As always, the above should only be taken as indicative; other possibilities for data collection techniques can and have been suggested. For example, Henry Thompson has recently suggested using statistical methods to assess the degree to which a particular machine translation output differs from a set of translations of the same input produced by a number of human translators, and Hans Karlgren once suggested that the ease with which a translation could be read aloud provided a good indicator of intelligibility and stylistic felicity.

1.5 Envoi and an Acknowledgement

The foregoing has tried to set a framework within which the individual panelists can be asked to put flesh on abstract bones by describing what their choices would be in a given situation. Whilst writing, I have been conscious of a very great debt to the participants in the Evaluators' Forum in Les Rasses in April of 1991; many of the questions raised here were discussed there, and the discussion served to bring order to my own mind. (I am of course solely responsible for misconceptions, misrepresentations and otherwise general stupidity). It would be impossible to acknowledge each individual's contribution to my discussion of particular points. I hope to compensate for this by having a summary of the discussions available at the panel.

Similar considerations have prevented me from trying to give exhaustive references. As a way into the literature, good bibliographies can be found in the references cited below:

[1] Lehrberger, J. and Bourbeau, L. Machine Translation. Linguistic characteristics of MT systems and general methodology of evaluation. John Benjamins Publishing Company, 1988.

[2] Falkedal, K. Evaluation Methods for Machine Translation Systems: An Historical Overview and a Critical Account. Report to Suissetra, 1991. Available from ISSCO.

•

The panelists were asked to identify themselves with one of the five types of customer distinguished in the introductory text, and to suggest what some of the main issues in that particular context might be. The audience is invited to do the same, and to contribute to the discussion by making explicit their agreement or disagreement with the panelists.

2 Yorick Wilks's Statement

Determining the development potential of machine translation systems: the role of evaluation techniques

I have an uneasy feeling that I am the only panel member here who is not playing by the rules. That is to say, all my instincts tell me to reject the assumption underlying the rubric of this panel: Namely, that, since we all know that there are many types of evaluation techniques, and that there are many sorts of translation needs (which is true enough), we should therefore expect a correlation between needs and evaluation methods. Indeed, until I began to refuse to play in the proper way with the other children, my assigned task had been to examine the evaluation needs of research workers. I am also painfully aware right now of the fact that having spent less time thinking about evaluation issues than have my panel colleagues, and certainly less than the panel chair, I am not qualified to put forward such a criticism. But, in any event, I shall follow my instincts.

First of all, it seems to me that research workers (RW's), (and system's developers for that matter), are not working in the vacuum that some imagine them to be; they already operate with particular languages in mind, as well as particular text types and domains. And they will almost certainly have chosen initial texts already. If they haven't done so, then they are probably not serious about their task. If I am right about this, then it means that they are not so different from the other types of customers we are being asked to consider (such as research sponsors).

Our RW group may also have a basic list of grammatical structures to be tackled, but they may not, depending on their methodology. That is to say, proponents of MT based solely on statistics (i.e. derived from very large bilingual corpora) will certainly not have any such list in their possession, and would in fact dismiss the value of any such. They would also be lighthearted about fine distinctions between "seeing into the workings of the MT system" or not, since their only workings are huge trigram tables and you are welcome to look at them for all the good it will do you. So it turns out that some of these issues are highly dependent on MT methodology, after all.

Secondly, I tend to be quite sceptical about special "test suites". To me, they seem to be part of the unfortunate linguistics legacy of MT; and very little functional MT owes anything to linguistics as practiced in university linguistics departments, (By this I do not mean statistics-based MT; but SYSTRAN, PAHO, and most working MT in Japan.) Almost by definition, test suites fit into the traditional line of thinking about MT: that well-chosen examples are more important than tackling boring old texts that tend to have the bad habit of not fitting neatly with current theories.

But I fear 1 must say: Out with the theories and suites, and in with the unseen, robust, but very real text. In the end, there can be no freedom of choice unless one can be sure one is operating within a limited domain of vocabulary and structure. And, as far as that goes, if test suites give us a way of defining the domain and its associated structure and vocabulary, then I have no objection to them.

I also have this unreconstructed feeling that there are not really a number of sui generis evaluation methods: there is simply the basic default method which measures the percentage of sentences translated correctly (or "acceptably"), and any deviation from this methodology requires an explanation.

On the other hand (and now some of my high-minded dogmatism must slip away), there clearly do exist test environments that are real-world (as some would say), but which function with something less than correct-percentages, as our panel rubric indicates. And as practically everyone in the funding/sponsoring world would be quick to remind us, everything comes down to the issue of cost in the final analysis; so we cannot remain aloof from the consideration of what resources will be needed to establish a chosen measure.

So, if we were asked to support up a weak bilingual, translating on-line at a workstation; then large lexical translation aids, statistically-based generation promptings for the target language, and the like, could very well improve performance, and in a way not measurable by "percentages of sentences correct." Hence, these results might not satisfy a certain type of RW, since such improvements in performance would probably not need a very high-powered MT core engine, And to the extent that that is so, the assumptions behind this panel are correct, and fly in the face of my own prejudices.

Since I have been less than enthusiastic about test suites, "glass boxery" (examining the system's workings), pure MT research, and so on, allow me to end on a more constructive note. First, I think that the notion of extensibility is crucial, and that every class of customers should be concerned with it, not only commercial ones. My own attempt at a fairly large-scale evaluation project (of the SYSTRAN Russian-English translation system back in 1979) was done on the assumption that SYSTRAN worked, an assumption that had been validated before, but which was not tested in the method I set up. What was being tested were the limits of its extensibility to new text types. That is a central notion: and its significance will grow as MT systems mature and increase in number: old systems seem not to die, they just move away from center stage.

Secondly, I think that people need to re-examine the purportedly well-attested correlation between fidelity and intelligibility, and its implication that monolingual evaluation is as good as bilingual evaluation (i.e. monolinguals are able to assess intelligibility but not fidelity, so if there is a strong enough correlation between the two, then one is as good as the other). One clue we have about the fallacy of this assumption is that the IBM statistical MT project I referred to earlier works by effectively guaranteeing intelligibility, since the output obeys standard stochastic norms for the target language. Yet experience has shown us that it does not yield over 48% correct fidelity at the present time. How can that be? Which of our assumptions are wrong, and what are the consequences for evaluation methodology?

This is no idle question, as can be seen from Henry Thompson's suggestion in the panel rubric that a method of MT evaluation should be designed in which the output is tested for its statistical divergence from human translation of the same texts. This would, in effect, cut down on evaluation costs, but what interpretation could possibly be assigned to the measures it came up with? Depending on what statistical bases the method had access to, it might just as easily serve as a critique of the hand-translators' intelligibility, as of the quality of the MT. In other words, his suggestion as such says nothing, and I suspect that the statistics needed to bear it out would render it as costly as any other evaluation method. This is not meant as a personal criticism (as I have not studied the details of his proposal), so much as a general reminder that statistical methods may be as disruptive and revolutionary to evaluation, as is their re-introduction into MT itself.

3 Sture Allen's Statement

MT Evaluation from the Standpoint of a Research Sponsor

As a research sponsor I should choose my evaluator(s) with great care. Anybody entrusted with the task would have to be aware of the facts that (a) translation, as a matter of principle, is impossible and (b) translation, to the extent that it is feasible, is a linguistic problem, basically.

My instructions would include the appraisement of the relevance of the linguistic model with special reference to its lexical capacity, its performance with respect to knowledge, and the amount of context taken into consideration in analysis.

Furthermore, I should emphasize the checking of whether the architecture of the computational system permits insights into its operation as well as whether interaction is made use of in disambiguation.

Any attempts at internal evaluation should be noted and taken advantage of.

The organization and management of the project, the state of affairs in relation to the project goal, and the economic situation should be investigated and commented upon.

My orders would also make clear that I expected the evaluator(s) to break off the evaluation and report back to me if there were any tendencies towards preediting.

4 Ulrich Heid's Statement

MT Evaluation from the Standpoint of an MT Developer

The following is a short summary of some ideas and views on evaluation, trying to take the point of view of a (fictitious) software company interested in acquiring a prototype of a machine translation system in order to develop it further and make a marketable product out of it. The basic question is: having one or several candidates to evaluate, what are the goals and methods of the linguistic evaluation?

The situation:

- an embryonic system exists and is demonstrated;
- the problem is to become convinced that
- the system can be extended in linguistic coverage, with respect to its performance and conviviality
- the system can be "customized" in order to meet particular needs.

Prerequisites for the evaluation:

- the system with its knowledge sources and intermediate representations should be accessible (no "black box" situation)
- the producer of the prototype to be evaluated should be available and co-operative.

Tasks: what is evaluated?

- at the level of linguistic description:
- is there a descriptive model underlying the linguistics of the system?
    - can I learn it?
    - is it operationalizable and communicable?
- is the description extensible?
    - adding new lexical entries
    - adding new grammar rules, monolingual and contrastive
    - are there (un)predictable interactions between old and new, between grammar and lexicon?
    - what does it cost to learn about this? what to handle this?
- is the description modular?
    - can I "extract" parts for certain applications?
    - is there a way of providing different kinds of descriptive fragments for different needs? ("tuning" of the system)
- at the level of system architecture, environment and conviviality:
- is the linguistic description "hard-wired" with the user environment; can I add and modify?
- is it necessary, and if so, is it known to be possible or likely to be possible to re-implement (part of) the system in a more efficient way? Does this affect the linguistic description?
- (at the ergonomic level:) is the system easy to handle or do the people involved need a lot of training? (time and cost intensive)

Some proposals for "linguistic evaluation":

1. Getting information about the learnability of the underlying description is vital. This is only possible in close collaboration with the producer of the prototype. The essential thing here is to learn not only the "syntactic" choices to be made (how to write rules and entries), but the descriptive intuitions (what to describe). Having a collection of data supposedly representing observable linguistic differences (i.e. a test suite) would help; the question then is not (only); "can the system treat this?", but: "how do you tell the system to treat this?"

2. Testing modularity and extensibility is most difficult, since any assessment has to be based on the certainty that in adding some new description you do it in the "best" way with respect to the model underlying the system; might it not be possible to ask the producer of the prototype to convince us by demonstrating that adding new descriptions in an efficient way is feasible?

Preparing material for the tests:

- A test suite may be constructed in such a way that it contains instances of phenomena which we would like to see treated in the system.
- Running the test suite shows what is treated (in which way) now,
- Looking at the phenomena which are not treated or which are wrongly treated gives input to two types of assessments:
- can the producer of the prototype "repair" the "errors"? (consider also the time, effort, knowledge required).
- can the producer communicate to me what he does when repairing?
- Adding new phenomena to an existing test suite and using it again with the system and its developer as above, as well as with the "trained" potential customer (is he able to repair or add himself?) may

give hints as to the extensibility and upgrading possibilities.

A general remark for future developments:

There are currently several machine translation systems under development in academic and industrial research laboratories. Their chances of being further developed up to marketable products seems to depend on whether the researchers manage to convince developers. They may be more convincing when their systems are based on general assumptions easy to communicate and agree upon. Sticking to some "trend" may be an advantage, e.g. in the use of representation formats, in the basic elements of the linguistic description, etc. There is room for guidelines based on a minimal consensus about such choices: they would facilitate evaluation and the process of convincing developers.

5 Doris Albisser's Statement

Evaluating machine translation systems in a real work environment

As a potential customer of MT-software, I would evaluate commercially available systems using the following approach, which reflects the strategies employed for the evaluation of MT-systems at Union Bank of Switzerland:

When evaluating machine translation systems for productive use within a company, the underlying principle is to evaluate an MT-system as an overall system and not only for the quality of the MT-output. Also, the evaluation criteria should be designed so that they provide a true basis for comparison.

In this respect, the evaluation criteria to be taken into account can be subdivided into four main categories:

- the linguistic capabilities of an MT-system
- the technical environment provided
- the organizational changes involved
- the corporate situation of the MT-supplier

As a preliminary remark, it has to be pointed out that the evaluation criteria and, in particular, their weighting are company-specific and thus subjective to some extent. Furthermore, quality issues are not quantified, they are rated according to their degree of importance. As regards the procedure, evaluations are carried out inhouse using company-specific texts. Below illustration of the four main categories briefly outlines the evaluation strategies employed,

As for the linguistic part, we have developed a method to assess the quality of the raw MT-output, First, sentences of a given text are categorized according to their degree of complexity (ranging from I to IV), Second, the mistakes found in the raw translation are scored according to a list of criteria which could be explained in detail at the panel. Basically, the determining factors for scoring the mistakes are whether they can easily be corrected, whether they seriously hamper understanding, and whether they violate basic grammatical structures. It should be noted that the linguistic evaluation is largely language-dependent and to some extent even specific to the text type.

Furthermore, the technical environment offered by the MT-software supplier has to meet certain requirements HO as to comply with the corporate information technology strategy (e.g. open systems architecture). This may include portability, interfaces to sophisticated word processors (WPs) and desktop publishing systems, access to terminology from WP, import/export of terminology, options for information retrieval (e.g. for recurring texts, updates), single vs multi-user system, and — most important — user friendliness. Another important factor is the capabilities for further enhancement of a system. Since commercially available systems tend to lend themselves for specific text types, the question arises to what extent an MT-system could be customized.

The third and very often neglected evaluation criterion refers to the organizational changes involved. An evaluator has to determine the required user profile. Questions arising in this context: Are terminologists needed? Who does the system administration? Can present translators be trained (and if so, what is the learning curve)? Another important factor comprises the cost/benefit analysis. Thus, what is the price of the system, what is the minimum translation volume to justify MT, and what is the throughput per day, including both the volume of MT-output and the pre-/post-editing time required. The latter can only be estimated. As a consequence, the increase in productivity can be assessed during the evaluation phase to a limited extent only.

Finally, the corporate situation of the MT-supplier plays an important role in terms of future development and cooperation. Issues raised in this respect include the size of the company (resources for development), the importance of MT-software within the overall product range, the market share, management, the financial situation, and — very important — customer support.

In view of future integration of MT-systems into a corporate environment two general questions might be worth reflecting upon. First, what is the potential of an MT-system to be integrated into a translator workstation? Second, does the MT-supplier take into account that translation is only part of the entire document production process or does he offer the MT-system as an isolated component?

In conclusion, I should like to emphasize that all four parts (linguistic, technical, organizational, and supplier-related) are of equal importance. Thus, an MT-system is evaluated as an overall system. Finally, subjectivity cannot be avoided in an evaluation because each company has its own needs and priorities. What might be generalized to some extent is the evaluation criteria as such, but not their weighting.