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

Constructive Machine Translation Evaluation

Stephen Minnis
Mitsubishi Electric Corporation

Abstract

It is now acknowledged that the evaluation of the quality of MT output is inextricably linked to the purpose to which the translation output will be put. It is also true to say that the value of the evaluation is inseparably linked to the purpose to which the evaluation results will be put.

For the developer of an MT system, the evaluation of the quality of the MT output must be approached from the viewpoint of increasing the knowledge about the MT system. Resultant measures must be analysed, so that practical feedback to improve the system is feasible. However, for the manager, evaluation of the MT output is often viewed in terms of comparison. Measures are compared against previous measures, or against those obtained by other systems, in order to gauge progress, or to assess the systems ability.

Measurement can be viewed as a tool for increasing the knowledge of some object or entity. From both viewpoints described above, the measurement is required as a means to increase the knowledge about the system, whether it is knowledge about the systems errors, or performance.

However, there is a more fundamental level at which measurement should be applied as a tool for increasing knowledge; that is, to increase the knowledge of the properties we are trying to measure (in this case intelligibility and fidelity). Such measurement is a precursory requirement for more general uses of evaluation measures, as described above.

When surveying the many methods currently employed in MT evaluation, it is not immediately obvious that the methods used serve to increase the knowledge of the properties being measured. This report describes a *constructive* machine translation evaluation method, aimed at addressing this issue.

Introduction

The move towards the use of measurement is sometimes justified by quoting the physicist, Lord Kelvin [Cook 1982] (see figure 1).

The premise is that measurement serves as a tool for increasing our understanding or knowledge. There is however, a danger in attempting to measure an entity when that entity is not fully understood, as it is possible to measure incorrectly (see figure 1, second quote [Hamer 1985]).

Measurement as a tool for increasing Knowledge

Lord Kelvin

> "When you can measure what you are speaking about, and express it in numbers you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.."

George Miller

> "In truth, a good case could be made that if your knowledge is meagre and unsatisfactory, the last thing in the world you should do is make measurements. The chance is negligible that you will measure the right things accidentally"

Figure 1

Current MT evaluation methods do not immediately lend themselves to increasing the knowledge of what we are measuring. As an example, do the methods for measuring the intelligibility serve to increase the knowledge of what makes a sentence intelligible? In fact this statement begs the question 'what is intelligibility?'.

If this issue is not addressed, then evaluation techniques centering on measuring intelligibility and fidelity can only be steeped in subjectivity. These vague notions will always mean different things to different people. Furthermore taking 'measures' by rating sentences on a subjective scale (cf ALPAC and other evaluation methods) only adds to this subjectivity. In order for MT evaluation to progress from this predicament a more constructive machine translation evaluation method is required.

A constructive evaluation method is one in which the evaluation method is applied in a manner which increases the knowledge about the translation system (from a variety of viewpoints), in addition to enhancing our knowledge of the properties being assessed (intelligibility and fidelity).

To develop this evaluation method, the approach taken in this document is to (briefly) review basic measurement theory, and then attempt to apply the important concepts to MT evaluation. The end result is an evaluation method which seems to offer some advantages over current evaluation methods.

Basic Measurement Theory

Measurement is one mechanism we use to describe particular properties of entities around us. Statements such as ‘I am taller than you’, ‘My pen is redder than yours’, ‘I am more knowledgeable than him’ etc. are all based on some underlying theory of measurement. An informal definition of measurement (given in [Finkelstein and Leaning 1984]) is shown in figure 2.

To illustrate by way of example, if ‘people’ are taken as the object of interest, we observe that some people are *taller than* others. Therefore a property (or attribute) that people possess is height. When we assign numbers to peoples height, the relationship ‘taller than’ is preserved in the measurement chosen.

The assignment of numbers requires standard procedures of measurement, and standard units representing the measurement.

Basic Measurement Theory

Definition:

> “Measurement is the process of empirical, objective assignment of numbers to the properties of objects and events in the real world in such a way as to describe them”

Example:

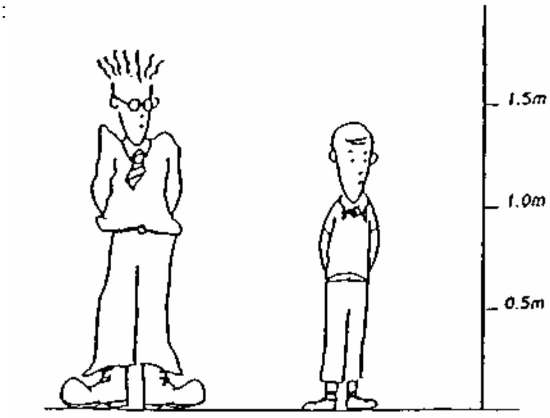


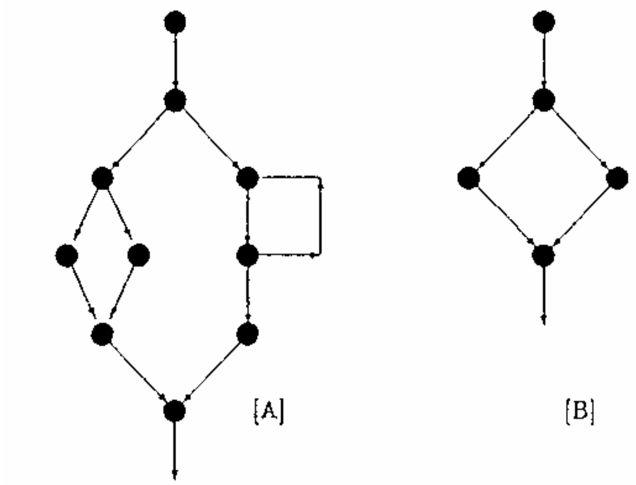
Figure 2

For example, if we measure a person standing upright against a wall, using the metre scale, we can state that a person who is 1.6m tall, is taller than someone who is 1.4m tall. Therefore the observed property ‘height’ is captured by the measurement (as 1.6 > 1.4). Note that the procedure of measurement should be more strictly defined, for example, to take into account whether shoes are worn, whether hair height is counted and so on.

This simple description of the application of measurement must be revised when the object[1] to be measured is very complex. In this case the usual approach is to simplify the object of interest, or at least some particular feature of the object, in the form of a model.

As an example, suppose we want to measure the ‘complexity’ attribute of a software program. It is obvious that many factors will affect this complexity, including the control structure, the data structure, and the length of the program. To measure the control structure, one approach could be to develop a graphical model of the control flow, and measure the structure of that graph[2] (see figure 3). The assumption is that the structure of the graph has an intimate correlation with the ‘complexity’.

Using Models to Abstract Properties



The control structure of [A] is more ‘complex’ than that of [B]

Figure 3

There are essentially two approaches to modelling; the first one, demonstrated above, is the structural approach, where the problem is decomposed

[1] All references to objects in the remainder of this document, should also be read as applying to events

[2] This approach has been taken in the field of Software Measurement [McCabe 1976, Fenton-Whitty 1987]

into manageable components, eg. the control structure is one component of the overall complexity. The second approach is statistical in nature, where many aspects of the object (eg the code) are measured (eg its length, number of goto statements etc.), and these are utilized in a statistical model. Typically, statistical models are used when the object/property being measured is not well understood. In practice, often an amalgamated model is developed, employing features of both these approaches.

Both types of models are useful, but all such models must be validated.

Validation is when something can be confirmed as relating to reality, usually shown by objective and repeatable experimentation. Some of the statements given at the start of this section can be validated, eg the statements relating to height and colour, however knowledgeability is more difficult to validate. It might be possible to validate some aspects of knowledgeability, by using particular models which describe knowledgeability in *measurable* terms. As an example, the winner of a mastermind quiz about 'Chinese sauces used in Mongolia between the years 1912-1918' might be considered more knowledgeable than the losers (assuming they get the same questions etc.), but strictly speaking the measurement only shows that the winner was more knowledgeable on that topic, over that particular set of questions. It does not constitute a validation of a wider notion that the winner is more knowledgeable than the losers[3].

Clearly, interpretation of the results is important, and must be done carefully with full analysis of the implications of the measurement method employed. This is particularly important when the property being measured is complex and abstract. It should also be clear that there is an advantage in defining vague attributes in measurable terms, in order that we can draw some conclusions about them.

Usually once a model is developed, some aspects of features of the model are used to hypothesise or explain certain relationships observed for attributes of the object. For example in the program complexity example, it has been hypothesised that the nesting of control loops increases the complexity of the control structure. The measures used attempt to capture this increased 'structural complexity' by assigning appropriate weighting to nesting in the measurement employed.

To summarize, the tasks in the application of measurement are shown in figure 4.

It is important to remember that each attribute of interest will possess its own specific assumptions about the object or model, and will therefore

[3] If the set of questions in the quiz was always the same, then eventually everyone would become familiar with the answers. The initial quiz then becomes useless (cf MT benchmark corpus evaluation?)

require its own specific metric, or measurement procedure.

Applying Measurement Theory

- * to define the object or event of interest (this may be a model)
- * to establish some intuitively observed property of that object or event
- * to define the standard measurement procedures, rules and units, in order to capture the relationship that the attribute imposes on the object or event

Figure 4

Measurement of MT Output

The three steps in the application of measurement are approached in the following manner. First, the attributes of interest are introduced, as they have more or less already been decided by work in MT evaluation over the past two decades.

After this the object of interest, and how it can be modelled is discussed. This step will necessarily include a description of how the attributes can be defined in measurable terms, on the particular model chosen.

Note that in order to restrict the scope and length of the paper, there has been little attempt to pursue, in detail, the third (and most difficult) task, that of defining the standard measurement procedures, rules and units. The tentative approach presented, aims to outline only the main features of the proposed measurement procedure.

Identification of Intuitively Recognised Properties

The choice of the two attributes of interest, intelligibility and fidelity, is dictated by the work in MT evaluation over the past few decades.

The attribute of intelligibility naturally comes from the intuitive observation that some machine translated texts are more intelligible than others. We want our resultant measures to represent the relationship ‘more intelligible than’. A text with a higher intelligibility rating will be more intelligible than one with a lower rating[4].

[4] Note that a measurement will have appropriate units and scales. We can say that a person of height 2m is twice as tall as someone who is 1m tall. For intelligibility and accuracy, at least initially, the scales will be arbitrary, and we will not consider such aspects.

Similarly, the fidelity of a translation arises from the observation that some translations are more ‘accurate’ than others. The notion of fidelity may be more difficult than intelligibility. This is because intelligibility is restricted to one cultural and linguistic domain, ie for one language. Fidelity must be assessed by a bi-lingual, and hence is more open to interpretation, as the knowledge of a second language is not often accompanied by a full understanding of the languages inherent culture and usage. The correspondence a bi-lingual makes between the two languages depends on this essential knowledge, and increases with experience. Also, often a literal translation does not convey the full ‘meaning’ of the source language.

By restricting the domain of the language under translation, for example to a particular technical domain, then the problems involved with bi-lingual assessment may be reduced in scope[5].

Definition of the Object of Interest

For MT evaluation, is the object of interest the translated text, produced from the input text, or is it the translation process we are really interested in evaluating? This distinction is discussed within the framework of the model shown in figure 5.

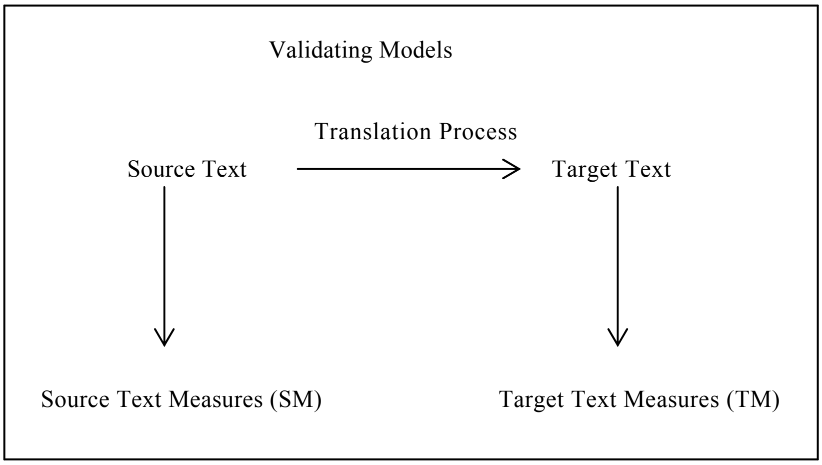


Figure 5

In the case of the intelligibility attribute, the object of interest is in fact only of the target text *product,* ie measure TM; (Target Measure of intelligibility). This measurement can be extracted in isolation to any knowledge about the translation process or of the source language.

[5] Actually, decreasing the scope may increase the problems due to more specific nature and knowledge required of that domain.

In the case of fidelity, what we want to measure is actually the relationship of some measure on the target text, TM_f, to some measure on the source text SM_f. This means that the fidelity measure is actually capturing some aspect of the translation *process*.

The measures TM_i and TM_f are not equivalent. The fidelity is not a case of relating the intelligibility of the target text, to the intelligibility of the source text.

To validate these measures, it is necessary to show that the measures on the text correctly capture the relationship being measured, ie that the numbers assigned reflect the relationship observed. For accuracy, in addition, the relationship between the source and target text measures must be validated.

What exactly is the text 'object' of interest? The domain of interest for MT is written text, therefore possible text 'objects' could be paragraphs or sentences, for example. It would be possible to discuss the intelligibility or fidelity of paragraphs, or even larger 'objects', but the next assumption is that our object of interest is the sentence in isolation. This is because most current machine translation systems translate on a sentential basis, and therefore the measurements should be trying to capture the effectiveness of this intra-sentential translation.

This is a practical assumption that reduces the complexity of the evaluation task, as it should be easier to assess the sentence in isolation.

However it is apparent that even for a sentence considered in isolation, it is not immediately obvious what the intelligibility or the fidelity will be, as we do not know exactly what these properties are. This is probably a terminology problem; if we continue to talk about the intelligibility defined as 'the ease at which the meaning of a sentence can be understood' and then proceed to develop scales to measure this (eg ALPAC), we get nowhere, as we never increase our knowledge of the attribute, and furthermore the measurements and results are always steeped in subjectivity.

The sensible approach seems to be to define the attributes in measurable terms, so that thereafter we can reason about them objectively. Furthermore, we then have the means to validate any models we may propose. This provides an opportunity to progress in increasing our knowledge of what affects/constitutes the attributes, and only then can we propose scales, such as those that have been suggested in other methods.

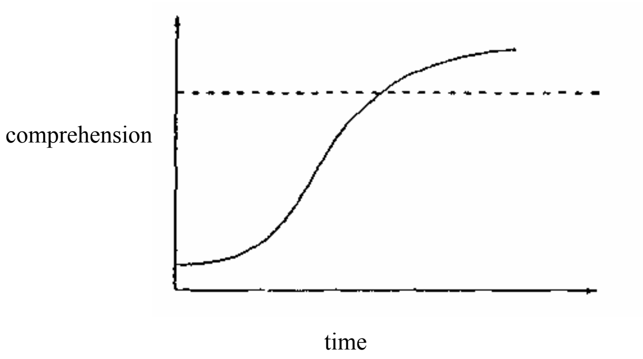
Two approaches are suggested, both aiming towards this goal. They are both directed at implementing objective and repeatable measures.

The first is to define the intelligibility/fidelity attributes in terms of comprehension time. In figure 6 the dotted line represents the threshold value at which the subject has grasped the meaning of the translated text[6]. Tests would need to be carried out periodically to show that the correct meaning

[6] This threshold value being assigned as a standard measurement procedure.

was read. Without going into such an experiment in detail, it is obvious that it would take some time to organise, but with careful design, it is possible that the results could provide useful measures. For example, the rate of understanding could be studied (the gradient), the area under the curve might have some significance, as well as the time to reach the threshold. Such aspects could provide information as to the capabilities of the subjects (if more than one was used), for example for normalization purposes.

Defining Attributes in Measurable Terms: Approach 1



measurable features

* rate of comprehension (gradient of curve)

* area under curve

* time to reach cut-off point

Figure 6

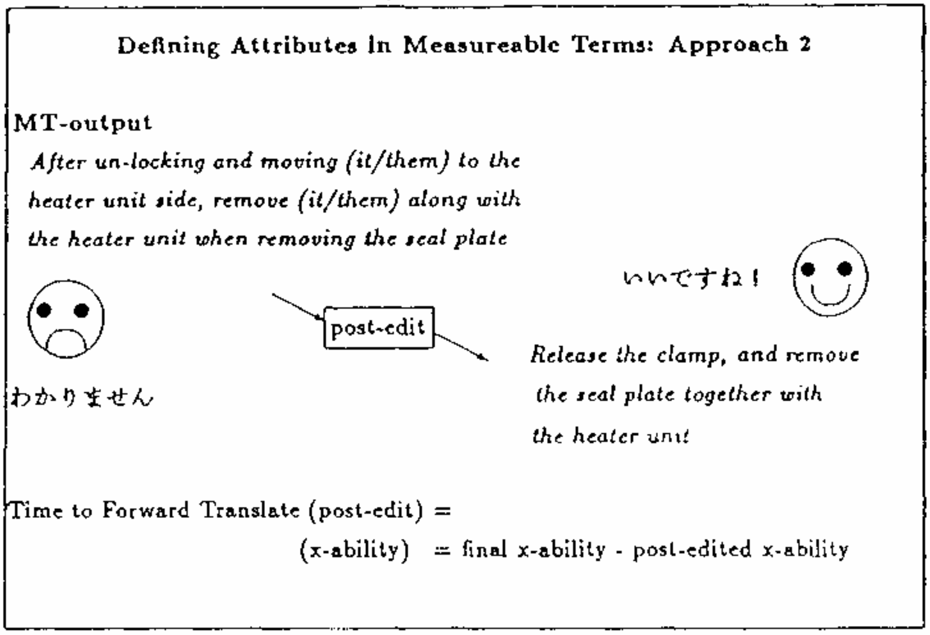
For accuracy, the same method could be used, except with reference to the original text. The method is similar to the informativeness measures proposed in [King], except it tries to be more objective in the measurement (time is better than simply assigning an arbitrary number).

This approach does suffer from the resources it requires, as well as practical difficulties, for example, in showing text comprehension, especially for machine translated sentences.

The following approach is more practical, but is an indirect approach. It is based on the method of ‘forward translation’, or for practical purposes, post-editing. In case of intelligibility, the time taken to make the sentence intelligible (by post-editing) could be used. For fidelity, the time taken to make the sentence accurate (with original text for reference) could similarly be used[7]. In this case we are in fact measuring the difference in the x-ability

[7] This raises the question as to whether it is possible to post-edit only for accuracy. In fact, 'real' post-editing is geared to producing accurate and intelligible output. The

of the pre-(post-edited) text and the final post-edited version (see figure 7[8]).



Defining Attributes In Measureable Terms: Approach 2

MT-output

After un-locking and moving (it/them) to the heater unit side, remove (it/them) along with the heater unit when removing the seal plate

わかりません

post-edit

いいですね！

Release the clamp, and remove the seal plate together with the heater unit

Time to Forward Translate (post-edit) =

(x-ability) = final x-ability - post-edited x-ability

Figure 7

This approach could be verified by using a variety of post-editors to confirm the final translations are understandable/accurate[9]. Although different translators might produce a slightly different end-translation, agreement could be made as to whether the final translations are acceptable (perhaps by using guidelines etc.). The time taken could be averaged to give a final value, if more than one post-edit is done.

In both approaches described, there is a need to standardise the measurement procedures used. Post-editing time might be affected by a variety of factors, including tool use etc. not to mention the post-editors ability. These topics are not discussed any further in this report.

Since fidelity is the relation of accuracy of the target text to the source text, the assumption is that when the text is post-edited for fidelity, the relationship is 1:1.

For intelligibility, since the post-editing is done in isolation by a monolingual, the resultant text might be widely inaccurate. This is not a problem from the measurement viewpoint. In fact the difference between the post-edited intelligibility version and the post-edited fidelity version, will indicate

difference between the post-editing time for intelligibility and the post-editing time for intelligibility/accuracy could provide an acceptable accuracy measure.

[8] Example translation and post-edited text taken from [Nikkei 1990].

[9] Note that verification is not validation. Verification is showing that something is correct. In this case we want to show that the final post-edited text is correct.

the quality of the translation in conveying the required meaning correctly. Developing another measure to capture this difference would be useful.

To summarize so far, it has been proposed to measure the two attributes intelligibility and fidelity in measurable terms. The preferred method (for practical purposes) is one in which we measure the attributes indirectly, therefore our assumption is that the time measured correlates with the attributes of interest. This is indirect measurement. A similar example which helps to clarify the approach, is that we can indirectly measure the temperature of a room by measuring the length an iron bar expands. The expansion of an iron bar has been validated as being related to the temperature, through extensive experimentation, similarly, the above indirect approach must also be validated.

Constructive MT Evaluation Method

In the last section a definition for intelligibility and fidelity was proposed which was *measurable.* If we accept such a definition then we increase the objectivity in our reasoning about the attributes. This allows us to investigate what affects the attributes, and therefore increase our knowledge about them.

To increase understanding of an object, it was mentioned in the introduction that there are two approaches; to develop theoretical models, or to attempt to measure certain aspects, and see if we can deduce significant trends or patterns through statistics. As an example, if we wanted to measure the temperature of a room, we could develop a theoretical model relating the expansion of an iron bar to the temperature, taking into consideration such factors as the bars size and density. Alternatively at the other extreme, we could measure these entities and try to relate them to the temperature statistically. Typically a mixed approach is employed, where a model is hypothesised, but refined through empirical observation and measurement.

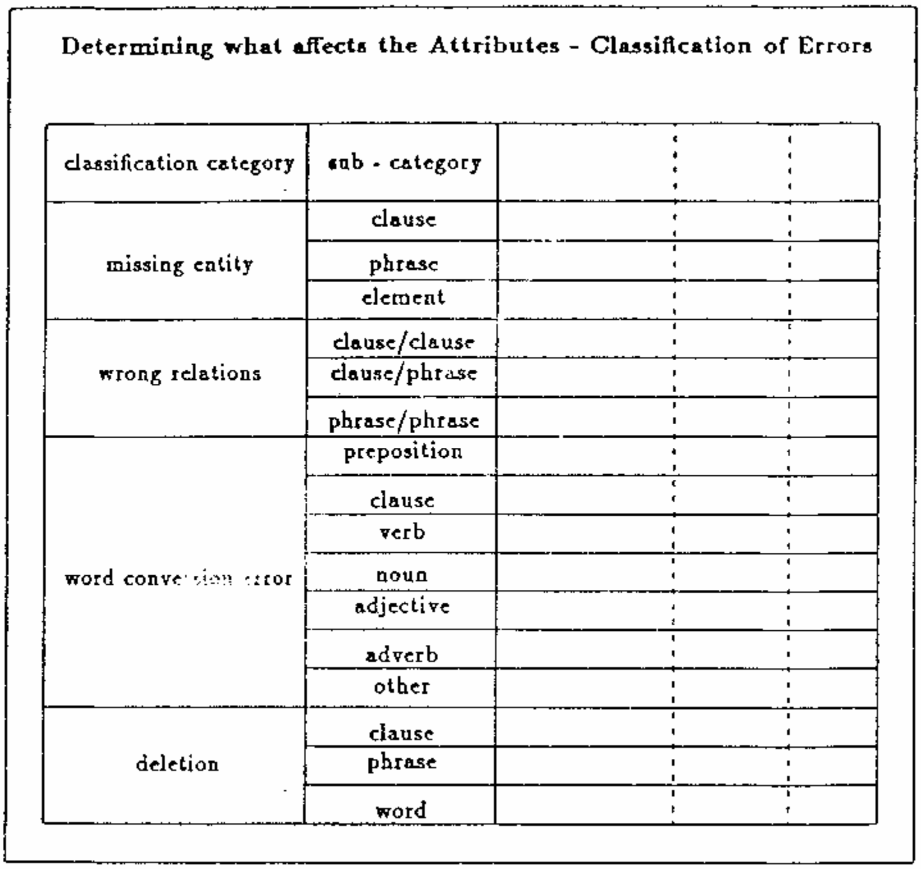
We have some intuitions as to ‘factors’ that might affect the attributes of intelligibility and accuracy, for example, missing words, incorrect sentence structure, wrong translation of verbs, and so on[10]. However, we do not know the inter-play between the factors affecting the attributes.

The following practical approach is proposed; to get the post-editor/evaluator to classify what is wrong with a sentence (see Figure 8[11]). Although it is possible that many evaluators will possibly classify different aspects as being wrong, this is only likely to happen when the translated sentence is

[10] Other possibilities might include style, grammaticality etc. but these are more difficult to detect in typical MT translated sentences. The above approach is deliberately kept simple.

[11] This classification being taken from [Nagao et al. 1988].

very unintelligible[12]. Note that we should be able to measure how bad the classifications differ between post-editors; this could serve as another measure/indication of the intelligibility/fidelity.



Determining what affects the Attributes - Classification of Errors

classification category	sub - category			
missing entity	clause			
	phrase			
	element			
wrong relations	clause/clause			
	clause/phrase			
	phrase/phrase			
word conversion error	preposition			
	clause			
	verb			
	noun			
	adjective			
	adverb			
	other			
deletion	clause			
	phrase			
	word			

Figure 8

Furthermore, even if different evaluators classify different aspects, as it is proposed that the post-editor must correct what has been classified (this time being measured), the classification will be ‘tested’. If the changes dictated by the error classification are not sufficient to make the sentence accurate/understandable, the results are discarded (or, more practically, the error classification is changed). Therefore we have a means for verifying the classification made by the post-editor. As explained before, showing that the final result is acceptable in terms of the attribute of interest, through agreement reached by a group of evaluators provides further verification of the measures.

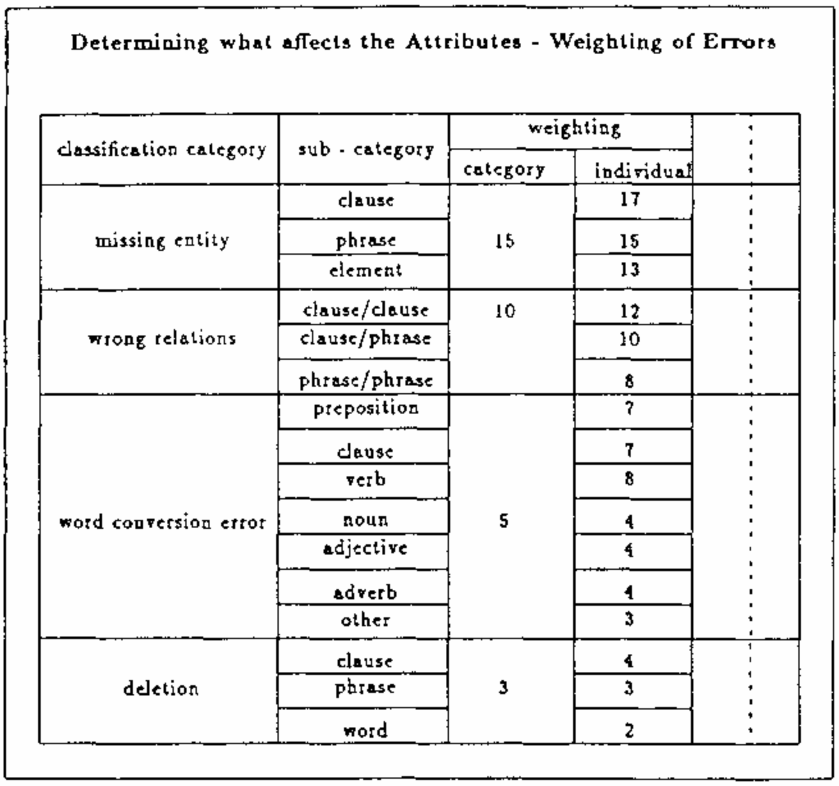
[12] The presumption is that the sentences are of a sufficient quality to be evaluated.

The classification shown in figure 8 has been deliberately left relatively high level, so that the method is practical. The idea is that the evaluator can decide the levels, and the detail required, as necessary. For example, a more detailed classification would be to divided the phrase classification into noun phrases and verb phrases.

At this point it is assumed that the time for post-editing has been recorded, and that the above error classifications have been verified as described.

Since we do not know how each of the above entities affects the intelligibility or fidelity, it is proposed that weights are assigned to each category.

These weights will be assigned values using our own intuition at the onset of the evaluation, but the aim is to *tune* them as to actual importance, using a database of collected measures. If we consider the indirect measure of temperature again, a theory relating the expansion to properties of the iron bar might have been developed in such a way. It is noted that the indirect measurement of temperature has far fewer variables, and is more amenable to measurement being a 'hard' physical quantity, than does a property such as intelligibility. Although the attributes of intelligibility and fidelity are complex, with experience it should be possible to identify the key factors affecting them.



Determining what affects the Attributes - Weighting of Errors

classification category	sub - category	weighting		
		category	individual	
missing entity	clause	15	17	
	phrase		15	
	element		13	
wrong relations	clause/clause	10	12	
	clause/phrase		10	
	phrase/phrase		8	
word conversion error	preposition	5	7	
	clause		7	
	verb		8	
	noun		4	
	adjective		4	
	adverb		4	
	other		3	
deletion	clause	3	4	
	phrase		3	
	word		2	

Figure 9

Different weights will probably have to be assigned for each attribute, as each attribute will be affected differently by different factors.

Note we could go further and sub-assign weights, for example a badly transferred verb could be assigned a weighting of 8, and an adjective of 4, as a verb is generally perceived as being more important for accuracy. However, it might be better to err on the side of caution, and not be too ambitious with regard to details, until key factors have been identified.

The weights do not have to be known to the post-editor. This will contribute to the objectivity of the classification and the post-editing process.

There is one final aspect of the measurement which we need to consider. This is the requirement to normalise sentences according to their complexity, so that the measures taken are put in perspective. This is necessary, because generally speaking, longer sentences will be more difficult to understand or be accurate, than shorter sentences. Normalization typically requires some quantification of frequency of occurrence. One possibility is to normalise within each error category. A second possibility is to normalise according to the number of verbs in the total sentence, as the number of verbs is generally seen as being related to the sentence complexity[13]. A more practical idea is to simply count the number of words in the sentence. A target language parser could be used to parse the post-edited sentence, and extract various information. The number of 'levels' in the parse tree could also be used as an indication of complexity (or even the parse tree itself – cf the earlier example on control flow program complexity)[14].

Assigning the final attribute rating is initially assumed to be a simple additive procedure. A more sophisticated model may be developed once suitable data is gathered. Alternatively, we could try a more detailed model initially using intuition based on empirical observations.

An example of tailoring of the weights is now given (normalization ignored), to show how the method might work in practice. Suppose we have the following four sentences, assumed all of the same 'complexity' (see Figure 10).

From this it is possible to see that some correlation exists between the resulting attribute score (**X**) and the actual time. In this case, one attribute rating is approximately 5.9 seconds. Therefore, we can say that if we get another attribute rating of 100, the time to post edit for that attribute will be approximately 590 seconds.

Although this example is too simple, it is hoped that weights can be tailored and roughly accurate models developed after a large database is

[13] Although, again, the question of standard definitions must be considered. For example 'the car is red' contains a verb 'is', whereas 'a red car' does not. Both could be valid translations. Also contrast 'the men destroyed the house' and 'the men's destruction of the house'.

[14] Again, would need to standardise the use of the parser employed for this purpose.

constructed.

The type of measure taken is useful in that it is meaningful to the system managers. To utilize the data results, statistical feedback can be easily extracted from the database. There is the possibility to integrate tools, graphical displays etc. for effective analysis and presentation of results. The developer can use the database to investigate which errors occur most often, and also, could identify those errors which cause most problems. It would then be possible to plan an improvement in the system. The evaluation is therefore playing a constructive role in the development of the machine translation system.

Tailoring Weights

example
word conversion error (WE) = 10
wrong relation error (WR) = 20

(1) —— 10 word errors ——
post-edit time 600 secs
(2) ——4 relation changes——
post-edit time 500 secs
(3) — 2 relation changes, 4 word errors —
post-edit time 550 secs
(4) — 1 relation changes, 7 word errors —
post-edit time 400 secs

10WE	= 100X	= 600 secs
4WR	= 80X	= 500 secs
2WR + 4WE	= 80X	= 550 secs
1WR + 7WE	= 90X	= 400 secs

1X ≈ 5.9 secs

Figure 10

Assessment of the Method

The evaluation method has the following advantages:

- it is amenable to automation, which is an important factor in evaluation, as it is such a tedious job.
- it has a certain degree of built-in verification
- it allows tailoring. If the weighting method changes, there is no need to re-do earlier experiments, merely to re-calculate the values

- the resulting measure is simple and meaningful. Furthermore the method yields data which is useful for a variety of purposes
- it facilitates validation of the measures proposed

It has (might have) the following disadvantages:

- need to collect data before measuring
- initial subjectivity in weighting, may be difficult to tune the weights, etc
- practical difficulty in classifying errors
- subjectivity involved with post-editing ability, needs to be addressed. The use of consistent practice and standards or large scale evaluation (sample size) should aid in reducing this subjectivity.

Conclusion

The constructive MT evaluation method described contains many aspects of other methods that have been suggested over the past twenty years. In that sense it contains few new approaches. What does characterise the method is the use of measurement in a pragmatic and constructive role, whilst remaining fully aware of the practical difficulties associated with evaluating machine translated text.

At this stage, although this report only outlines a simple framework for an evaluation method, it is hoped that others will see some advantage in taking the ideas much further, hopefully resulting in a field trial of the method. The main work to be done to meet this goal is the definition of standard measurement procedures and units, that can be practically applied.

If the method is pursued, it is hoped that the key drivers for affecting intelligibility and fidelity can be identified. It might also be possible that different weights could be developed for particular language pairs, with the possibility that the resultant weights could be used as standards for future evaluation.

Acknowledgements

The author would like to thank Koji Miyahara and Yasuhiro Kawata for many useful discussions, which helped to shape the final content of this report. Additional credit is due to Yasuhiro Takayama and in particular to Yasuyuki Mochizuki, for providing the opportunity to present the paper. Finally the author would like to thank Fuyuki Maruyama and Tadashi Iwase for granting permission to attend the evaluation forum.

References

Balfour, Richard. W. 1986. “Machine Translation: A Technology Assessment: An Introduction, an Evaluation, and a Study of Market Position present and Future”, *BMT Consultants, London.*

Cook, M.L. 1982. “Software Metrics :: An Introduction and Annotated Bibliography.” *Software Engineering Notes,* 7:41-60, 1982.

Fenton, N.E., Whitty, R.W. 1986. “Axiomatic Approach to Software Metrication through Program Decomposition.” *Computer Journal,* 29(4):330-340, 1986.

Finkelstein, L., Leaning, M.S. 1984. “A Review of the Fundamental Concept of Measurement.” *Measurement,* 2(l):25-34, 1984.

Hamer, P.G. 1985. “Software Metrics – the Curate’s Egg.” In *Proc. IEE Colloquium on Software Reliability* : *Models and Measurement,* Digest no.21, pp 4/1-4, Tokyo, Japan, 1985.

Kawata, Y. 1991, "”A Report on MT Evaluation Methods,” internal research report *Mitsubishi Electric Corporation,* (in Japanese).

King, M. (not yet published). “A Practical Guide to the Evaluation of Machine Translation Systems,” *Interim report to SUISSETRA,* ISSCO, Geneva.

Lehrberger, J., Bourbeau, L. 1988. “Machine Translation : Linguistic Characteristics of MT Systems and General Methodology of Evaluation,” *John Benjamins Publishing Company, Amsterdam/Philadelphia.*

Miyahara, K. 1989, “Evaluation of the MELTRAN J/E Translation of a Personal Computer Manual”, internal research report *Mitsubishi Electric Corporation,* (in Japanese).

McCabe, T. 1976. “A Complexity Measure”. *IEEE Transactions on Software Engineering,* 2(4):308-320, 1976.

Nagao, M., Tsujii, J., Nakamura, J. 1986. “Machine Translation from Japanese into English”, *Proceedings of the IEEE,* Vol 74, No. 7, July 1986, pp. 993-1012.

Nagao, M., Tsujii, J., Nakamura, J. 1988. “The Japanese Government Project for Machine Translation”, *Machine Translation Systems,* pp. 141-

186, Edited by Jonathan Slocum, Cambridge University Press.

Nishiyama, H. 1990. “Case Study : Vendor Collaboration and Practical Operation of a Japanese to English Machine Translation at Matsuda” (my translation of title), Nikkei AI journal, special issue, winter edition, pp. 128-134 (in Japanese).

Palmer, M., Finin, T. 1990, “Workshop on the Evaluation of Natural Language Processing Systems”, *Computational Linguistics,* Vol 16, No. 3, September 1990, pp.175-181.

Suzuki, K. 1988, “Evaluation Criteria for MELTRAN J/E Translation System”, internal research report *Mitsubishi Electric Corporation,* (in Japanese).