[*Machine Translation Today: Translating and the Computer 15*.
Papers at a conference…18-19 November 1993 (London: Aslib)]

Evaluation of MT software and methods

Margaret King
ISSCO and ETI, 54 rte des Acacias, CH-1227 Carouge, Geneva.

In this paper we first discuss the wide variety of possible scenarios in which an evaluation may be carried out. On this basis, an attempt is made to pick out some general characteristics relevant to the design of evaluations and the search for a general framework for evaluation methodologies is motivated. Some resources for use in data collection for evaluation purposes are briefly described and discussed, and the work of the EEC EAGLES Evaluation and Assessment Group is summarized.

INTRODUCTION

Although the title talks explicitly of machine translation, I hope that much of what is said will apply to the evaluation of language industry products in general: although less ambitious in their aims than systems which aim to produce a draft or finished translation, a growing list of products which can greatly facilitate the translator's task and relieve him of some of its more tedious aspects exists. Amongst the more modest are spelling checkers and automated dictionaries; the range continues through grammar and style checkers and terminology servers to the specialised work stations now being developed, which aim at providing access to previous translations as well as document preparation services specially conceived with the translator in mind.

Evaluation is relevant to all of these, and, from an end-user's point of view, perhaps especially to those which, precisely because they are relatively limited in their aims, rely on a stable linguistic technology and appear in considerable numbers on the market; choosing which spelling checker to buy can require quite a lot of thought.

EVALUATION IS VARIOUS

A starting point for thinking about evaluation is to notice how wide is the variety of those who might be interested in evaluating something; there are those who wish to buy a finished product, but tracing the life time of that product backwards, there are also those who developed it, those who

invested in the development, those who are trying to sell it, as well as those who did the original research and those who funded them for doing it. At the extreme ends of the spectrum, there are the policy makers who decide that now is the moment to invest heavily in a particular line of research, and those who want to decide whether a system that has been installed for some time is really earning its keep.

Another way to look at this variety is to consider what the objects of evaluation might be: if this time we work in the opposite direction, we can distinguish a range going from research proposals, through research (or development) work in progress, through prototype systems intended to demonstrate the validity of the technology underlying them to actual market products or systems that have already been installed. But thought of this way, a quite important point begins to emerge. Except in the case of research proposals or research prototypes, it is quite rare for a system in isolation to be of interest. In most other cases, what has to be evaluated is a system in a context - what Galliers and Sparck Jones (1) call a "set-up". A couple of examples will help to make this idea more concrete.

A system which in a single user situation is perfectly acceptable may turn out to be disastrously slow when many users are simultaneously involved. For example, the requirements on a system providing information on railway timetables will be radically different if it is to be accessible to the general public rather than used by an information clerk in the railway travel office. Similarly, there is little point in introducing even a satisfactory machine translation system into a translation service if it requires that the text be typed in and every translator in that service is refuses to touch a keyboard. King (5) discusses the particular contexts relevant to a translation service.

TOWARDS GENERALISATION

Once we think about all the different kinds of people who might carry out an evaluation, the variety of objects to be evaluated, the range of purposes for which the evaluation might be carried out and the diversity of contexts which might be relevant to the evaluation, it might seem hopeless to try to generalise at all. We can make a start however by thinking first about why people evaluate. Although the borderlines are rarely totally clear, we can distinguish three main reasons.

First, they may evaluate in order to check progress towards a goal: a typical example would be a research worker or a group carrying out development of a system. They have some idea, stated or unstated, about what that system should be able to do. They evaluate their own work in order to discover how far they have progressed towards that goal.

The same group may evaluate their work in order to discover what problems persist and diagnose the reasons behind

malfunctioning of the system. Another example of diagnostic evaluation may occur when a potential customer is evaluating *a* machine translation product: it is quite rare to be able to buy a system off the shelf which does exactly what any given customer requires. (For example, at a simple level, if he has specific terminology that should be used with his translations, it is unlikely that a general purpose system will already include it). Thus, when evaluating the system, he might well be interested in trying to find out what errors can be repaired by modifying the dictionary, what errors would require more extensive work.

The potential customer typically has a set of needs he wants a system to fulfill: his main reason for evaluating *a* system will be to find out if the system, once introduced into the context he foresees, will be adequate to those needs.

Essentially, then, we have distinguished three types of evaluation: progress evaluation, diagnostic evaluation and adequacy evaluation.

Once the type of evaluation has been determined, the next stage in designing an evaluation is to decide what dimensions of the system are relevant. Obvious dimensions are things like whether the system does what it is supposed to do, and whether it does it efficiently, but considerations of context can be important here too. For example, a large company considering introducing a machine translation system into a service which already makes substantial use of computing, may be insistent that any candidate system can easily be integrated with the hardware and software platform which already exists.

Having decided what the relevant dimensions are, the next step is to look for criteria which will provide information on how well the system being evaluated performs on that dimension. Let us pause once more here to make things concrete through an example. Let us imagine that we are designing an evaluation for a potential customer of a machine translation system. Since we shall clearly be involved mainly in adequacy evaluation of whether a system meets his needs, determining what dimensions are relevant will involve examining his needs very closely. Notice that this can be a long and difficult task, since it involves sorting out from his, perhaps naive, perception of his needs, and the evaluator's, perhaps misinformed, perception of his needs what the real needs actually are. It may even be the case that different parts of the same organisation - say the computing service and the translation service - have quite different views of what the real needs are. But let us imagine that this has been done, and we know that one relevant dimension is the efficiency of the system. One possible criterion for efficiency is speed; how long it takes to translate some pre-determined amount of text.

We are now faced with the problem of defining a way in which speed can be measured. A variety of possible measures are conceivable. For example, it is a well known characteristic of machine translation systems that processing speed tends to deteriorate as a function of sentence length. Thus, (I hope

this is caricature), if it takes two seconds to translate a two word sentence, it may take 10 seconds to translate a three word sentence and three hours to translate a thirty word sentence. So let us take just one aspect of our criterion and make it not simply speed, but the speed with which a sentence can be translated. As a measure of this we might imagine taking a text typical of the kind of text we are interested in translating, and choosing from it a set of sentences with length ranging from the shortest we can find in the text to the longest. Thus, if the shortest sentence is three words long and the longest fifty, we will take a three word sentence, and a four word sentence, and a five word sentence, gradually increasing the length up to fifty. (Of course, we might not be able to find a sentence of thirty seven words, but we will ignore that complication). We can then, as a method for obtaining the relevant data, submit each sentence to the system, discover how long it takes to produce a result for each of our sentences and perhaps plot a graph to show how processing time increases with sentence length.

Unfortunately, the measure we have chosen may not be so straightforward as it first seems. Measures in general should be both valid and reliable; that is, they should measure what they are supposed to measure, and they should do it reliably - we should not get different data on different occasions. The measure we have chosen may not be valid; certain syntactic constructions are typically harder to deal with than others, and their presence in the input may have a direct influence on the processing time. Thus "Who did John see?" may, for good reasons, take longer to process than "John saw the dog" despite the fact that both are four words in length, and the longer the sentence, inevitably, the harder it will be to disentangle the effect of length from the effect of different structures. Similarly, the measure may not be reliable. If the system is being used in conjunction with other computing processes, we may get different answers depending on what demands are being made on the computing power available by those other processes. To give a simple example, we would almost certainly get different answers if we ran the test via a windowing interface than if we ran it without one.

There is no space here to discuss in any detail what valid and reliable measures might be, even in the specific case of our chosen example. The important point is that it can take time, care and sometimes considerable imagination to come up with measures that are relevant to the criteria chosen and are also valid and reliable. Similarly, the method used to obtain the data used for measuring must also be valid and reliable; it is no use, for example, taking as a measure the length of time taken to correct a translation via a text processing system on a screen if in some cases the screen has sunshine falling on it and sometimes does not.

To summarize this section: designing an evaluation can be thought of as a top-down process. First the type (or types) of evaluation in question are determined. Secondly, the dimensions relevant to this particular evaluation are defined. Then a set of criteria refining and defining each dimension in more detail

are defined, and a set of valid and reliable measures pertinent to each criterion defined. For each measure, a method of obtaining the data relevant to that measure is also defined. The method too must be valid and reliable. King (6) discusses the overall design of evaluations and examines some past evaluations in the light of the design suggested here.

SOME MORE ABOUT MEASURES.

Before we return to more general questions, it is worth remarking on one further distinction: measures may be qualitative or quantitative. Qualitative measures are those which involve some kind of subjective judgement, as when, perhaps, we want to evaluate the user interface to a system and we simply ask a representative group of people to use the system for a while and then rate the interface on a three point scale as being good, acceptable or bad. Quantitative measures, as the term implies, involve counting something, for example the time from input to output in our previous example. A look at the literature on evaluation will reveal how important it is not to get these mixed up; it is only too easy to give a spurious air of objectivity to what is basically a set of subjective judgements. This is especially true when extraneous factors may influence the judgements made. One anecdote recounts an evaluation where both human and machine produced translations were to be rated. If the human translations were typed and the machine translations printed on a line printer, the human translations were consistently judged to be better than the machine translations. Unfortunately, if the machine translations were typed and the human translations printed, the result was reversed and the machine translations consistently judged to be better - even when the same translations were put up for judgement in the two cases.

This too brings us to a general point, which has already been touched on a little when we were talking about methods; questions of good experimental design are critical to designing a good evaluation. Just as it is of doubtful validity to ask three people whether they prefer black cars or red cars and produce only cars of whichever colour gets most votes, it is not very reliable to base an evaluation of a machine translation system on one person's going to see a demonstration of it. This point, and other related issues, are discussed in detail in Falkedal (2), which contains a valuable critical assessment of some past evaluations.

TOWARDS A GENERAL FRAMEWORK.

A first motivation for investing effort into trying to define a general framework for evaluation methodologies emerges from what has been said so far: evaluation is difficult. There is a very strong argument for sharing experience, discussing the strengths and weaknesses of different techniques, recognising and publicising mistakes, all with the aim of coming to a

common understanding of what is involved in designing an evaluation and in carrying it out. Far too often in the past every new evaluator has started from scratch; he has, of course, gone over what is in the literature, examined it critically and then either rejected it or tried to build on it. But until comparatively recently, most evaluations were carried out under contract to a particular organisation and the results presented only in the form of an internal report to that organisation, so that the literature is sparse. And even the best intentioned evaluator, if there is no publicly available common fund of wisdom to draw on, is necessarily limited by his own background and experience and potentially tainted by his own prejudices.

Furthermore, building such a fund of common wisdom would have almost as a side effect that it would become easier to share the results of evaluations. Even if one did not agree with the way the evaluation had been done, or disagreed with the conclusions reached, there would be shared grounds for arguing about it, or for rejecting some parts and accepting others. This would help to avoid considerable waste; how many separate organisations have independently evaluated any single one of the well-known machine translation systems?

One possible obstacle to sharing results might be reluctance on the part of the manufacturers and vendors of particular systems to let the results of the evaluation be known; but there does seem to be a growing awareness amongst the manufacturers, manifested by a willingness to participate in discussion of what constitutes a valid evaluation, that ultimately it is to their advantage to have well-established and commonly accepted evaluation methodologies.

The organisation of an Evaluators' Forum in 1991 (see Falkedal (3)) and of a workshop on evaluation in 1992 (organised by AMTA with the support of IAMT) is a concrete acknowledgment of the importance of sharing experience and working towards a common understanding. A further workshop in 1992 was organised by EAMT in conjunction with the Saarbrucken Technology Fair.

A further motivation for collaboration begins to emerge when the cost of a serious evaluation is taken into account. Even what has been said so far implies a considerable investment, and we have not so far looked at all at what kind of resources might be needed to carry out a valid evaluation and how much they might cost. Let us turn to that question now.

RESOURCES FOR EVALUATION.

As can be deduced from previous discussion, evaluation is becoming more and more of an expert task. Apart from the cost of this expertise in itself, the cost of an evaluation depends also on the resources required to collect the pertinent data. In previous work, four types of resources have been used or proposed.

One classic technique is to use rating scales, for example by asking a group of people to rate translations for their intelligibility or for their faithfulness to an original. Some problems have already been mentioned with the use of such techniques, but if we set aside any reservations about their use, it remains true that rating scales cost money to set up. This cost is increased when good experimental design aimed at removing or at least mitigating some of the problems is also required.

A second classic technique is to submit a text or a collection of texts to the system to be evaluated and to examine the results. Such a corpus may be specific to a particular context, for example when a potential customer is primarily interested in a system's ability to translate his own technical documentation, or may be more general. In either case, if the corpus has to be constructed, it will cost time and money to do so. Even if the texts are available in machine readable form, they will typically have to be prepared for input to the system, for example by removing formatting commands or photo-composition codes. Especially in the case of general corpora, there are also problems of representativity which may call for the investment of additional expertise. For example, one of the well-known corpora publicly available is the proceedings of the Canadian Parliament, especially valuable because parallel English and French versions exist. This corpus is clearly representative of the language used by Canadian members of Parliament, but not, say, of the language of written minutes or of technical documentation. Several recent initiatives, such as the European Corpus Initiative, the Linguistic Data Consortium and the Data Collection Initiative of the ACL aim at collecting a wide range of corpora and making them available to the community as a whole.

Recently, quite a lot of interest has been shown in the use of test suites as a tool for evaluation. A test suite is a carefully constructed set of inputs, where typically each input is designed to probe the system's behaviour with respect to some specific phenomenon. For example, if we were interested in knowing whether a system could deal with verbs like "give" which allow the pattern "Give x y", we might have an input "John gives Mary a book", where possible interference from other linguistic phenomena is minimised by keeping the noun groups very simple. Such test suites are both difficult and costly to construct, and become more so if semantic or translational phenomena are to be taken into account. (See King and Falkedal (4) for a discussion of some of the problems). On the surface, at least, they therefore constitute a prime candidate for a resource to be developed collaboratively and subsequently shared. A current project (TSNLP) in the context of the EEC's Linguistic Research and Engineering programme is aimed at investigating the feasibility of this.

In the context of fact extraction, considerable effort has gone into developing test collections for use within the ARPA funded projects represented in the MUC conferences. (See, for example, Lehnert and Sundheim (7). A test collection is a

collection of inputs where with each input is associated a correct output, usually accompanied by guidelines justifying the definition of the correct output. As might be imagined, test collections are even more costly to develop than are test suites. Except in very specific cases it is hard to imagine how a test collection could be developed for use in evaluating a machine translation system, simply because of the difficulty of defining a "correct" translation, but they deserve mention here if only because their development in the ARPA context has contributed much to the growing awareness of the importance and difficulty of evaluation.

THE WORK OF THE CEE EAGLES GROUP ON EVALUATION AND ASSESSMENT.

Considerations such as those discussed in the last two sections are at the basis of a recent initiative of the Commission of the European Communities. With the aim of working towards re-usable resources, the Expert Advisory Groups on Language Engineering Standards started work early in 1993. There are five groups in all, on Lexica, Corpora, Linguistic Formalisms, Speech and Spoken Language and Evaluation and Assessment.

In a desire to reach concrete results as soon as possible, the Evaluation and Assessment Group has decided to limit its work during the first two year period of activity to developing evaluation methodologies for market or near-market products in application areas where the underlying technology is relatively stable. Thus work will concentrate on methodologies for evaluation of writing aids and translators' aids, and a preliminary investigation of what might be required from information management systems.

In an attempt to produce results which will be useful to a large number of people, the paradigm adopted is rather like that of the reports to be found in consumer organisation magazines like "Which?". Different dimensions potentially relevant to adequacy evaluation of the products in question are determined and criteria, measures and methods given for evaluation along those dimensions. The user of the evaluation can then determine for himself which dimensions are relevant to his particular case and tailor the evaluation to suit his needs.

Despite the comparative modesty of these ambitions, the work is intended to contribute to the definition and validation of a general framework for the design of evaluations, which, it is hoped, will contribute to resolving some of the problems set out in this paper. The results of the work will, of course, be disseminated as widely as possible.

(1) Galliers, J.R., and Sparck Jones, K., 1993. "Evaluating Natural Language Processing Systems", Technical Report No. 291, University of Cambridge Computer Laboratory.

(2) Falkedal, K., in progress. "Evaluation Methods for Machine Translation Systems: An Historical Overview and a Critical Account". Draft report, ISSCO.

(3) Falkedal, K. (ed.), in press. Proceedings of the Evaluators' Forum, Les Rasses, 1991. ISSCO, Geneva.

(4) King, M. and Falkedal, K., 1990. "Using Test Suites in the Evaluation of Machine Translation Systems", Proceedings of Coling '90, Helsinki, pp. 211-219.

(5) King, M., 1992. "L'évaluation des systèmes de traduction automatique dans le cadre d'un service de traduction". In D. Estival and M. Cormier, eds. Special Edition on Machine Translation, Meta. Vol 37, No. 4, pp. 817-828.

(6) King, M., in press. "The Evaluation of Natural Language Processing Systems". In Y. Wilks (ed) Special Edition on Natural Language Processing, CACM.

(7) Lehnert, W. and Sundheim, B., 1991. "A performance analysis of text analysis technologies". AI Magazine 12 (4), pp. 81-94.