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

Evaluation of MT systems
A programmatic view

Steven Krauwer
Research Institute for Language and Speech (OTS)
University of Utrecht

1 The Human Translator Metaphor

Most MT systems seem to aim at simulation of the behaviour of human translators, although this is not very often stated as explicitly as in e.g. [JW87, page 136]). Like human translators they translate texts normally not written with the specific goal of translation in mind and just as in the case of human translators their translations will normally be revised prior to distribution to the end user.

This predominance of what we will call the *Human Translator Metaphor* has – as a natural consequence – led to a view of MT evaluation where the main purpose of evaluation is to determine to what extent the makers of a system have succeeded in mimicking the human translator.

Yet it has to be noted that this interpretation of the notion of MT system is the one least likely to be successful over the next 10 or 20 years. There has been no real breakthrough over the last decade, and there is little evidence that the next 10 years will present us with a revolution in MT.

It is very tempting to object that progress *has* been made, and that the products that are available *are* dramatically better than the ones in existence some 10 years ago, but on closer inspection one will see that the main progress has been made in technology, i.e. that what seems to be an increase of MT quality is nothing but a gain in performance or capacity (faster hardware, better price/performance ratio), enabling a dramatic increase in size of grammars and dictionaries.

This type of progress may be attractive for a while, and in some environments even longer term cost effectiveness may result, but in the end it will turn out that the actual translation quality will hardly improve, and that after attaining a local optimum the quality of such systems will decrease by the simple fact that the addition of more and more lexical and grammatical distinctions will lead to more and more ambiguities not resolvable with the means currently at our disposal. That is, in the end all systems will run into the same brick wall, and after a relatively short period of improvement collapse under their own weight.

We certainly don't want to be pessimistic about the ultimate outcome of the MT enterprise, but an enormous research effort will be necessary for MT of this type, i.e. based on the human translator metaphor, to become feasible.

At the same time one has to observe that there are indications that new directions in MT are under exploration, directions that are much more promising given the current state of our knowledge. The main lines here are (a) systems of limited scope and/or ambition (which is by no means a totally new concept, cf TAUM-METEO), and (b) systems with limited translation tasks integrated in bigger environments (multilingual document production and revision, information distribution).

2 Current evaluation practice

2.1 Gear box evaluation

The MT evaluation literature seems to be entirely focused on the Human Translator Metaphor. The (correct) observation that an MT system is just a metaphor (rather than a copy) of the human translator leads to a situation where much of the evaluation effort goes into trying to determine to which extent the metaphor fails to be a copy, and the result is measured in terms of the effort needed to compensate for this (cf [Tuc87, page 28]). One could call this the Bad Metaphor view – the MT system as a bad imitation of the human translator.

Even if it can be argued that for a potential end user, who is faced with the choice between employing a human translator and acquiring an MT system, this is the only thing that really matters, it should be noted that as a consequence most of the criteria adopted are rather peripheral to the MT problem as such, and focus on e.g. the ease of post-editing, dictionary extension and debugging, and the irritation of the post-editor.

The danger here (especially for the MT research community) is that such criteria are very likely to be taken over by funding authorities and – as a natural consequence of this – by the system designers and developers themselves.

A natural (more puritan from an MT point of view) reaction to this is to go to the other extreme, and evaluate an MT system in isolation, and find other measures to check to what extent this system's functional behaviour is identical to the translator's (the Good Metaphor – how good a translator is the MT system).

The danger here is of course that it is not quite clear how to measure this, and that for the time being the result will always be very disappointing.

Following the tradition to explain evaluation methods in terms of *X box* evaluations we introduce another type of evaluation, *gear box evaluation,* to illustrate this.

Suppose we have just invented gear boxes for cars, and that we want to evaluate our products.

Since the potential customer wants a car rather than a gear box, the Good Metaphor approach in evaluation would most probably lead to an experimental setup where (under well-controlled weather conditions) a car (without gear box) and a gear box would be placed next to each other on the road, with one driver in the car, and one on top of the gear box, after which a number of objective measurements would take place (e.g. speed, behaviour in curves etc etc). It requires little imagination to predict what the outcome would be.

Alternatively we could adopt the Bad Metaphor approach, and take a gear box, build a car around it, and compare the whole car with another car without a gear box, which is necessarily different in design and construction. Even if this seems to make much more sense than the previous approach, it has to be noted that under such circumstances there is hardly any guarantee that what is actually measured is the contribution of the gear box to the behaviour and performance of the car, rather than the quality of the seats, the radio, the length of the gear stick etc etc. This is not to say that it would be absolutely impossible to make objective, comparative measurements, but it is obvious that this method may mislead – not only the potential customers, but also the designers themselves.

In either case we are handicapped by the fact that our invention is only capable of a partial contribution to the overall functioning of a car, and evaluation in isolation is extremely difficult, and totally irrelevant from the end user's point of view. We will call this problem the *gear box evaluation syndrome.*

The lesson to be learned from this is that what is needed in the first place is a set of methods to evaluate all and only properties of the gear box, in addition of course to the evaluation of the interaction between the car and its gear box.

If we apply this to MT systems it should be clear that we will have to find methods to evaluate all and only the properties of the MT system proper, i.e. the component that performs the actual translation. We will come back to this in section 3. At this point we would like to stress that the existence (and general acceptance) of such methods would not only have direct practical relevance for the system designer, but would also have a beneficial effect for the MT research community in the sense that the present situation inevitably suggests to funding bodies (be they public or industrial) that investment in e.g. more sophisticated windowing facilities or touch-screens for the post-editor would be more profitable than investment in MT proper.

2.2 Evaluation as a discovery procedure

Another specific feature of current MT system evaluation practice is that a large amount of time is needed to discover what exactly the functional behaviour of the system is. One possibility is of course to ask the designer to provide this information, but it turns out that in many cases this information is not obtainable (as a matter of fact probably does not even exist), and the evaluator is forced to feed the system with vast amounts of text, in order to be able to make hypotheses concerning the system's functionality ([LB88, page 136]).

The obvious question to ask is whether this is reasonable at all: the vendor offers a product without telling (or being able to tell) what this product does, thus leaving it to the potential customer to figure out what exactly the product is, and whether it suits his needs.

This is a very peculiar situation, applicable to hardly any other product of modern (or even old-fashioned) technology.

How can this situation be justified or explained? The main problem here is that the whole concept of translation is still very ill-understood. Even if there are many people who are capable of translating texts between various languages, very little is known about the nature of the relation between source and target language texts established by the translator. This makes it very difficult to describe the (probably very partial and imperfect) functional behaviour of a translation system. To come back to the car metaphor: if someone produces an object which he claims to be a car, it is normally rather easy to verify whether or not this claim is valid, by checking the properties of this object against a list of known properties of cars (some of which are even prescribed by law).

In the absence of an explicit theory of what translation is, the system maker has no mechanism to provide an explicit description of what his system does and this leaves the evaluator completely in the dark. He has no basis for comparison, and there is little he can do except trying to figure out how the system behaves in a number of cases known as translation problems to him. And this leads to a lengthy and expensive discovery and evaluation procedure – to be carried out by the evaluator.

3 Directions to take

In order to arrive at some sort of workable notion of evaluation of MT systems we will have to create our own fixed point for comparison and evaluation.

The most preferable option would be to develop a formal theory of MT, which would provide us with a formal definition of the tasks an MT system should perform. Although we would strongly recommend such activities to be carried out, we don't believe that this will lead to any directly usable results in the immediate future.

If the fixed point that is required for proper evaluation cannot be obtained by a formal theory describing what translation is, we should look for alternatives.

It is beyond the scope of this paper to explore the whole universe of possibilities here, and we will concentrate on what could be gained by aiming at a reduction of the translation task. Thus, possibly promising approaches such as evaluation based on huge corpora of translated material will not be addressed here.

One possible option is to reduce the problem to sublanguage translation (as opposed to unrestricted language translation). At first sight this looks very attractive, but unfortunately there seems to be no clear evidence that a theory of sublanguage translation is essentially more feasible than a theory of general language translation, and hence the fixed point for evaluation cannot be expected to be easier to establish (even supporters of the sublanguage approach admit that the number of sublanguage-based systems is small, and that a lot of basic research remains to be done [Kit87, page 64]).

The option that we want to advocate here is that the fixed point is given apriori, in the form of a full specification of what the system does. This specification should be seen as part of the MT system, and thus the system becomes a pair <Specs,Device>, where Device stands for the system proper (i.e. the thing that actually runs on a computer), and Specs is the functional specification of the Device, provided by the system maker.

We will discuss some advantages and problems connected with this approach in the following sections.

3.1 Advantages

The existence of a <Specs,Device> pair creates the possibility of subdividing the evaluation of a system into two separate evaluation activities, which cover entirely different aspects of the system, but which – to our knowledge – are never separated in evaluation discussions.

The first one is to check the conformity of the Device with its Specs. If the Specs are well-defined (i.e. if there is a clear explicit statement of what the system is supposed to do), it should be possible to check empirically to what extent the Device actually implements the Specs. This should allow for objective testing, both by the system designer/implementer and by e.g. potential customers or research programme funders.

The other one is to check to what extent the Specs match the user's requirements. Note that the Specs are the basis for this evaluation, rather than the actual Device. This type of evaluation refers to the system's usefulness (from the user point of view), and not necessarily to the system's inherent qualities.

This decomposition of the evaluation has the obvious advantage that there is a fixed point of reference, which can be used for both conformity and usefulness checks.

Note that conformity alone does not necessarily imply usefulness: one can think of lots of systems that are fully specified, and correctly implemented, but which do nothing useful. But in this case the availability of the Specs would still help the user in making a preselection before he even starts worrying about conformity.

3.2 Problems

There are a number of problems connected with this approach. We will discuss the most notable ones here, in the form of a list of questions and answers.

(i) Isn't it unfair to require that the system maker does not only make a system, but has to provide the specs as well?

No, it is not unfair (although it may be far from trivial), since such specs ought to exist anyway (in order for the system designer to know what he is doing, unless the system is nothing but a piece of arbitrary hacking), and it is totally unfair to require that everyone else has to figure out for himself what exactly the system's capabilities are.

(ii) Isn't it unfair to require that the system maker reveals all his design secrets to third parties?

Such a requirement would be unfair, but what is needed is not a complete blueprint of the system, with all the technical refinements, but rather a specification of the functional and performance behaviour of the system as intended (and in the end guaranteed) by the designer, i.e. there is no need to open up the black box, as long as the functionality is properly described.

(iii) Isn't the production of specs as difficult as the production of a system, and isn't this approach to evaluation as unrealistic in the light of what has been said earlier about the feasibility of MT?

Yes, this is absolutely true. If we cannot reduce the complexity of the translation task as such dramatically, this whole exercise will remain vacuous. And here we touch upon the central point of our proposal, namely that the Specs should not only play a central role in evaluation, but they should be taken as a starting point for the definition of the translation task as well. Once one has accepted the fact that the translation task as a whole is too complex, and that sublanguage based approaches are not significantly simpler, the next step should be to define one's own notion of translation task, in such a way that the result is both implementable and useful. That is, the designer should offer something that he trusts he can specify and make, and the usefulness of this product will depend on his negotiations with the user concerning adaptations to the user's specific needs. We expect that the typical result of such negotiations will be a controlled language approach, where the maker can guarantee the proper behaviour of the system, as long as the user sticks to the rules. This seems to be fully in line with e.g. Church & Hovy's desiderata for a good niche MT application [CH91, page 149].

(iv) What do proper specifications look like?

This is probably the most difficult question. There are no standard techniques for specifying MT systems, and we expect that different types of technique may be relevant for different classes of MT systems (note that the systems we refer to are specialized systems). Possible ingredients are:

—Explicit specifications formulated in some specification language.

—General principles that ensure predictability of the system's behaviour (e.g. compositionality, the degree of which can be objectively measured, cf [GACR91, page 141]).

—Test suites. Note that the A- and B-list approach suggested in [KF90, page 214] already comes close to this idea if one is prepared to go one step further and interpret the A-list as (part of) the specifications (to be provided by the maker). Note also that the standard criticism that these suites may lead to systems geared towards the suites themselves no longer holds, since this gearing could be interpreted as a instance of specialization, and the suite itself as part of the specifications.

Note that the choice between such ingredients does not necessarily express any claims concerning 'real' translation, but rather reflects the designer's views on what could be seen as a feasible and useful translation task. If for instance in his perception the translation task in a certain environment can be redefined such that all the expressive power necessary can be obtained by means of an exhaustive list of primitive constructs and a number of compositional operations defined on those, he may be able to satisfy the translation needs of a user – who has to be prepared to express himself in terms of exactly those primitives and operations.

As for the specification language it should be clear that no such language exists, and it would certainly be a research enterprise to design one. The reason why we feel that such an enterprise would not be unfeasible is that one could do with a variety of such languages, geared towards the specific requirements of classes of specialized MT systems, rather than with one single specification language that would do the job for all possible MT systems.

4 Conclusions

The main conclusions of this paper are

- The Human Translator Metaphor is not the most fruitful approach to evaluation (and MT systems in general), and one should rather go for specialized systems, to be negotiated between designers and users.
- One should be aware of the gear box evaluation syndrome when evaluating MT systems.
- MT systems should be seen as pairs <Specs,Device>, where the Specs constitute the fixed point for evaluation.
- Evaluation is both conformity checking between Device and Specs, and usefulness checking via user requirements and Specs.
- Research should be stimulated into specification methods and languages for specialized MT systems.

References

[CH91] Kenneth W. Church and Eduard H. Hovy. Good applications for crummy machine translation. In *Natural Language Processing Systems Evaluation Workshop,* 1991.

[GACR91] Björn Gambäck, Hiyan Alshawi, David Carter, and Manny Rayner. Measuring compositionality in transfer-based machine translation systems. (Also in this volume) In *Natural Language Processing Systems Evaluation Workshop,* 1991.

[JW87] Roderick L. Johnson and Peter Whitelock. Machine translation as an expert task. In *Machine Translation: Theoretical and Methodological Issues.* Cambridge University Press, 1987.

[KF90] Margaret King and Kirsten Falkedal. Using test suites in evaluation of machine translation. In *Proceedings of COLING-90,* 1990.

[Kit87] Richard I. Kittredge. The significance of sublanguage for automatic translation. In *Machine Translation: Theoretical and Methodological Issues.* Cambridge University Press, 1987.

[LB88] John Lehrberger and Laurent Bourbeau. *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation.* John Benjamins, 1988.

[Nir87] Sergei Nirenburg, editor. *Machine Translation: Theoretical and Methodological Issues.* Cambridge University Press, Cambridge, 1987.

[NW91] Jeanette G. Neal and Sharon M. Walter. Natural language processing systems evaluation workshop. Technical report, Rome Laboratory, 1991.

[Tuc87] Allen B. Tucker. Current strategies in machine translation research and development. In *Machine Translation: Theoretical and Methodological Issues.* Cambridge University Press, 1987.