A Hands-On Study of the Reliability and Coherence of Evaluation Metrics

Marianne Dabbadie[1], Anthony Hartley[2], Margaret King[3], Keith J. Miller[4], Widad Mustafa El Hadi[5], Andrei Popescu-Belis[3], Florence Reeder[4], Michelle Vanni[6]

[1] EVALING, Paris (France)
[2] Centre for Translation Studies, University of Leeds (UK)
[3] ISSCO/TIM/ETI, University of Geneva (Switzerland)
[4] The MITRE Corporation (USA)
[5] Université Lille III - Charles de Gaulle (France)
[6] U.S. Department of Defense (USA)

Abstract

This section of the workbook provides the description of the MT evaluation exercise that is proposed to the workshop participants, including the specification of the metrics for MT evaluation that the participants are suggested to use at the workshop.

1. A Collective Hands-on Exercise

1.1. Motivation

The motivations behind the LREC 2002 MT Evaluation workshop are grounded in previous work in the field, described at length in the previous section. The workshop is the sixth in a series of hands-on workshops on MT Evaluation, organized in the framework of the ISLE Project.

The goal of these hands-on evaluation workshops is to carry on a collective effort towards the standardization of MT evaluation. The ISLE taxonomy has been designed for standardization, but it would have not reached the present state without feedback from the participants at the workshops. Conversely, the participants have broadened their view of MT Evaluation, through the concrete use of the ISLE taxonomy for the design of toy evaluations, but also through extensive discussions with the organizers and other participants.

Some of the workshops have focused more on the setup of an evaluation depending on the desired context of use, others on metrics, others on reporting results obtained in this framework. As pointed out in the previous section, the need for a clear view of the performances of various metrics has prompted the organization of the present workshop, «Machine Translation Evaluation: Human Evaluators Meet Automated Metrics». Through hands-on application of selected metrics from the present workbook, the participants will be able to familiarize themselves with the current problems of MT Evaluation, to get a first-hand experience with recent metrics and to contribute to research in this field by their own observations of the metrics' behaviors.

1.2. Description of the exercise

The participants to the workshop are suggested to register with the organizers well before the day the workshop will take place (May 27, 2002). Thus, both organizers and participants will be able to prepare in advance an evaluation exercise (requiring several hours of work), so that the workshop itself can be devoted to the exploitation of those results.

The evaluation study that all participants are kindly required to carry on can be summarized as follows:

1. Select two evaluation metrics among those described below, preferably one «human-based» and one «automated» (more than two is welcome!).

2. Optionally, add one of the metrics that you have used before in MT evaluation, or any personal suggestion for a metric.

3. Using the test data provided by the organizers, apply the selected metrics and compute the scores of each translation, on a 0%–100% scale.
   The test data is described in the next document of the workbook and can be downloaded from `http://www.issco.unige.ch/projects/isle/mteval-may02/`. It consists in two source texts in French, each with a reference translation and about a dozen translations to be evaluated, from various systems and humans.

4. Send the results by email to the organizers (e.g., `Andrei.Popescu-Belis@issco.unige.ch`), together with any comments you believe useful.

5. Prepare a brief account of the evaluation (about 10–15 minute talk) to be presented at the workshop, for instance by first answering the question «what are the strongest and the weakest points in the measures that you used?»

1.3. Exploitation of the Results

The results of these evaluations will be discussed and highlighted at the workshop from the perspective of present research goals. Regarding individual metrics, the scores obtained by different evaluators using the same metric will inform the community about the reliability of that metric (cf. preceding document, 5.2), by computing standard deviation and inter-annotator agreement.

The other important result of the pre-workshop evaluations will be data on cross-metric correlation, i.e. the agreement between pairs of metrics. This is important both for metrics based on human judges (it illustrates how well the specifications are defined or how coherent the judges are) and for automated metrics (for which agreement with a reliable human judgement is almost the only proof of coherence). These meta-evaluation

considerations will be analyzed at the workshop by the organizers, based on the results sent to them by the participants. These considerations will constitute the basis for discussion and conclusions of the workshop.

2. Specifications of the Metrics

2.1. Preamble

The metrics that are proposed in this application illustrate a broad spectrum of those that were synthesized for the ISLE MT evaluation framework. The two categories identified below parallel of course the title of the workshop, «Human Evaluators Meet Automated Metrics». In the history of MT evaluation, given the difficulty of the task, most of the quality judgments, and later 'metrics', we carried on by humans. However, as explained in the previous chapter, the utility of automatic measures has always been clear: they provide cheap, quick, repeatable and objective evaluation. 'Objective' means here that the same translation will always receive the same score, as opposed to human judges that may have fluctuating opinions. However, since human judges are the final reference in MT evaluation, the results of automated metrics must correlate well with (some aspect of) human-based metrics.

The metrics specified below must of course be integrated in a broader view of evaluation, since none of them is sufficient to determine the overall quality of a system. As stated in the ISLE taxonomy, it is the desired context of use of the evaluated system that determines a 'quality model', namely a set of useful features, to which several metrics are associated. It is only the combination of these scores that provides a good view of the quality of the system in the given context.

Documentation about the metrics below (apart from the references quoted) can be found in several papers available over the Internet. The ISLE evaluation workgroup has a webpage at `http://www.issco.unige.ch/projects/isle/ewg.html`, with links to previous workshop material for MT Evaluation, and to electronic versions of Van Slype's (1979) report and of the MT Evaluation workshop held at the MT Summit VIII conference. The ISLE taxonomy can be found at `http://www.issco.unige.ch/projects/isle/taxonomy2/`.

Below is a synopsis of the metrics that will be described in the remaining part of this document.

(A1)	IBM's BLEU and the NIST version
(A2)	EvalTrans
(A3)	Named entity translation
(A4a)	Syntactic correctness
(A4b)	X-Score / parsability
(A5a)	Dictionary update / number of untranslated words
(A5b)	Translation of domain terminology
(A6)	Evaluating syntactic correctness from the implementation of transfer rules
(H1)	Reading time
(H2)	Correction / post-editing time
(H3)	Cloze test
(H4a)	Intelligibility / fluency
(H4b)	Clarity
(H5)	Correctness / adequacy / fidelity
(H6)	Informativeness: comprehension task

2.2. Automated/automatable metrics

2.2.1. IBM's BLEU and the NIST version (A1)

We mention first the most recent proposal of an automated metric for MT Evaluation, namely the BLEU algorithm proposed by a team from IBM (Papineni et al., 2001; Papineni, 2002). The principle of this metric, which was fully implemented, is to compute a distance between the candidate translation and a corpus of human «reference» translations of the source text. The distance is computed averaging *n*-gram similitude between texts, for $n = 1, 2, 3$ (higher values do not seem relevant). That is, if the words of the candidate translation, the bi-grams (couples of consecutive words) and tri-grams are close to one or more of those in the reference translations, then the candidate scores high on the BLEU metric.

Apart from intuitive arguments, the method to find out whether this metric really reflects translation quality is to compare its results with human judgements, on the same texts. In-house data (Papineni et al., 2001), as well as the DARPA 1994 data (Papineni et al., 2002), were used to test the coherence between human scores and BLEU scores, and this was found acceptable.

The metric was also adapted for the recent NIST MT Evaluation campaign (Doddington, 2001). The main changes were: text preprocessing, a differentiated weight associated to N-grams based on their frequency, and the use of tri-grams only. These modifications must still be discussed by the community, but the NIST provides yet the scripts implementing the BLEU metric as well as its adaptation, at: `http://www.nist.gov/speech/tests/mt/mt2001/resource/`.

We do not describe further this metric, but would like to refer the participants to the documentation quoted above, which provides enough resources to apply it.

2.2.2. EvalTrans (A2)

Automatic corpus evaluation extrapolation using EvalTrans (Niessen et al., 2000) gives statistics, such as the average Levenshtein distance standardized to the length of the target sentence. The tool can be downloaded at `http://www-i6.informatik.rwth-aachen.de/HTML/Forschung/Uebersetzung/Evaluation/`.

The first step is to load and save the human translations. For the present workshop, the reference translation as well as the other human translations of the same source text will constitute the «reference set». When the system is set up to work automatically, it will search this reference database for sentences which are most similar to the machine translated sentence that must be scored.

However, in order for the extrapolation to be performed, the Levenshtein distance algorithm needs to be seeded with scores for some (at least one) manually evaluated sentence. For this, a baseline machine translation (for instance) needs to be loaded and some sentence pairs need to be evaluated.

Next, the «test corpus» sentences need to be loaded. These are the machine translations for each source text. For each set of «test corpus» sentences, which comprise each machine translation of a source text, subjective sentence error rate (SSER) and multi-reference word error rate (mWER) will be calculated by the automatic metric.

- Several statistics of interest will be produced:
- Average number of «perfect» (scored 10) reference sentences per evaluation sentence pair (to indicate how reliable the mWER is).
- (average-score) / (value of all (evaluated/ extrapolated) sentence pairs)
- Standard deviation of the score
- Subjective sentence error rate (i.e., 100% * (1 – average-score)). An average score of 0.0 results in a SSER of 100%, an average score of 10.0 in a SSER of 0%.
- Subjective sentence error rate weighted by the length of the target sentences
- Average extrapolation distance: average Levenshtein distance (per target word) of all extrapolated sentences

The SSER indexes each sentence, then uses the mWER, the number of perfect reference sentences, the absolute Levenshtein distance to each sentence, and the Levenshtein distance to that sentence v. the length of current sentence.

The mWER is the word error rate against the most similar reference sentence which has been evaluated as «perfect» (i.e., has been assigned a score of ten). It is calculated as Levenshtein operations per reference word (and can thus exceed 100%). Average mWER for an evaluation corpus is calculated word-wise, not sentence-wise.

Another measure, the information item error rate, is not included because it relies heavily on manual scores, use of which would defeat the purpose of the automated metric.

2.2.3. Named entity translation (A3)

The NEE metric (Named Entity Evaluation) is described for instance in (Reeder et al., 2001). Since automated software to support this metric is available, it has been considered here an automated metric. Participants to the workshop may of course apply it manually, given the small amount of test data.

The process for utilizing this metric is relatively straightforward: a) identify the named entities within a given test corpus; b) pull unique entities from the document; c) find the entities in the system output text; and d) compare entities in the output text with those identified in the reference text (see Figure 1 below). Identifying the named entities in the reference translation requires human annotation, and is the only stage of the process to do so.

In a concrete example of this metric, to prepare the corpora for evaluation, two expert annotators used the Alembic Workbench (Day et al., 1997; see also `http://www.mitre.org/technology/alembic-workbench/`) annotation tool to tag occurrences of named entities according to the MUC annotation guidelines. After the named entities are tagged in the reference translation (designated here by ANNO), the metric can be applied.

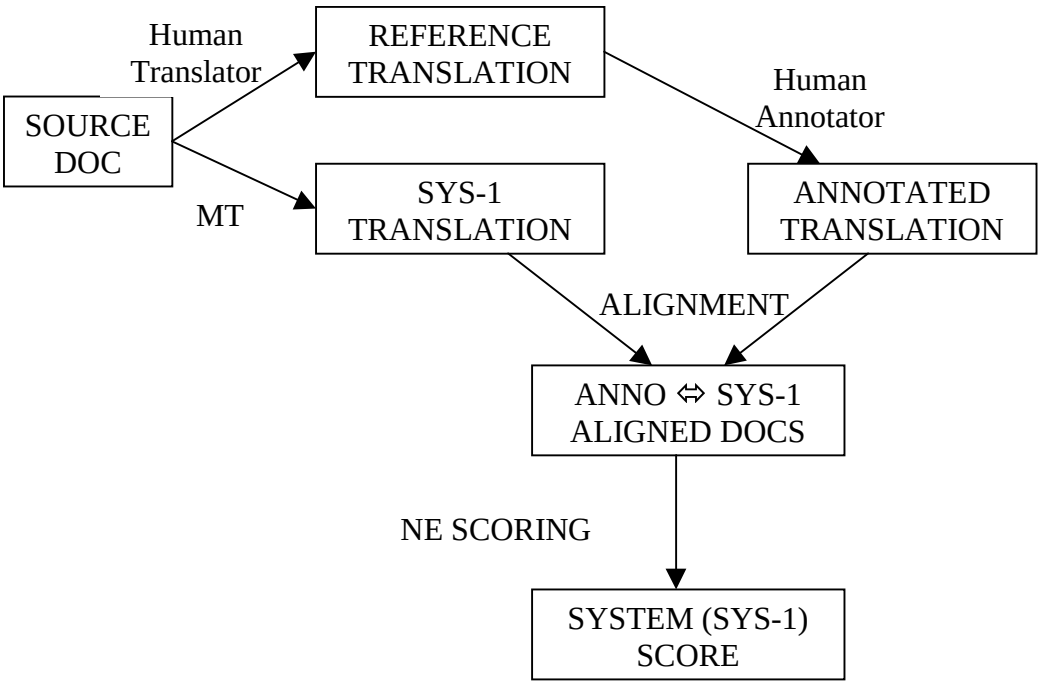

Figure 1. Scoring technique for the NEE metric.

The next stage is to align the ANNO translation text with the evaluation text (the output of the system SYS-1 for this example). To score the translation, for each article in the aligned pair, the tagged named entities are pulled from the ANNO and a list of unique names for the comparison unit (paragraph or article) is prepared. This is followed by normalization. At this time, the normalization steps applied are: (a) substitution of non-diacritic marked letters for the equivalent diacritic mark character for Romance languages (for instance **ã** becomes **a**); (b) down-casing; (c) the normalization of numeric quantities (particularly for numbers under 100) and (d) the removal of possessives. Other normalization steps may be needed, as well as the incorporation of partial match scoring (see Reeder et al., 2001). Once the named entity list and the SYS-1 tokens have been normalized, the search for named entities in the token lists is straightforward. Only exact matches given the normalization steps described are considered at this time and all results here reflect this.

2.2.4. Syntactic correctness (A4a)

The following describes a syntax metric based on the minimal number of corrections necessary to render an MT output sentence grammatical. Each evaluator must transform each sentence in the MT output into a grammatical sentence by making the minimum number of replacements, corrections, rearrangements, deletions, or additions possible. The syntax score for each sentence is then defined as the ratio of the number of changes for each sentence to the number of tokens in the sentence. For the purposes of this test, a token is defined as a whitespace-delimited string of letters or numbers. Additionally, individual punctuation marks, since they are subject to correction, are also counted as separate tokens. Each item of punctuation that occurs in pairs (e.g. brackets, braces, quotation marks, parenthesis) is counted as a separate token. Thus, in the following sentence, there are 24 tokens:

- *Mary, who had gone to see the fountain (in the center of town), said that it was turned off.*

It is important to remember that the final edited sentence need only be syntactically correct. That is, the final result may be semantically anomalous. Raters should endeavor to produce a syntactically correct sentence by making as few changes possible to the original MT output. Deletions, substitutions, additions, and rearrangements are counted by totaling the number of words deleted, substituted, added, or moved. In the event that there are combined operations, for example, moving a phrase consisting of four words, of which one has been deleted, the move is computed *after* the deletion is counted, thus the above-mentioned operation would result in one deletion and 3 moves. Finally, errors in inflectional morphology are not counted in the syntax metric. In applying this metric to test data, it was found that even when evaluators arrive at the same score for a given sentence (that is, they have the same total number of changes), they often choose a different combination of the four operations to arrive at their final grammatical sentence. The metric as it stands has not been automated, and would indeed be very difficult to automate; however, partial automation, such as automatic tracking and counting of necessary edit operations, would greatly assist in applying this metric in an efficient manner.

2.2.5. Automatic Ranking of MT Systems by X-Score (A4b)

Background: The X-Score metric aims to rank MT systems in the same order as would be given by a human evaluation of the Fluency of their outputs (Hartley & Rajman, 2001; Rajman & Hartley, 2002). The metric is especially adapted to rank machine translations relative to one another, rather than comparing human and machine translations. This metric was derived from experiments conducted on the French-English segment of the corpus used in the 1994 DARPA MT evaluation exercise. In that exercise, human evaluators scored translations of 100 source texts by 5 MT systems for their Fluency (among other attributes). To establish the present metric, the F-scores (Fluency scores) for individual texts were converted into rankings of systems using the aggregation technique of ranking by average ranks (average rank ranking or ARR). Using the same ARR technique, rankings were computed on the basis of the X-score for each document. The X-scores were found to represent a very good predictor of the ranking derived from the human evaluations (H-rankings). The distance between the H-ranking and the X-ranking is 1, corresponding to a similarity of 93.3%, a precision of 93,3% and a recall of 93.3%. If restricted to the most complete partial ranking, these values improve to a distance of 0.5, a similarity of 96.7%, a precision of 100% and a recall of 93.3%.

Computing the X-Score: The X-score is taken to measure the grammaticality of the translations. For any given document, the X-score is obtained as follows. First, the document is analyzed by the Xerox shallow parser XELDA in order to produce the syntactic dependencies for each sentence constituent. For example, for the sentence The Ministry of Foreign Affairs echoed this view, the following syntactic dependencies are produced: SUBJ (Ministry, echoed); DOBJ (echoed, view); NN (Foreign, Affairs); NNPREP (Ministry, of, Affairs).

On the corpus used in (Hartley & Rajman, 2001), XELDA produced 22 different syntactic dependencies, among which:

- RELSUBJ: for example, RELSUBJ(hearing, lasted) in «a hearing that lasted more than two hours»;
- RELSUBJPASS: for example, RELSUBJPASS(program, agreed) in «a public program that has already been agreed on ...»;
- PADJ: for example, PADJ(effects, possible) in «to examine the effects as possible»;
- ADVADJ: for example, ADVADJ(brightly, colored) in «brightly colored doors».

After each document has been parsed, we compute its dependency profile (i.e. the number of occurrences of each of the 22 dependencies in the document). This profile is then used to derive the X-score using the following formula:

- *X-score = (#RELSUBJ + #RELSUBJPASS – #PADJ – #ADVADJ)*

Note that several formulae would have been possible for computing the X-scores. The above-mentioned one

was selected in such a way that, if applied to the average dependency profile, it correctly predicted the average rank ranking (ARR) derived from the F-scores. In this sense, one can say that the computation of the X-score was specifically tuned to the test data and so it was considered quite ad hoc in (Hartley & Rajman, 2001). However, this is not true of (Rajman & Hartley, 2002). This second experiment retained exactly the same formula for the X-scores, while completely changing the human evaluations – evaluators directly assigned rankings to series of translations instead of assigning individual scores to each of the translations. Moreover, a new MT system was added, not present at all in the data that was used for the tuning. Thus, there is no reason to believe the X-scores to be ad hoc, which strongly increases their chances of being highly portable to other experimental data.

Computing the Rankings: For each of the documents, the scores of the systems are first transformed into ranks and the average ranks obtained by the systems over all the documents are then used to produce the final ranking.

2.2.6. Dictionary update (A5a) and domain terminology (A5b)

Dictionary update (also known as *non-translated or untranslated words*) and *domain terminology* are two potentially automatable metrics. Although related, these two metrics are not identical, as can be seen from their descriptions below. There are many ways in which a dictionary update measure could be calculated, but it seems obvious to use two objective and easy to observe features of MT output:

- the number of words not translated;
- the number of domain-specific words that are correctly translated.

It is these two features that have been described in previous related work, including (Vanni & Miller, 2002), and that will be specified below.

2.2.7. Number of untranslated words (A5a)

This metric makes use only of the target text. It is based on the intuition that translation quality is linked to size of vocabulary. In its simplest form, the number of words left untranslated is counted. By untranslated, we mean simply that a word which should be translated is not, and is simply copied over untouched into the target text. (This reflects the behavior of many machine translation systems). There are, of course, words which should not be translated (most proper names are a good example): not translating these items is not counted as an error. A score is obtained by the following calculation:

- (number-of-untranslated-words) / (total-number-of-words-in-text) x 100 = percentage-of- untranslated-words… *high is bad*

One possible way to automate this metric would be to run a spelling checker over the target text and count the number of mistakes found. This would, of course, pick up any spelling mistakes in translated words which might exist, as well as finding words which were not legal words of the target language; however, this amount is probably low for translations programs, which generate words based on valid dictionaries. On the whole, this automatic measure might not invalidate the metric as an indicator of overall translation quality.

In discussing the automation of this measure, it is worth noting that some MT systems provide as ancillary output statistics concerning the numbers of untranslated words in the output. However, this is not the case for all systems. In these cases, other automated means must be developed for computing this measure. In cases of languages using a non-Roman script or containing characters outside the standard lower-ASCII range found in typical English text, one possible way of counting non-translated words (for systems that simply pass untranslated words through in the translation) would be to locate and count tokens containing these characters that do not appear in English text. However, even in the case of the Japanese-English systems, some systems did produce a romanization of the untranslated words, and did not leave them in the native script. The romanizations contained only characters found in the lower portion of ASCII.

Given that this metric is intended to compute the number of words that the MT system was unable to translate, another possibility would be to use a tool such as *ispell* in order to identify non-English strings within the output translation. Counting these strings and comparing with the output of a utility such as *wc* (Unix word count) could provide a ratio of untranslated words in the output text.

Two potential problems with this last approach could both lead to undercounting the number of untranslated words in a text. First, included in the untranslated word count for Japanese – English translation were Japanese particles and other bits of non-English material, which may or may not have been the result of romanization of text found in the source. Examples of this include *na*, *re*, *X*, and *inu*. Another Japanese particle, *no*, did not appear in this context in the translation, but had we relied on an automated spelling-based identification of untranslated words, words like *no*, which also happen to be valid English strings (although with a different meaning) would be left uncounted. Secondly, untranslated word scores would likewise be affected for languages that share a high number of cognates with English. For these languages, the string in the source and target language may be identical, and thus not counted as an untranslated word, regardless of whether the system actually translated the word or simply passed it through.

The application of this metric to translations produced by human translators is somewhat doubtful: human translators when faced by a gap in their lexical knowledge try to work round the problem, and do not, normally, simply transcribe the problematic word or leave a gap. It is possible though that the spelling mistake variation might be informative.

It is also worth noting that while untranslated words certainly have an impact on the usability of MT output, such output often contains sentences that are completely unintelligible, but in no way due to untranslated words. Thus, this test should clearly not be used in isolation to provide a picture of overall MT quality, whether quality is defined along the lines of clarity, fluency, adequacy, or coherence.

2.2.8. Translation of Domain Terminology (A5b)

The domain terminology score is calculated as the percentage of correctly translated pre-identified domain terms. The procedure for this test is as follows: First, a list of key term translations is extracted from the human translation. To accomplish this, raters individually select key terms from the human translation, and then the separate key term lists are reconciled before application of the test to the MT systems' output. This step is amenable to automation, but has not as yet been automated. During the test application, systems receive a point for each term for which the translation matches the human translation exactly, and no point otherwise. The final score is the percentage of exactly-matched translations of key terms.

There are two divergent directions in which this test could be developed in the future. First, it could be made more sensitive to acceptable variation in translation of key terms by application of the ACME Cloze test methodology as described for instance in Miller (2000). This methodology simulates basing lexical tests on multiple human translation, while sufficiently constraining the structure of the translation to enable automated comparison.

2.2.9. Evaluating syntactic correctness from the implementation of transfer rules (A6)

This metric proposal is the result of two previous studies. In the first former study, the authors chose to count the number of NPs (noun phrases) and VPs (verb phrases) in source text and target texts, a first indication being given by non parallel data (Mustafa El Hadi, Timimi, Dabbadie, 2001). Another study presented the results on the same corpus after terminological enrichment (Mustafa El Hadi, Timimi, Dabbadie, 2002).

Nevertheless, the use of finer grained criteria such as adjectives or prepositional phrases count could also be envisaged. Any overlap of this threshold might then be considered as an indication that MT system may have failed to analyze source syntactic structure and that therefore, the initial figures require further analysis. But this methodology is still imprecise and limited to a first indication of MT system's analysis failure, when a gap is observed on non parallel data. The use of this methodology also implies that the test is carried out on relatively syntactically isomorphic languages such as French and English. A methodology including a test tool that would implement source and target transfer rules might probably prove more accurate and also apply to non isomorphic languages.

We propose here the following steps for the application of the metrics:

1. Deduce a set of French / English transfer rules from the source text and the reference translation (this part involves manual processing).
2. Write a script (e.g., in Java or Perl) to implement these rules (if not, go to point n. 3)
3. Check that these rules apply through the various candidate translations from the test data (automatically with the script or manually).
4. Generate an output failure file (or else carry out a manual check) and work out syntactic correctness.

2.3. Human-based measures

2.3.1. Reading time (H1)

Reading time can be defined in one of two ways: oral reading time or closed reading time.

Oral reading time (Van Slype, 1979) tends to measure more closely with intelligibility and also tends to be more relevant to higher quality translations. Therefore, for each document, the evaluators should read out loud the first paragraph and time the length of time that it takes to read each sample. The number of words then can be used to calculate a words per minute (WPM) rate:

- *WPM = number-of-words / reading-time*

The closer the WPM rate is to the WPM of natural language (depending on the evaluator), the higher is the quality of the translation (on a scale to be defined by each participant).

Closed reading time relates to the amount of time that a user needs to read a document to a «sufficient» level of understanding. The sufficient level is often paired with other measurements such as comprehension score on a test. Still, the instructions can be given that the readers measure the amount of time necessary to arrive at an understanding they consider to be sufficient to answer basic questions about the text. Words-per-minute rate can be calculated in the same way.

2.3.2. Correction / post-editing time (H2)

This metric is based on the intuition that the time required to produce an acceptable translation from a raw translation (whether produced by a human or by a machine) is inversely proportional to the overall quality of the raw translation.

It can be measured fairly easily by noting when the person responsible for the revision/post-editing starts their task and when they finish it, normalizing the result by taking into account the size of the text measured in words, then multiplying by a fixed factor in order to obtain a number on a wider scale. For this exercise, the following calculation is suggested:

- (number-of-minutes-spent-in-correction) / (total-number-of-words-in-text) x 10 = correction-time… *high is bad*

Note that this metric can only sensibly be applied to a whole text: timing correction to smaller text elements is both annoying for the person doing the timing and difficult to do reliably.

A variation on this metric is to count not the overall time but the number of key strokes made by the corrector.

It should be noted that this metric is somewhat problematic both with respect to validity and reliability for a number of reasons:

- The amount of correction needed depends in part on the ultimate use to which the translation will be put: a text destined for publication will probably be treated with more care than a text intended for information assimilation, for example
- The errors corrected differ in their nature. There will be straightforward grammatical or lexical errors, as well as more complicated stylistic errors. This will affect the amount of time needed to carry out the correction. This would not matter

so much if those doing the correction always agreed on what corrections are needed. But, inevitably, where matters of style are concerned, no such agreement exists.
- There is considerable variety amongst correctors and the way they work. Some work quickly and decisively, others are more hesitant and sometimes change their minds.
- Correctors may be influenced by knowing whether they are dealing with a human produced translation or a machine produced translation. One anecdote tells of correctors correcting far more on machine produced translation but spending comparatively less time in doing so because they felt no need to take into account the computer's feelings.

Participants who choose to work with this metric are invited to reflect on these issues and on possible improvements to the simple metric defined here.

2.3.3. Cloze test (H3)

This metric is reported by Van Slype (1979) as a test of readability. It may however also be thought of as a test of fidelity or of intelligibility, since it is based on the ability of a reader to supply a missing word correctly, which intuitively relates both to readability and intelligibility when the target text alone is considered and to fidelity when the source text is taken into account.

The method is simple. Every *n*-th word in the translation is deleted (in the Van Slype Report (1979), $n = 8$, but other values appear also in the literature). The translation is then given to a group of readers, who are asked to supply the missing words. Two scores are normally computed, one based on the number of answers which comprise exactly the suppressed original word, the other based on the number of answers with a word close in meaning to the original word. The second score has to be interpreted partly in the light of the first score

- (number-of-exact-answers) / (number-of-deleted-items) x 100 = percentage-of-exact-items-supplied… *high is good*

- (number-of-close-answers) / (number-of-deleted-items – number-of-exact-items-supplied) x 100 = percentage-of-close-items-supplied… *high is good*

A possible weakness of this metric is that it potentially also tests the intelligence and wealth of vocabulary of the reader supplying the missing words. This weakness can be mitigated by controlling the size and type of the group of readers.

A second possible weakness appears if the translated text is technical in nature: the readers have to have sufficient knowledge of the subject matter to make it plausible that they should be able to supply the missing items.

Van Slype (1979) also points out that some texts are more redundant than others in the way they carry information, and that if translations of several texts are to be compared, it is important to take this factor into account. He suggests that this can be done by carrying out a Cloze test also on the original text.

2.3.4. Intelligibility / fluency (H4a)

Intelligibility is one of the most frequently used metrics of the quality of output. Numerous definitions (or protocols for measuring it) have been proposed for it, for instance in Van Slype's report or in the DARPA 1994 evaluations. We outline here the definition proposed by T.C. Halliday in (Van Slype, 1979, p. 70), which measures intelligibility on a 4-point scale (0 to 3).

Intelligibility or comprehensibility expresses how intelligible is the output of a translation device under different conditions (for instance, when the sentence fragments are translated while being entered, or after each sentence). Comprehensibility reflects the degree to which a complete translation can be understood. Intelligibility can be based on the general clarity of translation, or the output can be considered in its entirety or by segments out of context.

The following scale of intelligibility has been proposed, from 3 to 0, 3 being the most intelligible:

- 3 – Very intelligible: all the content of the message is comprehensible, even if there are errors of style and/or of spelling, and if certain words are missing, or are badly translated, but close to the target language.
- 2 – Fairly intelligible: the major part of the message passes.
- 1 – Barely intelligible: a part only of the content is understandable, representing less than 50% of the message.
- 0 – Unintelligible: nothing or almost nothing of the message is comprehensible

To apply the metric, the following steps are suggested:

1. Take the reference translation of a text (or the source if you are proficient in that language).
2. Separate and number the sentences.
3. Take a candidate translation and do the operation (2) on it. Match sentences with those in the reference/source translation.
4. Rate sentences from the candidate translation using the 0 to 3 scale described above.
5. Optional: to normalize scores, calculate intelligibility on a 0% to 100% scale, by averaging sentence ratings over the whole text.
6. Produce a final score for each translation

2.3.5. Clarity (H4b)

In work described in (Vanni & Miller, 2002) a metric called *clarity* is proposed that merges the ISLE categories of comprehensibility, readability, style, and clarity into a single evaluation feature. This measure ranges between 0 and 3. Raters are tasked with assigning a *clarity* score to each sentence according to the following criteria:

Score	Criterion
3	meaning of sentence is perfectly clear on first reading
2	meaning of sentence is clear only after some reflection
1	some, although not all, meaning is able to be gleaned from the sentence with some

effort

0 Meaning of sentence is not apparent, even after some reflection

Since the feature of interest is clarity and not fidelity, it is sufficient that some clear meaning is expressed by the sentence and not that that meaning reflect the meaning of the input text. Thus, no reference to the source text or reference translation is permitted. Likewise, for this measure, the sentence need neither make sense in the context of the rest of the text nor be grammatically well-formed, since these features of the text would be measured by tests proposed elsewhere, namely the *coherence* and *syntax* tests, respectively. Thus, the clarity score for a sentence is basically a snap judgement of the degree to which some discernible meaning is conveyed by that sentence.

2.3.6. Correctness / adequacy / fidelity (H5)

This evaluation metric reprises the DARPA 1994 *adequacy* test (Doyon, Taylor, and White, 1996). As with that test, the reference translation or "authority version" is placed next to each of the translations of the source text, to be used as a comparison against each one, human or machine. Before the test is performed, both the "authority version" as well as each of translations should be segmented, with each text separated into sentence fragments to appear next to the corresponding fragment in the translation.

Once each translation is lined up with its equivalent, evaluators grade each unit on a scale of one to five, where five represents a paragraph containing all of the meaning expressed in the corresponding text. The *Adequacy* scale is as follows:

- 5 – All meaning expressed in the source fragment appears in the translation fragment
- 4 – Most of the source fragment meaning is expressed in the translation fragment
- 3 – Much of the source fragment meaning is expressed in the translation fragment
- 2 – Little of the source fragment meaning is expressed in the translation fragment
- 1 – None of the meaning expressed in the source fragment is expressed in the translation fragment

2.3.7. Informativeness: comprehension task (H6)

There are two methods for testing comprehension. The most common of these is the reading comprehension exam (e.g., Somers & Prieto-Alvarez, 2000; DARPA-94; Tomita 1992). In this case, the evaluators design a set of questions, usually under 10, for the given texts. Sometimes, as in the case of Tomita, these tests are structured first and then applied to the translations. Tomita began with the Test of English as a Foreign Language (TOEFL) examinations which he then translated to Japanese and had students take. The theory being that the better scores on the exam will have resulted from the better translations. The big difficulty (Somers & Prieto-Alvarez, 2000) is that it is difficult to test only the reading without bringing a large amount of pre-existing world knowledge to the table. In addition, the design and structuring of such examinations is an art in and of itself.

The second method for a comprehension test takes instead the task of figuring out the kinds of questions that one might want to be able to answer from a translation and determining whether the translation can support answering said questions. For instance, one might want to know the people, places and organizations mentioned in an article. This is covered by the named entity metric. Yet, it is really only the first stage of measurement. The secondary measure would be to look to determine if the entity relationships are also preserved by the translation - that is, who belongs to what organization or who did what to whom. This is the question we began to study at MT Evaluation workshop organized at NAACL 2001, when we asked participants to fill in templates based on specific kinds of questions. The better systems would enable the successful template filling and scoring would follow Message Understanding (MUC) guidelines. It is this type of exercise you will be asked to do at this time. The previously identified named entities will be used here. You will fill out templates to answer specific details of events or relationships between parties.

3. References

D. Day, J. Aberdeen, L. Hirschman, R. Kozierok, P. Robinson, and M. Vilain. 1997. Mixed-Initiative Development of Language Processing Systems. In *Fifth Conference on Applied Natural Language Processing*, Washington, D.C.

G. Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *HLT 2002, Human Language Technology Conference*, San Diego, CA.

J. Doyon, K. Taylor, and J.S. White. 1998. The DARPA MT Evaluation Methodology: Past and Present. In *Proceedings of the AMTA Conference*, Philadelphia, PA.

A. Hartley and M. Rajman. 2001. Automatically Predicting MT Systems Rankings Compatible with Fluency, Adequacy or Informativeness Acores. In *Workshop on MT Evaluation "Who did what to whom?" at MT Summit VIII*, Santiago de Compostela, Spain, p.29-34. See `http://www.issco.unige.ch/projects/isle/MT-Summit-wsp.html`.

K. J. Miller. 2000. *The Machine Translation of Prepositional Phrases*. Unpublished PhD Dissertation. Georgetown University. Washington, DC.

W. Mustafa El Hadi, I. Timimi and M. Dabbadie. 2001. Setting a Methodology for Machine Translation Evaluation. In *Workshop on MT Evaluation "Who did what to whom?" at MT Summit VIII*, Santiago de Compostela, Spain, p.49-54. See `http://www.issco.unige.ch/projects/isle/MT-Summit-wsp.html`.

W. Mustafa El Hadi, I. Timimi, and M. Dabbadie. 2002. Terminological Enrichment for non-Interactive MT Evaluation. In *LREC 2002, Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain.

S. Niessen, F.J. Och, G. Leusch, H. Ney. 2000 An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *LREC 2000, 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, pp. 39-45.

K. Papineni. 2002. Machine Translation Evaluation: N-grams to the Rescue. In *LREC 2002, Third*

International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain.

K. Papineni, S. Roukos, T. Ward, J. Henderson, and F. Reeder. 2002. Corpus-based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results. In *HLT 2002, Human Language Technology Conference*, San Diego, CA.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2001. BLEU: a method for automatic evaluation of MT. Research Report, Computer Science RC22176 (W0109- 022), IBM Research Division, T.J.Watson Research Center, 17 September 2001. See `http://domino.watson.ibm.com/library/CyberDig.nsf/home`, and search for 'RC22176'.

M. Rajman and A. Hartley. 2002. Automatic Ranking of MT Systems In *LREC 2002, Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain.

F. Reeder, K.J. Miller, J. Doyon, and J.S. White. 2001. The Naming of Things and the Confusion of Tongues: an MT Metric. In *Workshop on MT Evaluation "Who did what to whom?" at MT Summit VIII*, Santiago de Compostela, Spain, p.55-59. See `http://www.issco.unige.ch/projects/isle/MT-Summit-wsp.html`.

H. Somers and N. Prieto-Alvarez. 2000. Multiple Choice Reading Comprehension Tests for Comparative Evaluation of MT Systems. In *Workshop on MT Evaluation at AMTA-2000*.

M. Tomita. 1992. Application of the TOEFL Test to the Evaluation of Japanese-English MT. In *MT Evaluation Workshop at AAMT*.

G. Van Slype. 1979. Critical Study of Methods for Evaluating the Quality of Machine Translation. Technical Report BR19142, Bureau Marcel van Dijk / European Commission (DG XIII), Brussels. See `http://issco-www.unige.ch/projects/isle/van-slype.pdf`.

M. Vanni and K. J. Miller. 2002. Scaling the ISLE Framework: Use of Existing Corpus Resources for Validation of MT Evaluation Metrics across Languages. In *LREC 2002, Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain.