[*Proceedings of the Evaluators' Forum*, April 21st – 24th, 1991, Les Rasses, Vaud, Switzerland;
ed. Kirsten Falkedal (Geneva: ISSCO).]

Declarative Evaluation of an MT system: Practical Experiences

Lorna Balkan, Matthias Jäschke, Lee Humphreys,
Siety Meijer & Andy Way
Department of Language and Linguistics
University of Essex

1 Introduction

The authors recently had the opportunity to evaluate the performance of a small commercial MT system – Globalink Translation System (GTS) – which runs on PC-type machines. A review which we published in the popular UK small systems journal *Personal Computer World* (1) was very much directed towards the needs of potential users. The present paper is intended as a rather fuller account of some of the difficulties encountered in trying to construct an appropriate evaluation method within a realistic time-scale. The moral of the paper is plain enough: evaluating an MT system from a user-perspective is a much trickier business than most Computational Linguists might suppose.

2 Types of Evaluation

There are two groups of people interested in the evaluation of an MT system – system developers and (potential) system users. The developer typically wants to check on the number and types of linguistic deficit the system has e.g. whether or not it handles pseudo-clefts. We shall call this type of evaluation a *typological evaluation.* The currently favoured tool for typological evaluation is the *test suite* – a structured set of test sentences which individually embody some specified linguistic construction and collectively constitute a sampling of overall linguistic performance (3). Note, however, that this type of evaluation is unlikely to be of much help to the majority of potential users: it presupposes a knowledge of linguistic theory and the expected frequency of the various linguistic constructions in the user's chosen text type; if a particular construction is of very low frequency, the system's failure to handle it is unlikely to be a serious practical problem. Hence although perspective evaluation clearly has its place and forms the subject of active research (e.g. in what sense can a suite be constructed to probe specifically translational problems), we shall have no more to say about it in the present paper.

We can discern two broad evaluation strategies of particular interest to the potential user:

Declarative Evaluation which seeks to specify how an MT system performs relative to various dimensions of translation quality (6),and

Operational Evaluation which seeks to establish how effective an MT system is likely to be (i.e. in terms of cost effects) as part of a given translation process (4,2,5,7,8).

In the first case one attempts to answer the layperson's question "Just how good are the translations it produces?" directly; the cost-benefit question – "Will it help to reduce my translation costs?" – is addressed indirectly. With operational evaluation the priorities are reversed: cost-benefit comes first, with quality assessment being used solely as a check on whether some criterion performance has been achieved. An operational evaluation usually entails finding out whether or not the productivity of a translator is improved when s/he is given MT output to post-edit rather than doing "pure" human translation *ab initio*; a verification procedure is required to ensure that, in each case, the quality of output is the same.

We think there is a lot to be said in favour of an operational evaluation – it provides the potential user with very direct purchasing guidance. However, it does have the drawback that nothing is said directly about the actual quality of the translations produced, such as the proportion of sentences which are "good" translations, and to what degree the remainder are less than good. If a potential user has an application in mind which differs from that identified in an operational evaluation, the results will not easily be applicable.

Since the readers of a Personal Computing Journal constitute a rather varied group of potential users, we decided that it would be appropriate to provide quality assessment by some sort of declarative evaluation[1].

Examination of the literature suggested that the early work by J. Carroll for the (in)famous ALPAC report on MT was worth considering as a possible experimental model.

[1] It would have been attractive to conduct *both* types of evaluation – but this would have been well beyond our resources. We also investigated some properties of the GTS system as a whole e.g. user-friendliness and dictionary-updating; this was the least problematic part of our evaluation and we have nothing to say about it in the present paper

3 The Carroll Study

The principal published declarative evaluation of MT systems is to be found in J.B. Carroll's elegant contribution to the US Government ALPAC report (6).

A group of 18 English monolinguals and 18 English native speakers with a "high degree of competence in the comprehension of scientific Russian" were asked to evaluate 3 human and 3 machine translations on 10 point scales along the fidelity (multi-lingual) and intelligibility (mono-lingual) dimensions. Thus, for example, the most intelligible sentence was one which was

> "Perfectly clear and intelligible. Reads like ordinary text; has no stylistic infelicities."

Fidelity was determined by asking raters to score for the "relative informativeness" of an original sentence (or a good quality human translation, in the case of the monolinguals) when compared to its machine translation. The translated material had as its source 5 passages from a Russian technical work (Machine and Thought, by Z. Rovenskii, A. Uemov and E.Uemova, Moscow 1960). All the monolingual raters were described as having high (tested) verbal intelligence and "good backgrounds in science".

The results of the evaluation were principally presented as histograms plotting the mean intelligibility rating (abcissa) for all the sentence sample against frequency (ordinate). Carroll found that, when averaged over sentences, passages and raters, fidelity and intelligibility scores were very highly correlated; only for a few particular sentences did the mean ratings of intelligibility and informativeness convey different information. This is hardly surprising: if an MT system succeeds in producing a translation which reads plausibly, it is unlikely to be completely erroneous[2].

4 Design of GTS Evaluation

Carroll was evaluating several MT systems at once; in the present study, only one system was investigated.

Multiple-Language Pairs The Carroll study was conducted solely with translations from Russian into English. For the GTS evaluation, we needed to extend the methodology to handle language translation in both directions

[2] That is, MT systems don't generally exhibit the translational misdemeanours of Monty Python phrasebooks

for two language pairs (Spanish-English, German-English – these being the modules supplied for review). Such an extension ought to be simple, but in fact it led to our first mistake – and hence our first warning:

> *If you use non-native speakers as raters, make sure that they are supplied with scoring instructions etc in their own language.* !

We used non-native speakers to assess the quality of translations *out of* English into Spanish and German. (In a trial, we did try using a German native speaker with very good spoken English to assess some translations from German into English: his acceptability judgements on the English was completely at variance with both monolingual native English speakers and native English speakers with a good knowledge of German.) Since all our non-native speakers had good English skills, we supposed that they would be able to read and comprehend our instructions without any difficulty. And, indeed, informal "post-experimental chats" with these raters tended to confirm our view. However, it is impossible to be sure that these evaluators really did have the same grasp of the instructions as their native-English speaking colleagues would have done; moreover, since all the *examples* for Intelligibility were given in English, they would have had to invent equivalent examples for themselves in their own native languages. In the event of finding significant differences between the performance of the system depending on its direction (e.g. into or out of English), the effect of English-only instructions could not be ruled out as a factor.

During trials, all raters were given questionnaires asking whether or not they considered the instructions they had been given to be clear or unclear or unintelligible etc. Since the responses during the (later stages of the) trials were almost uniformly positive, no such questionnaire was used during the evaluation proper. However, in general it *is* desirable to achieve structured feedback from raters since this enables the experimenter to be sure that they consider the task well-defined and tractable; if they do not, some sort of redesign is clearly necessary.

> *Make sure that structured feedback is available from raters e.g. via a questionnaire.* !

Scale Design Carroll's study represents a fairly substantial deployment of resources (18 raters per group); we were interested in the possibility of using rather fewer raters (5 per group) to achieve useful results, since it is unlikely that 18 skilled translators can be employed in many practical evaluations. The manpower problem becomes even more acute when attempting to assess translation performance on more than one language pair in both directions. Using fewer raters suggested that we should simplify the scales in order to achieve the maximum possible inter-rater scoring consistency.

Whilst recognising Intelligibility and Fidelity (or Accuracy) as appropriate quality dimensions for the assessment of translation, we chose to start developing scale points and scale-point descriptors from scratch. Various trial scale descriptions were constructed and tested on practice texts or sentence sets by members of the experimental team and others. The object of each trial was to reduce raters perceived uncertainty as to which score each given sentence should be assigned.

No attempt was made to provide any formal test for the perceived interval-equality on either of the scales. However, care was taken during the development of the scales to ensure that the results were felt by individuals to have such a property. The ordering principle of both scales was the same i.e. the higher the score, the poorer the quality[3].

One particular problem with scale design is that there are very few "natural" constraints. For example, should both scales have the same number of points? The possibility of different scale lengths was left open from the start and, ultimately, was reflected in a choice of 1-3 for Accuracy and 1-4 for Intelligibility. The use of different scale lengths can cause some presentation difficulties – results for both dimensions cannot conveniently be presented on the same graph (unless one of its dimensions has two scales). On the other hand, Accuracy and Intelligibility are dimensions which, whilst correlated, are in fact incommensurable: even if results for both dimensions are presented together, no particular significance can be attached to their comparison.

How large should the scale range be? As noted, we opted for small ranges in the interests of promoting inter-rater consistency. However, it is not obvious that finer-grained scales than those we chose are actually necessary. During post experimental informal discussions with raters, not one ever suggested that s/he felt the scales to be too coarse to capture gradations in quality s/he felt important.

Training Feedback Carroll reports that "raters attended a 1-hr session in which they were given instruction in the rating procedure and required to work through a 30-sentence practice set"; presumably they received unstructured guidance and feedback on evaluation performance.

> *Make sure that any training of evaluators is structured and similar for each of them.*

The use of training with feedback within a single study is perfectly acceptable if raters receive a roughly similar training effort, matched responses

[3] Rather strangely, Carroll's scales go in opposite directions i.e. 9 represents highest or best intelligibility, whereas 0 represents best accuracy/fidelity.

to queries etc etc. However, unstructured training sessions can compromise any attempts to compare replications of the evaluation conducted with a different rater group and different experimenters, since there is no guarantee that the training regime will be similar. Since it is difficult to fully articulate sets of conditions and circumstances under which one score rather than another should be assigned to a test sentence, feedback on training examples can reflect individual experimenter preferences and hence affect individual rater behaviours.

In the present study, we attempted to reduce such effects by avoiding any experimenter-directed training and feedback. Participants were given a sample text on which to practice scoring before evaluating the experimental material. No feedback was offered in either practice or experimental sessions.

Text Context One particular feature of Carroll's experimental design is that raters were asked to score on both scales sentence translations which were

> "..selected randomly from a translation and interspersed in random order among other sentences from the same translation and also among sentences selected at random from other translations of varying quality" (p.68)

In the cited report no reason is given for presenting material this way rather than in a way which preserves the textual context of each scored sentence. One supposes that Carroll, presuming that the degree of contextual support provided for a given sentence varies significantly from sentence to sentence and/or text to text in a manner difficult to control, sought to ensure that scores were minimally affected by raters' contextual inferences.

De-contextualising sentences à la Carroll is likely to depress mean intelligibility scores (and possibly accuracy scores as well) by some constant factor compared to mean scores achieved on material with context preserved.

We were unable to discern any persuasive reason for de-contextualising the translation task and hence we chose to use continuous (unshuffled) text samples for all our evaluations. Hence

> *Make sure that the design of the evaluation reflects as far as possible the way in which the evaluated system is likely to be used.*

Text Type The Carroll study was clearly conducted with the premise that MT and HT should be compared with respect to a specific text-type, viz. Russian "scientific" texts. We had no such brief. It was our task to choose material that would be representative of the sorts of translation tasks for which an MT system like GTS might conceivably be used.

Re-examination of the ESPRIT document (used initially as a test corpus for EUROTRA) revealed that such a highly complex bureaucratic document was probably unsuitable, given that many of the sentences tended to be too long, elaborately structured and indifferently punctuated. It was decided that translated samples of this text would only be suitable as rater practice material.

> *Make sure that any source texts are* good *examples of the written language in question; a depressingly large proportion of distributed texts are badly punctuated, over-wordy and contain over-long sentences.*

Eventually, we settled on two types of experimental material: a short series of business letters produced by the Overseas Department of Essex University, and a few sections from the GTS operator manual. The latter, with its relatively short and simply structured sentences, was thought to be representative of the generality of high-quality technical documentation. The sections chosen were adjudged to be relatively comprehensible to persons not necessarily familiar with technical software documentation.

Source Material – Prepared Translations In order to investigate multilingual performance, it was necessary to prepare some translations of source material; we commissioned such translations from individuals within the Department of Language and Linguistics at Essex who were native speakers of the language they translated into and who, in our opinion, had a very good knowledge of English (the usual source language) and some experience of translation. Limited resources meant that we were not able to arrange quality checks on the translations we were actually supplied with - a clearly undesirable state of affairs.

> *Make sure that* all *source material is subject to quality control – not just English source text.*

We gave no special instructions to our translators. However, it would probably be desirable to encourage them to preserve the *number* of English source sentences as far as possible in their translations – this ensures that the number of scored units in the evaluation remains pretty much the same regardless of which language pair the system translates.

Source Material – Length The length of the texts we used for evaluation was determined largely by the amount of time we felt we could reasonably ask our raters to spend - in practice, a 2 hour session.

Overall, the length of the text samples evaluated was almost certainly too small. Larger samples would have produced more data points and hence clearer trends for both Intelligibility and Accuracy scores. The situation was aggravated by our choosing two different text types for comparison, rather than just using a longer sample of one type.

> *Given limited resources, go for a larger sample of a single text type rather than smaller samples of different text types.*

Dictionary Coverage Typically, not all word forms encountered in an input text are covered in an MT system's dictionary. In the present study we wished to try and separate out lexical coverage issues from overall translation quality issues: we were interested in the question of how the system would perform assuming that all the words in the input were in its dictionary. This separation of dictionary coverage from other quality issues reflects – in our opinion – the fact that whilst dictionary extension for technical terms etc. is relatively straightforward and can be carried out during the lifetime of the system by the user, linguistic rules which affect non-terminological aspects of quality (e.g. handling of agreement etc.) are not amenable to any user upgrade or revision and hence constitute part of the "absolute" system performance.

In general, the system dictionaries for GTS exhibited good coverage of general language terms; where items were missing, the following policy was adopted:

- Missing proper names and acronyms (e.g. IBM) were left as produced by the system (i.e. the same as their input form, prefixed with @@). In all cases these forms were deemed sufficiently interpretable from the context.

- Monolinguals were supplied with translations of missing terms and items of general vocabulary on a separate wordlist.

- Bilinguals and those with a good knowledge of the source language were required to infer the correct translation from the untranslated source language string left in the GTS output.

It might be objected that supplying translations of missing words (and, in general, isolating quality scoring from lexical coverage) tends to produce an overestimate of system performance since the experimenter-supplied translation will always be appropriate for the context. However, the vast majority of missing words encountered were single or multi-word *terms* from various sublanguages: such terms tend to have unique (context-independent) translations.

Another possible objection is that the presence of lexical gaps is likely to degrade the overall quality of a sentence translation. To the extent that such a claim is true – and we have no particular evidence on this matter – it may well be that the experimental design would be improved by updating the dictionary with the missing words before the translations for evaluation are produced. The principal reason for not carrying out this procedure was a shortage of time.

Make sure you have a sensible policy for coping with the possible effects of missing words on system performance and rater decisions.

Scoring Interaction between Scales During trials, informal discussion with participants made it clear that they felt their scoring behaviour for Accuracy was strongly influenced by the (prior) scores they had assigned for Intelligibility. Accordingly, we decided that raters should be presented with fresh unmarked copies of the evaluation text after they had scored it for Intelligibility; raters thus assigned Accuracy scores without having their Intelligibility scores in front of them.

A further problem was that, in the view of the trial participants, there was no obvious rationale for assigning an accuracy score to a sentence which had already been deemed to be more or less unintelligible. For this reason, we decided that Accuracy scores should only be assigned to sentences which achieved a score of 3 or better for Intelligibility; sentences scoring 4 for Intelligibility by a given rater were removed from the evaluation texts (and replaced by their source language sentence) before it was returned to individual scorers. This manipulation was intended to ensure that full context for other sentences in the text was maintained. Unfortunately, one result was that the number of sentences scorable for Accuracy varied a little from rater to rater, depending on how many sentences they initially identified as unintelligible.

Make sure you have some reasonable policy for assigning Accuracy scores to unintelligible sentences.

Presentation of Evaluation Material What is a sentence? Is it something terminated by a full-stop, question-mark or exclamation mark – or is it something also terminated by colons or semi-colons? We considered that evaluation would actually be slightly simplified if all main-clause elements (plus any subordinate clauses) were marked independently – there would be fewer cases when one part of the sentence was good and the other rather less so. Although all our raters had some exposure to linguistics and clearly knew something about punctuation, trials revealed that raters were liable to be rather erratic in their identification of sentences in properly punctuated continuous text. In the experiment itself, we found it convenient and practical to pre-edit all evaluation material so that strings which we wished to be considered as a scorable sentence were clearly separated by white space.

> *Make sure that the raters task is as simple as possible.*

The material was also provided with a left-hand margin in which raters were instructed to place their scores.

The Scoring Process Raters were presented with typeset instructions in English describing how to score sentences for Intelligibility and two sets of material to be scored (Business Letters and GTS Manual). After an experimenter had told them briefly the purpose of the exercise viz. to evaluate the performance of an MT system[4], they were invited to read the Intelligibility instructions and then, when they felt they understood the instructions fully, proceed to score each sentence in the training material. Raters were able to refer to the typeset instructions throughout the rating period.

Since they were given both sets of evaluation material at the same time, raters were free to score the two texts in whatever order they chose. Although they had already undergone practice sessions, one could by no means rule out the possibility that further practice effects might affect their scoring of whichever text they scored second. Had we ensured that half the scorers received one text first, and the the other half the other first, any practice effects would tend to cancel out when scores where averaged out across the evaluator group.

> *Throughout the design of an evaluation, ensure that a balanced presentation paradigm is used wherever appropriate to control for order effects.*

On receipt of the material (about 5 minutes after their completion of intelligibility scoring), raters were invited to read the Accuracy instructions at their own pace and then to score the material for Accuracy at their own pace. Once again, raters were able to refer to the typeset instructions throughout the rating period and were free to score the two texts in any order.

Rating for Accuracy and Intelligibility was generally conducted in one session lasting about 1.5-2 hrs in all. All raters were paid *£5* or £10 depending on whether they scored for Intelligibility and Accuracy or Intelligibility alone.

[4] We did not feel it necessary to conceal the source of the translations. Indeed, this would hardly be possible, given their occasionally bizarre content compared to HT texts.

No rater scored more than one translation of a given sentence for Accuracy or Intelligibility.

5 Difficulties in Interpreting Declarative Evaluations

When evaluating anything, it is best to evaluate it with respect to the competition. We did not have access to any comparable MT system; we hoped that, nonetheless, some graphic presentation of translation quality achieved by GTS alone would be informative. However, even if one does have the opportunity for comparing several MT systems in a declarative evaluation, it is far from clear that the results will be readily interpretable by potential users. There are two main problems:

Which things cost more to Fix? ALPAC-style scale point descriptors mention many factors which help to degrade score performance e.g. “poor style”, “poor word choice”, “incorrect grammatical arrangements”, “syntactic arrangement”, “critical words untranslated”, “stylistic infelicities”. Consider a potential system usage in which all output translations are post-edited; will an “incorrect grammatical arrangement” be more difficult (time-consuming) to edit than a “poor word choice”? This is likely to be a problem however restricted the language used for scale point description.

Furthermore, the term “incorrect grammatical arrangement” covers a (very great) multitude of potential sins; we don’t know *without further data* whether correcting (say) systematic gender agreement deficiencies is more or less difficult than correcting (say) systematic deficiencies in the selection of modal verbs. In short, one doesn’t know *without further data* whether translational deficiencies which contribute equally to degradations in some quality score contribute equally to post-editing costs.

Which performance curve? Perhaps the most important presentation of the results of an ALPAC-style study is a series of bar charts plotting mean intelligibility rating (abscissa) for all the sentence samples against frequency(ordinate) for each of translation treatment. (As mentioned, in the present study only one MT system was being evaluated; normally the same procedure would be applied to a set of competing products.) Such graphs give a picture of the spread of perceived quality (on the intelligibility dimension) within each translation.

Faced with a set of such charts, how should a user choose between competing MT systems? Suppose one system has a monomodal plot (one hump) and another has a bi-modal plot (two humps); the significance of this difference for a given usage is completely unclear. Alternatively, monomodal MT plots might centre on the same average score but differ in their deviation; the significance of such differences remains obscure.

Sufficient experience linking declarative scores with observed system use would perhaps in time allow rational interpretation of such scores; however, the necessary experience is unlikely to be available.

So quite apart from the difficulties in constructing suitable scales and the problems of designing appropriate scoring procedures, declarative evaluation offers the potential user little in the way of concrete purchasing guidance. For this reason, we are now investigating in some detail the ways and means of conducting a well-principled operational evaluation of MT systems. However, we should make it clear that operational evaluation is not likely to be easy either. To obtain a measure of the improvements in translator productivity offered by a particular system, it is necessary to compare material produced by pure HT or MT+PostEditing which has been produced by translators of comparable quality: either one grades participating translators in advance, or one lets the same translator translate the same source material by both methods (in which case somehow or other the rather long-term practice effects must be balanced out or neutralised), or one tries to evaluate translations of different-but-similar source texts produced by the same translator using each of the two methods – highly problematic. Note also that we must take into account the different working speeds of translators, their different abilities in tackling post-editing and their different competences with respect to different text types. We fully expect a model operational evaluation to demand considerable evaluator and experimenter resources.

6 Conclusions

Our experience as MT research scientists evaluating a small-scale commercial system has made us painfully aware of two particular points:

1. In the absence of an off-the-shelf procedure, the design and implementation of any evaluation procedure is likely to require considerable resources

2. Any variety of declarative evaluation (or the use of test suites) probably has a very limited usefulness for the potential user.

Conducting *any* sort of user-oriented evaluation is likely to require design and experimental skills which are not, perhaps, routinely found in the intellectual armamentarium of the average Computational Linguist.

References

(1) Essex MT Evaluation Group (1991), The Globalink Translation System, *Personal Computer World,* 162-166, January 1991, London.

(2) M. King & K. Falkedal (1990), Using Test Suites in Evaluation of Machine Translation Systems, *13th International Conference on Computational Linguistics,* Helsinki, 211-216.

(3) Dan Flickinger, John Nerbonne, Ivan Sag & Tom Wasow (1987), Toward Evaluation of NLP Systems, *Ms delivered at Session of 25th Annual Meeting of the Association for Computational Linguistics.*

(4) Margaret King (1989), *A Practical Guide to the Evaluation of Machine Translation Systems,* ms, ISSCO, Geneva.

(5) John Lehrberger & Laurent Bourbeau (1987) *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation,* Amsterdam, John Benjamins.

(6) John R. Pierce & John B. Carroll (1966) *Language and Machines – Computers in Translation and Linguistics (Alpac Report),* Washington D.C.

(7) G. van Slype (1982) Conception d'une méthodologie générale d'évaluation de la traduction automatique, *Multilingua,* 1(4), 221-237.

(8) Muriel Vasconcellos (1989), Long-term Data for an MT Policy, *Literary and Linguistic Computing,* 4(3), OUP, 203-213.