An experiment in comparative evaluation: humans *vs.* computers

Andrei Popescu-Belis

ISSCO/TIM/ETI, University of Geneva
40, bd. du Pont d'Arve – CH-1211 Geneva 4 – Switzerland
andrei.popescu-belis@issco.unige.ch

Abstract

This paper reports results from an experiment that was aimed at comparing evaluation metrics for machine translation. Implemented as a workshop at a major conference in 2002, the experiment defined an evaluation task, description of the metrics, as well as test data consisting of human and machine translations of two texts. Several metrics, either applicable by human judges or automated, were used, and the overall results were analyzed. It appeared that most human metrics and automated metrics provided in general consistent rankings of the various candidate translations; the ranking of the human translations matched the one provided by translation professionals; and human translations were distinguished from machine translations.

1 Introduction

The quality of machine translation (MT) systems can be measured using a variety of techniques, which often depend on the context in which the MT system will be used. Whereas many parameters are relevant to the global quality of a system, it is often on *output quality* that developers focus. The quality of a set of texts translated by a system is measured using judgments similar to those employed for human translations. Since there is no unique perfect translation of a given text, but many acceptable variants, the challenge is to measure the quality of an MT-translated text in the most *objective* and *economic* manner. Many techniques, either based on human judges or on automatic procedures, have been proposed. The goal of this paper is to report on an experiment that applied several of these techniques to a set of systems, and compared the results. The test data consisted of both human (imperfect) translations and machine translations.

The paper is organized as follows: we first describe the organization of the experiment (as a workshop at a major conference in 2002) and the definition of the task and evaluation metrics. We then describe the test data that was used, and give the evaluation results, which are discussed in the end. The main questions are whether human metrics and automated metrics provide consistent scores (or at least rankings), and whether they show different results on human *vs.* machine translations.

2 Evaluation context

The experiment reported on here was designed as an informal test-bed for various evaluation metrics. Therefore, it pertains to "meta-evaluation", rather than to evaluation. The main goal is to analyze the behaviour of evaluation metrics, not to evaluate in detail a given system, or to set up a competition among systems.

2.1 Organization of the experiment

The experiment was staged as a workshop at the LREC 2002 conference (Language Resources and Evaluation), as part of a series of workshops on MT evaluation, supported by the joint European and US project ISLE (International Standards for Language Engineering, 1999-2002) through its Evaluation Work Group (*http://www.issco.unige.ch/projects/isle/ewg.html*). Given that the 2002 workshop had to take place in just one day, most of the work had to be done before the workshop, which was mainly devoted to the report of individual results. Therefore, the evaluation experiment had to be carefully planned about three months before the workshop. The following tasks were accomplished.

2.1.1 Preparation of metrics to be tested

One of the organizers prepared a list of metrics for MT, distinguishing two categories: automated (or at least automatable) *vs.* intrinsically human, picked from the literature on MT evaluation (cf. list in Table 1 below). The organizers then gathered evaluation guidelines for the application of each metric (in one page or less, plus references). Finally, a workbook for the participants was edited.

2.1.2 Preparation of the data

A set of different translations of two articles in French was assembled. The articles were translated from French into English by students (as a graded university translation test), and also by several online MT systems. A “reference translation” for each article was also prepared from the student translations, taking into account the corrections made by their professor. The test data was prepared for distribution, by removing information on the authors of the translations (humans or systems) and replacing them with numbers.

2.1.3 Pre-workshop evaluation

Organizers and pre-registered participants proceeded then to evaluate the translations, by choosing two metrics among those described by the organizers (one manual and one automatic), applying them, and sending back the results to the organizers. A preliminary analysis of the results was prepared before the workshop.

2.1.4 Workshop

At the workshop, the evaluation results were presented and discussed. Some time was left at the beginning of the workshop so that all participants familiarize themselves with the evaluation task, in case they did not pre-register.

2.2 Definition of the task

Since the participants were suggested to register with the organizers well before the workshop took place, they were able to prepare in advance the evaluation experiment. The individual task that was described to them was the following:

1. Select two or more evaluation metrics among those described in the experiment’s guidelines, one “human-based” and one automated.
2. Optionally, add other metrics that the participant 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 of each of the two texts, if possible on a 0%–100% scale (it turned out that this was not always respected).
4. Send the results back to the organizers.
5. Prepare a brief account of the evaluation (about 10–15 minute talk) to be presented at the workshop.

The test data, described in more detail below, was made available online. The participants knew that the English translations of the two French source texts (the object of the evaluation) came from various MT systems and from students in translation, but did not know exactly the origin of each translation.

2.3 Metrics proposed

The metrics proposed in the guidelines of the experiment illustrated a broad spectrum of those that were synthesized for the ISLE MT evaluation framework (*http://www.issco.unige.ch/projects/isle/taxonomy3/*). In the history of MT evaluation, given the difficulty of the task, most of the quality judgments and metrics were carried on by *human judges*. However, the utility of *automated metrics* 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 reference in MT evaluation, the results of automated metrics must correlate well with (some aspect of) human-based metrics.

The participants were given a broad list of MT evaluation metrics, and were asked to apply an automatic and a human one. Below is a synopsis of the metrics, with the code names that will be used further on, and important references (Table 1).

(A1)	IBM's BLEU (Papineni 2002, Papineni, Roukos, Ward and Zhu 2001) and the NIST version (Doddington 2002)
(A2)	EvalTrans (Niessen, Och, Leusch and Ney 2000)
(A3)	Named entity translation (Reeder, Miller, Doyon and White 2001)
(A4)	X-Score / parsability (Hartley and Rajman 2001, Rajman and Hartley 2002)
(A5)	Dictionary update / number of untranslated words (Vanni and Miller 2002)
(A6)	Evaluating syntactic correctness from the implementation of transfer rules
(H1)	Reading time (Van Slype 1979)
(H2)	Correction / post-editing time (Van Slype 1979)
(H3)	Cloze test (Van Slype 1979)
(H4a)	Intelligibility / fluency (Van Slype 1979, p.70)

(H4b)	Clarity (Vanni and Miller 2002)
(H5)	Correctness / adequacy / fidelity (Doyon, Taylor and White 1998)
(H6)	Informativeness: comprehension task (Somers and Prieto-Alvarez 2000)

Table 1. MT evaluation measures proposed for the present experiment (A: automated; H: human).

3 Test data

3.1 Presentation

The human-translated texts were extracted from a corpus of translation study examinations that is under construction at the École de Traduction et d’Interprétation (University of Geneva). In the corpus, the translations are encoded using markup, together with the corrections made by professors, and most important, with the *grade* that has been attributed to each translation. In the present experiment, the data was stripped off its markup, that is, was restored to its initial state. The machine translated texts were obtained by submitting to several online translation tools the text that was translated by the students.

3.2 Source texts and reference translations

The two source texts (named in the experiment *10S* and *20S*) are excerpts from two longer essays, originally in French. Evaluators had access to them, and to a reference translation for each text (*10A* and *20A*) that we constructed from the best student translations, using also the teacher’s corrections. An important information, that was not initially given to the evaluators, is that reference 10A was based on translation 104 (also with insights from 105), while reference 20A was based on translation 203.

Even if it is clear that the “reference translation” is not meant to be “the perfect translation” (since there is no single perfect translation), some evaluators have given the reference a privileged role in evaluation. The reference was only meant to be a correct translation, close enough to the source text to help evaluators that do not understand French. The subjects of the texts are “Children and drugs” and “Taliban and women” and their references can be found on the workshop website (http://www.issco.unige.ch/projects/isle/mteval-may02/).

3.3 Human translations

The translations produced by translation students were numbered 101–106 and 201–204. More translations were not available, since the French-to-English examination was not passed at the ETI by many students, and we looked for English output only. The students of the 100 series had French as their first language, while those of the 200 series had English as their first language. The marks (cf. Table 2) were assigned by two faculty professors, from 1 to 6 (6 best, 4 needed to pass).

Code	Mark	Code	Mark
104	5.5	203	5.3
105	5.2	202	4.9
101	5.0	204	4.6
102	5.0	201	4.0
103	4.3	-	-
106	4.3	-	-

Table 2. Marks of the human translations, out of 6.

All the translations were made by different students, so for instance 101 and 201 are not made by the same student. The students were not instructed to use either of the particular varieties of English (British vs. American), hence some slight spelling variations. A sample of the translations produced for the first text (including source and reference) is provided in Table 3 below.

3.4 Machine translations

We used systems which were available online, on web sites that offer general-purpose translation tools. Our purpose wss not to evaluate intrinsically these tools, which are obviously a very helpful feature on the respective web sites (listed by Laurie Gerber’s at *http://www.lim.nl/eamt/resources/*). Therefore, we will neither disclose the precise origins of each machine translation we used, nor the websites. It appeared a posteriori that our seven translations came from only four different systems, sometimes parameterized differently[1] . The grouping is the following: x07, x08+x09, x10+x11+x13, x12 (so files 107 and 207, 108 and 208, etc., are from the same version). There are only

[1] In alphabetic order: Lernout & Hauspie’s T1, SDL Enterprise Translation Server, Reverso/ProMT, and Systran.

very minor differences in the output within each group (cf. Table 3, right column).

3.5 The evaluators' knowledge

The view presented until now discloses all the details of the test data. However, in order not to bias evaluation, the evaluators did not know everything about the test data. The evaluators *were given* the source texts, the reference translation (with the caveat above), and the candidate translations. They *knew* that output came from both humans and machines, but *did not know* precisely which numbers came from humans and which from machines. A fortiori, they did not know that 107/207, 108/208, etc. came from the same system, nor that x10, x11, x13 had very similar origins (though this was quite easy to discover). Several participants attempted to spot the humans, with some success of course: mistakes made by systems, such as untranslated words, or options such as he/she/it, are an easy hint. However, some evaluators proceeded directly with the evaluation, and the results sometimes led them naturally to separate humans from machines.

4 Results

4.1 Individual results: applications of metrics

The following scores were obtained by various participants to the workshop; however, not all of them are reported. We describe how each metric was applied, then provide the comparative results.

Regarding the automated metric, most of the participants applied BLEU (sometimes in the NIST version), since the existing software is freely available. However, the choice of the reference translations differed quite a lot, hence the scores differed too, and the scales as well. Regarding the human metrics, there was more variety, but the number of judges used was not always sufficient.

4.2 Applications of the BLEU/NIST metrics

The BLEU metric (Papineni 2002, Papineni, Roukos, Ward and Zhu 2001) uses of a set of reference translations to score a candidate translation, by estimating its overall "proximity" to the set. In our case, no such set was given, therefore the evaluators had to invent other solutions.

One of the evaluators, George Doddington, using BLEU modified by NIST (Doddington 2002), proposed to use 10A and 20A as single references (hence this metric is noted **NIST-1**). Unsurprisingly, his BLEU/NIST score of 10A and 20A was maximal (about 8.7), then followed the scores of the (human) translations on which they were based (104 at 6.75, and 203 at 7.80), then quite far behind the other translations (see Tables 4 and 6 for the ranking). Of course, given the somewhat arbitrary choice of the reference translation, it is not the case here that the human translations always score higher than the machine translations.

George Doddington also proposed a second application: for each of the two series, he used *each of the translations* as a single reference, and plotted together the score curves thus obtained on the same diagram. Considering the result, he asserted that "translations 107-113 (resp. 207-213) all share a similarity that makes them categorically different from the other translations" (Doddington, personal communication). Indeed, each score curve reaches a maximum for the translation that it uses as a reference, but besides that, when 10A, 101,…, 106 are used as references, the machine translations 107-113 all exhibit uniformly low scores, as well as the human ones (in some sense, "no other translation resembles a human translation"). But, when 107, …, 113 are used as references, the human translations still score low, while some of the machine translations score higher than before. Graphically, the difference is quite striking, due also of course to the particular ordering of the index numbers (humans first, then machines).

Cristina Vertan used BLEU to compute similarity between each of the candidate translations and the reference translation (10A, resp. 20A). The rankings shown in Tables 4 and 6 are similar but not identical to those obtained using the NIST version and the same protocol.

Bonnie Dorr, realizing the problem of using a single translation as a reference, proceeded to enrich the collection with three more reference translations done at UMIACS, University of Maryland. A first scoring protocol (**NIST-4**) used BLEU/NIST with four reference translations. Of course, 104 and 203 have again the highest scores (0.68 and 0.77, on an unknown scale), now followed by some other human translations. However, the ranking shown on Tables 4 and 6 still fails to separate humans from machines.

<table>
<tr><td colspan="2"><i>Source text</i>
Les résultats d'études récentes le démontrent clairement : plus la prévention commence tôt, plus elle est efficace.
Il n’est pas forcément nécessaire d'être un spécialiste des toxicomanies pour aborder ce sujet avec vos enfants.</td></tr>
<tr><td colspan="2"><i>Reference translation</i>
The findings of recent studies clearly show that the earlier prevention starts, the more efficient it will be.
You do not necessarily need to be an expert in drug dependence to talk about this issue with your children.</td></tr>
<tr><td><i>Translation 101</i>
The findings of recent studies clearly show that "the earlier the prevention, the most efficient it is."
You do not necessarily need to be a specialist in drug addictions to talk over this issue with your children.</td><td><i>Translation 107</i>
The results of recent studies show it clearly: more the prévention begin early, more she is effective.
It is not necessarily necessary be a specialist of the toxicomanies to approach this subject with your children.</td></tr>
<tr><td><i>Translation 102</i>
Outcomes of recent studies carried out recently, clearly demonstrate that the sooner the prevention begins, the better and the more successful it will be.
You needn't be a specialist in drugs to talk about it with your children.</td><td><i>Translation 108</i>
The results of recent studies demonstrate him(it) clearly: the more the prevention begins early, the more it is effective.
It is not necessarily necessary to be a specialist of the drug addiction to approach this subject with your children.</td></tr>
<tr><td><i>Translation 103</i>
Recent studies have clearly shown that the earlier the prevention begins, the more efficient it is.
It is not unavoidably necessary to be a specialist in drug addictions to talk about this subject with your children.</td><td><i>Translation 109</i>
The results of recent studies demonstrate him(it) clearly: the more the prevention begins early, the more it is effective.
It is not necessarily necessary to be a specialist of the drug addiction to approach this subject with your children.</td></tr>
<tr><td><i>Translation 104</i>
As recent studies have clearly shown, the earlier prevention starts, the more efficient it will be.
You do not necessarily need to be an expert in dependences to talk about this issue with your children.</td><td><i>Translation 110</i>
The results of the recent studies show it clearly: the more the prevention starts early, the more it is effective.
It is not inevitably necessary to be a specialist in drug-addiction to tackle this subject with your children.</td></tr>
<tr><td><i>Translation 105</i>
Recent studies have shown very clearly that the earlier prevention starts, the more effective it will prove.
You do not necessarily need to be an expert in addictions to talk about that issue with your children.</td><td><i>Translation 111</i>
The results of the recent studies show it clearly: the more the prevention starts early, the more it is effective.
It is not inevitably necessary to be a specialist in drug-addiction to tackle this subject with your children.</td></tr>
<tr><td><i>Translation 106</i>
Recent study results show this clearly: the earlier the prevention starts, the more efficient it is.
It is not completely necessary to be a specialist on drug addiction to discuss this subject with your children.</td><td><i>Translation 112</i>
The results of recent studies demonstrate it clearly : the earlier the prevention begins, the more efficient it is.
Him n ' is not inevitably necessary of to be a specialist of the drug addictions to approach this subject with your children.</td></tr>
<tr><td></td><td><i>Translation 113</i>
The results of the recent studies show it clearly: the more the prevention starts early, the more it is effective.
It is not inevitably necessary to be a specialist in drug-addiction to tackle this subject with your children.</td></tr>
</table>

Table 3. Excerpt from the test data: source text (French), reference translation, candidate translations from humans and from commercial systems available over the Internet.

Bonnie Dorr proposed two other protocols, applicable when separating human from machine translations. For human ones, the idea (noted **NIST-H**) is to score each candidate against all the other human translations (*n*–1); the scores vary here for instance from 0.85 for 10A, 0.73 for 104, to 0.36 for 102. For machine translations, the idea (**NIST-M**) is to score them against all the human ones (notwithstanding the imperfection of the student translations). The score range is narrower here, from 0.56 for 110, to 0.45 for 107. The ranking is shown on Tables 4 and 6. Note that using the two protocols in a single column is somewhat misleading, since it is unclear whether scores for **NIST-M** and **NIST-H** are comparable; we did this in order to reduce the number of tables.

4.2.1 *Applications of human-based metrics*

Unlike automated metrics, the metrics that need human judges are more varied, but take more time to apply. At the workshop, variants of fluency and fidelity, as well as reading time and correction time were measured, sometimes with just one judge implementing the procedure.

Eva Forsbom measured fidelity (H5) – how much of the meaning of the source is conveyed in the candidate translation, computed for each sentence – with four evaluators carrying on the task. The scores go from 25 (worst) to 45 (best), and the rankings are shown in Tables 5 and 7.

Cristina Vertan measured intelligibility (or fluency), that is, the clarity of the candidate, independently of its closeness in meaning to the source (H4a). Only one judge, not a native English speaker, applied this metric.

Finally, Michelle Vanni measured reading time (H1) and correction time (H2) for each of the candidates, using only initial excerpts of about 100 words and one judge. Both metrics are in principle related to fluency, since it has been shown that non-fluent texts take longer to read and longer to correct. The second metric is of course also related to fidelity.

NIST-1	NIST-4	NIST-H/M	BLEU
104	104	104	104
101	106	106	101
106	110, 111	110, 111	106
110,111,113	…	…	110,111,113
…	113	113	…
…	101	101	…
105	108	108	105
103	109	109	109
109	112	112	108
108	107	107	103
107	103	103	112
102	105	105	102
112	102	102	107

Table 4. Ranking of the 100 series translations following various applications of the BLEU/NIST metrics.

Fidelity (H5)	Fluency (H4a)	Reading time (H1)	Correction time (H2)
101, 104	102, 104, 105, 106	103	105
…	…	102, 112	104
103, 106	…	…	106
…	…	101, 110, 111, 113	111, 113
102	101, 103, 109	…	…
105	…	…	109
109	…	…	102
110	108	104,105,108	103
108	107	…	110
111, 113	110, 111	…	101
…	…	109	112
112	112, 113	106	107, 108
107	…	107	…

Table 5. Ranking of the 100 series translations following various human-based metrics.

4.3 Comparative results

We decided to use, for the moment, only the *ranking* provided by the previous measures, since the scores appear to be on heterogeneous scales. Tables 4 and 5, for the 100 series, and Tables 6 and 7, for the 200 series, display the ranking of the texts, from the best to the worst. When two candidates receive the same score, they are quoted together in the same cell, and the cell below is left empty. The results are meant to reflect the quality of the texts, in the first place – though caution is necessary since many metrics were applied in non-canonical conditions. The quality of the texts is of course related to that of the systems or people who produced them, but again, two texts are certainly not enough to score a system. Finally, the most relevant information is probably about the metrics themselves: the rankings often display agreement between metrics, and agreement with the “official” evaluation of the examinations (101-106, 201-204).

NIST-1	NIST-4	NIST-H/M	BLEU
203	203	203	203
204	211, 213	210,211,213	202
210,211,213	…	…	207
…	210	…	209
…	204	204	204
202	208	208	208

208	209	209	210
209	207	207	211, 213
207	212	212	…
201	202	202	212
212	201	201	201

Table 6. Ranking of the 200 series translations following various applications of the BLEU/NIST metrics.

Fidelity (H5)	**Fluency (H4a)**	**Reading time (H1)**	**Correction time (H2)**
203	203	201	202, 204
204	207	204	…
202	208	202	201
210	209	203	203
211	213	209	210
213	202	212, 213	213
208	204,210,211	…	211, 212
209	…	208	…
201	…	210	208
212	201	211	209
207	212	207	207

Table 7. Ranking of the 200 series translations following various human-based metrics.

5 Analysis

To analyze the results, we will first separate human from machine translations, since they are from the start of a different nature. We will compare them in the last subsection.

5.1 Ranking of student translations

The professor's marks ranked student translations in this order: 104 > 105 > 101 = 102 > 103 = 106, and respectively 203 > 202 > 204 > 201. Regarding the automated metrics, they fail to reproduce the professor's ranking of the 100 series: it is true that 104 has maximal score, but this is because it served to write the reference translation, often used as the only reference. Text 105, which has the second best mark, received low BLEU/NIST scores, as it differed notably from 104. Similar results hold for the 200 series. Unsurprisingly, n-gram proximity is not a good criterion to compare human translations, which often exhibit syntactic and lexical diversity. When using the human metrics, which do not require a reference translation, the human translations are consistently ranked among the first ones, though their ranking does not always match the professor's one (e.g., metric H1 on 200 exhibits the reverse order!). It is of course expected that humans score quite high on fluency, if their knowledge of the target language is reasonable.

5.2 Ranking of machine translations

The use of a single reference translation is somewhat less penalizing when scoring machine translations. Nearly all of the metrics consistently detected that there were in fact only four systems: 7, 8+9, 10+11+13, and 12, even with the slight differences between parameterizations of a single system. The results consistently show two groups: 8+9 and 10+11+13 score always higher than 7 and 12: this is probably one of the most certain conclusions. It is also quite interesting that human and automated metrics globally agree on this ranking. On the 100 series, the "average" ranking is 8+9 > 10+11+13 > 12 > 7 (or maybe 7 > 12). For the 200 series, the "average" ranking is 10+11+13 > 8+9 > 7 > 12. Obviously, there is not enough test data to produce a finer-grained ranking, but it is remarkable that even with scarce data, there is agreement on the two better systems 8+9 and 10+11+13 (this could be one of the reasons they are embedded in several web sites).

5.3 Comparison of humans and machines

The separation of humans *vs*. machines is not always visible on the scores (maybe an encouraging result for system designers) – but this was not the goal of our experiment anyway. The frequent use of only one (human) reference for the automated metrics explains why the other human translations score low. However, as we mentioned, a graphical analysis prompted George Doddington to distinguish the 1-6 series from the 7-13, a most interesting result obtained without knowledge of the translations' origins. Besides, most of the human metrics manage to separate programs from people: for instance, fidelity (H5) does this rigorously on the 100 series, and most of them do it on the 200 series.

6 Conclusion: meta-evaluation of metrics

The results presented here are an attempt to analyze the behavior of metrics and the resources they need in order to provide a score. The use of BLEU/NIST shows how important the reference set is, but also how one can use the metric on a smaller set. From a theoretic point of view, if the goal of evaluation is to provide a mapping from systems to scores ({*systems*} → {*scores*}), in reality the input to evaluation is {*systems*} × {*test data*} × {*metrics*}, where the last term means itself either {*a-metrics*} × {*reference data*}, for automated metrics, or {*h-metrics*} × {*judges*} for human metrics. The goal is then to factor out the subjective elements and to provide objective scores, independent of judges, test data, and reference data. Comparative criteria and coherence criteria for metrics are useful to attain this goal. For instance, it is helpful to know the standard deviation of a metric, the score/rank correlation between metrics, but also the cost or the time needed to apply a metric.

Regarding individual metrics, the scores obtained by different evaluators using the same metric inform the community about the reliability of that metric. The interval of values generated by the metric, its correlation with human judgments of the translation quality, and the inter-annotator agreement (for human metrics) are all useful information for evaluators who need to chose metrics prior to an evaluation.

The other important result of this experiment is 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 judgment is almost the only proof of coherence). Overall, the questions targeted by this evaluation experiment bring a better knowledge of the evaluation metric, which is useful in taxonomization efforts such as the one initiated in the ISLE project.

7 Acknowledgments

The author is grateful to all the participants to the LREC 2002 MT Evaluation Workshop, in particular to those who contributed with results. The author would like to thank in particular the participants explicitly quoted in the paper: George Doddington (NIST, USA), Bonnie Dorr (UMIACS, Univ. of Maryland, USA), Eva Forsbom (Uppsala Univ., Sweden), Cristina Vertan (Univ. of Hamburg, Germany), and Michelle Vanni (US Government). This work was supported by the ISLE project (NSF, European Union, and Switzerland). Sandrine Zufferey (Univ. of Geneva) kindly helped with the formatting of the final version.

8 References

Doddington G. (2002): "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics", *HLT'02*, San Diego, CA.

Doyon J., Taylor K. B. and White J. S. (1998): "The DARPA MT Evaluation Methodology: Past and Present", *AMTA Conference*, Philadelphia, PA.

Hartley A. and Rajman M. (2001): "Automatically Predicting MT Systems Rankings Compatible with Fluency, Adequacy or Informativeness Scores", *Workshop on MTEval at MT Summit VIII*, Santiago de Compostela, Spain, pp. 29-34.

Niessen S., Och F. J., Leusch G. and Ney H. (2000): "An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research", *LREC 2000*, Athens, Greece, pp. 39-45.

Papineni K. (2002): "MT Evaluation: N-grams to the Rescue", *LREC 2002*, Las Palmas, Spain.

Papineni K., Roukos S., Ward T. and Zhu W.-J. (2001): *BLEU: a Method for Automatic Evaluation of Machine Translation*, Research Report, IBM Computer Science Research Division, T.J.Watson Research Center, RC22176 (W0109-022).

Rajman M. and Hartley A. (2002): "Automatic Ranking of MT Systems", *LREC 2002*, Las Palmas, Spain, pp. 1247-1253.

Reeder F., Miller K. J., Doyon J. and White J. S. (2001): "The Naming of Things and the Confusion of Tongues: an MT Metric", *Workshop on MTEval "Who did what to whom?" at MT Summit VIII*, Santiago de Compostela, Spain, pp. 55-59.

Somers H. and Prieto-Alvarez N. (2000): "Multiple Choice Reading Comprehension Tests for Comparative Evaluation of MT Systems", *Workshop on MT Evaluation at AMTA-2000*, Philadelphia, PA.

Van Slype G. (1979): *Critical Study of Methods for Evaluating the Quality of Machine Translation*, European Commission / Directorate for General Scientific and Technical Information Management (DG XIII), BR 19142.

Vanni M. and Miller K. J. (2002): "Scaling the ISLE Framework: Use of Existing Corpus Resources for Validation of MT Evaluation Metrics across Languages", *LREC 2002*, Las Palmas, Spain.