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

Panel on Evaluation: MT Summit IV. Introduction.

M.King
ISSCO and ETI
University of Geneva.
54 rte des Acacias, CH-1227 Carouge Geneva
Tel: (41) +22 705 7114, Fax: (41) +22 300 10 86
e-mail: king@divsun.unige.ch

In order to structure the discussion, each of the panelists was asked to react to the following points:

1. Please summarize briefly your own experience with the evaluation of machine translation systems.

2. A recently published report (Galliers and Sparck-Jones, 1993) claims:

"The many elements involved in evaluation, perspectives and levels on the one hand and system structures and applications on the other, mean that it is quite unreasonable to look for common, or simple, evaluation techniques...., what can be common in the field is rather a set of methodologies and an attitude of mind. Thus evaluation in NLP is best modelled by analogy with training the cook and supplying her with a good batterie de cuisine."

Do you agree with this?

3. The programme of work for the European Commission sponsored EAGLES group on Evaluation and assessment distinguishes three types of evaluation, as follows:

Progress Evaluation: assessing the actual state of a system with respect to some desired state of the same system, as when progress of a project towards some goal is assessed. When this kind of assessment is applied to successive versions of the same system, it can also provide a way to measure progress. When a number of different systems have to be compared (as, for example, in the evaluation guided research paradigm exemplified by the DARPA SLS and SR programmes), this kind of evaluation can be used as the basis of comparison. In summary, this kind of evaluation is typically used to compare like to like, measuring system performance against some pre-established criterion. We can therefore further distinguish the choice of criteria, the measure used and the method employed to obtain the measure. This kind of evaluation tends to make large demands on resources in the form of test collections or annotated corpora, for example, which must be established and distributed.

Adequacy Evaluation: assessing the state of a system with respect to some intended use of that system, as exemplified by a customer investigating whether a system, either in its current state or after modification will do what he requires, how well it will do it and at what cost. ... Once again, this kind of evaluation may involve comparison between two or more systems. This kind of evaluation is very much oriented towards specific requirements, and may therefore require considerable work to establish the potential customer's needs. One familiar model of a more general version of this kind of evaluation is the consumer organisations which publish the results of tests, e.g. on cars and appliances, and identify "best buys" for certain price and performance targets.

Diagnostic Evaluation: assessing the state of a system with the intention of discovering where it fails and why, as exemplified by a research group examining their own system. Typically, this kind of evaluation requires intimate knowledge of the system examined, and is not done comparatively, although there may be comparative study of the effects of alternative versions of some system component. This kind of evaluation typically involves production of a system performance profile with respect to some taxonomisation of the space of possible inputs. The most familiar examples are the use of test suites in work on machine translation and natural language front ends. It tends to make large demands on manpower to undertake the taxonomisation and design and compile the test material.

Do you accept this distinction between different types of evaluation? Are they really different from one another in significant ways? Would you like to refine the distinction?

4. The passage quoted above makes frequent reference to the cost of providing test materials for evaluation (test corpora, test suites, test collections etc.). Do you see any way of reducing the cost, or of sharing the test materials across a number of evaluations?

5. Free comment: please feel free to add anything that you wish.

As will be seen from their individual contributions, different panelists put different emphases on each of the points. An attempt is made here to summarize areas of agreement and disagreement.

1. Experience.

All the panelists have extensive experience of evaluation. For completeness, I should add my own. I first became actively involved in evaluation of machine translation systems (as opposed to being a more or less passive victim of it) around 1987, when ISSCO was requested by the Swisstra Association to produce a set of guidelines for use by potential customers of machine translation systems. Typically, these potential customers were civil servants, with no expertise either in computational linguistics or in machine translation. It very quickly became clear that the situation and context of work of the potential customers, as well as the needs they expected to be fulfilled through use of a machine translation system, varied so greatly that no single evaluation method could be useful. This realisation pushed me into considering what could usefully be provided; a question that has remained one of my major preoccupations ever since.

On the practical side, I have been involved in the evaluation of both research projects and of commercial systems, and have been able to benefit from my colleagues' experience in long-term evaluation in collaboration with a large telecommunications company, in internal evaluation within ongoing research projects and in evaluation of other natural language processing products. In 1991, ISSCO organised an Evaluators' Forum, which brought together a group of about forty people interested in exchanging their experiences of evaluation.

In recent months I have been able to benefit from discussion connected with the work of the EAGLES Evaluation and Assessment working group, which is part of an EEC initiative aimed at working towards the production of standards for linguistic engineering.

2. The Cook and Her Tools.

Although most of the panelists seem to agree that evaluations have to be hand tailored for specific circumstances, there does seem to be some feeling that practical experience of evaluation will lead to common methodologies, at least for progress and diagnostic evaluation. Adequacy evaluation however, is felt to be involve more complex factors, where it may be harder to define easy to follow guidelines.

3. Types of Evaluation.

Most of the panelists point out that it is hard to distinguish clear borderlines between different types of evaluation, even when they agree that thinking in terms of different types of evaluation can be useful in keeping perspective and structuring the task.

4. Common Tools.

There seems to be widespread agreement that some common tools can be developed, although what these might be varies considerably across the different panelists.

All of these points will be discussed more extensively during the panel itself. The audience is invited to prepare their own interventions.

The Panelists Contributions.