The NIST 2008 Metrics for Machine Translation Challenge – Overview, Methodology, Metrics, and Results

Mark Przybocki, Kay Peterson, Sébastien Bronsart, Gregory Sanders
{*Mark.Przybocki, Kay.Peterson, Sebastien.Bronsart, Gregory.Sanders*}@nist.gov

Multimodal Information Group, National Institute of Standards and Technology, Gaithersburg MD, U.S.A.

May 15, 2009

Abstract

This paper discusses the evaluation of automated metrics developed for the purpose of evaluating machine translation (MT) technology. A general discussion of the usefulness of automated metrics is offered. The NIST MetricsMATR evaluation of MT metrology is described, including its objectives, protocols, participants, and test data. The methodology employed to evaluate the submitted metrics is reviewed. The general classes of metrics that were evaluated are summarized. Overall results of this evaluation are presented, primarily by means of correlation statistics, showing the degree of agreement between the automated metric scores and the scores of human judgments. Metrics are analyzed at the sentence, document, and system level with results conditioned by various properties of the test data. This paper concludes with some perspective on the improvements that should be incorporated into future evaluations of metrics for MT evaluation.

1. Introduction

It is not inconceivable to claim that IBM's introduction of BLEU (1) in 2001 has had a greater impact on the advancement of statistical machine translation (MT) technology than any other single contribution to the field over the succeeding five years. BLEU was the first automated, and more importantly repeatable, metric to demonstrate general correlation with human judgments of translation quality (1), (2). As such, BLEU provided a means for instituting large-scale MT technology evaluations.[1] As the popularity of these evaluations grew, BLEU quickly became the de facto standard metric for MT evaluation.

Automated metrics have advantages over human assessments of MT quality. They are typically quick to implement and can be used to score large amounts of data with minimal human effort. Also, scoring MT output with automated metrics is repeatable – running the metric over the same data more than once produces identical results. These advantages make automatic metrics an integral tool for large-scale evaluations when limited resources are available for human assessments. Automated metrics also allow system developers to quickly assess the impact of a system modification, by translating the same source language dataset and scoring the MT system's output produced before and after the modification.

On the other hand, human assessments, especially from bilingual judges, are often accepted as the standard for evaluating translation quality, and they can be crafted to meet the needs of a specific application. Thus, human assessments have the potential to indicate the usefulness of the MT output.

[1] While other automatic metrics, such as sentence error rate (SER) and word error rate (WER) were used in MT research prior to BLEU, no studies demonstrating their correlation with human assessments of MT quality are readily available.

Both manual and automated metrics have their known flaws, too. While showing general correlation with human assessments of MT quality, the scores produced by most automated metrics are not intuitive. A certain BLEU score does not allow conclusions as to the actual quality, and thus potential real-world usefulness, of the machine translation. Several studies have investigated shortcomings of or proposed changes to current automatic metrics. Doddington (3) and Babych et al. (4) suggest incorporating frequency weighting into measures based on n-gram co-occurrence statistics such as BLEU. Callison-Burch et al. (5) demonstrate cases where BLEU scores do not correlate well with human assessments; they argue that BLEU and other automatic metrics provide too rough a measure of translation quality, as they do not use models that allow the variation that can always be found in translation. Chiang et al. (6) also point to cases where BLEU produces counter-intuitive results and propose modifications to the metric. It has been shown that some automated metrics produce better scores for a particular type of MT system than others, where human assessments may produce opposite results. For example, BLEU has been shown to yield higher scores for statistical MT (SMT) systems than for rule-based MT (RBMT) systems even when the human assessment scores are higher for the RBMT system. This is one of the reasons why metrics such as BLEU may be viewed as most appropriate for measuring the impact of changes to an MT system over time or comparing very similar SMT systems, rather than measuring differences between substantially different MT systems. Also, most automated metrics to date have been built around, and tested most extensively on, output in the English language. Their usefulness for other target languages, especially non-Indo-European languages that are structurally very different from English (such as Arabic or Chinese) or whose orthography is less standardized, is not yet well understood. Condon et al. (7) have shown that for Arabic, normalizations that take apart the morphology of Arabic words will increase the correlation between BLEU scores and human judgments of semantic adequacy.

Human assessments have two main disadvantages. The first is their high cost, both in manpower and time. The second is that scores from human assessments are not completely repeatable; while one would expect system rankings to remain the same or nearly the same with repeated human assessments, the exact scores will differ due to the inherent subjectivity of human assessments. Different judges will supply differing assessments (inter-judge disagreements), and upon repeating an assessment later, judges do not always assign the same scores as before (intra-judge disagreement). To account for the subjective nature of human assessments, one relies on assessments from multiple judges and uses statistical analyses to help account for judge differences. Ideally, one wants the judges to be highly correlated on average. For data that are roughly numeric, one can use Pearson's correlation. For categorical (nominal-scale) data, one can use Cohen's Kappa to measure agreement between pairs of judges (8), (9) or Fleiss' generalization to measure agreement among more than two judges (10).

Realizing the many benefits automated metrics could provide for improving MT technology, and acknowledging the growing list of concerns and identified shortcomings of the current metrics available, researchers have been on a constant quest to *develop improved methods that automatically evaluate machine translation quality*.

1.1. NIST Metrics for Machine Translation Challenge

The interest in improving MT metrology is evident in that other recent initiatives have included analyses of the correlation between different human assessments and different automatic metrics, as was done in recent WMT workshops (11), (12). The unique goal of the NIST Metrics for Machine Translation Challenge (MetricsMATR) is to focus *exclusively* on MT metrology research, bringing together the numerous research efforts conducted in the field of MT metrology, and helping to promote innovations

in the development of automated metrics. MetricsMATR serves as a forum for researchers to exchange ideas. (13)

Successfully capturing the strengths and weaknesses of MT metrics requires their analysis over large and varied data sets. For instance, one may want to measure the relative performance of metrics, conditioned to specific criteria, such as the source language, the type of MT system (statistical, rule-based, hybrid), and the data genre. MetricsMATR makes use of many data sets assembled from various NIST-coordinated MT evaluations. Each data set contains one or more reference translation(s), one or more machine translation(s), and one or more type(s) of human assessments.

The rationale behind the analyses performed for MetricsMATR is that the closer an automated metric models human judges, the better the metric. Consequently, human assessment plays a key role in this challenge. Different types of human assessments are available and will be used to compare the metrics scores against. Finding the best way to do human assessments is itself a major research challenge. Achieving acceptable intra- and inter-annotator agreement is one challenge; designing assessments that can be performed with a reasonable amount of time and effort is another. Recent initiatives that have performed investigations of intra- and/or inter-annotator agreement, and demonstrated a need for improving it, include the IWSLT 2006 evaluation campaign (14), the WMT07 (11) and WMT-08 (12) workshops, and the NIST OpenMT 2009 evaluation. WMT-08 had a specific focus on improving human assessment methods by increasing intra- and inter-annotator agreement and reducing assessment time (by assessment at a sub-sentential constituent level, a different approach than what is done in MetricsMATR where assessment is at the sentence level).

1.2. Outline

The second section of this paper describes the MetricsMATR data, including the human assessment types used for the correlation studies. Section 3 provides an overview of the evaluated metrics, both the submitted metrics and a set of existing baseline metrics. Section 4 describes the evaluation protocols, section 5 the results and section 6 summarizes the first MetricsMATR evaluation and suggests future directions for the evaluation of automated MT metrics.

2. MetricsMATR 2008 Data

The term “data” is used to refer to both the machine translations that are scored against human reference translations by the automated metrics, and the human assessments that the automatic metrics are to be compared against. The MetricsMATR machine translation data is described in section 2.1 which is followed by a description of the human assessments in section 2.2.

2.1. Machine Translation Data

As coordinator of multiple MT technology evaluations, NIST has access to a large and varied set of MT data submitted for formal evaluation. Table 1 describes the MetricsMATR evaluation test set; the components from the individual technology evaluations are described in detail in the subsections below.

Evaluation Name	Source Language	Target Language	Data Genre	Documents	Segments	References	Systems
MT08	Arabic	English	Newswire	22	228	4	10
MT08	Arabic	English	Web	20	177	4	10
MT08	Chinese	English	Newswire	29	263	4	10
MT08	Chinese	English	Web	23	344	4	10
GALE Phase 2	Arabic	English	Newswire	22	207	1	3
GALE Phase 2	Arabic	English	Web	23	262	1	3
GALE Phase 2	Chinese	English	Newswire	25	183	1	3
GALE Phase 2	Chinese	English	Web	22	209	1	3
GALE Phase 2.5	Arabic	English	Broadcast News	20	210	1	2
GALE Phase 2.5	Chinese	English	Broadcast Conversations	21	267	1	3
GALE Phase 2.5	Chinese	English	Broadcast News	21	221	1	3
TRANSTAC JAN07	Arabic	English	Dialog	15	433	4	5
TRANSTAC JUL07	Arabic	English	Dialog	47	419	4	5
TRANSTAC JUL07	Farsi	English	Dialog	25	414	4	5

Table 1: MetricsMATR 2008 evaluation data statistics are shown. "Systems" represent the number of translations per segment. For the MT08 and TRANSTAC data sets, where multiple references are used, two references were included as systems in the single reference track.

Although the MetricsMATR evaluation encouraged the development of metrics applicable to a variety of target languages, the original MetricsMATR evaluation plan stated that analysis would be limited to the scoring of English translations. A secondary data set made available for use from ELDA/ELRA had the target language as French, but this data is not discussed in this paper. The ELDA/ELRA data is planned for full inclusion in the next MetricsMATR test.

2.1.1. MT08 data set

The MT08 data was drawn from systems that voluntarily processed the MT08 Progress test set used in the NIST 2008 Open MT evaluation.[2] The Progress test set is a special data set developed to allow more meaningful analysis of year-to-year progress in MT technology, by reusing the same test data with special protections in place to insure the test data remains strictly unseen. The systems included in the MetricsMATR test data were chosen to provide a range of MT translation quality, as determined by a set of commonly used automated MT metrics.

There were two motivations for selecting translations from the Progress test set. First, the data was used in an instantiation of the DLPT-star MT comprehension test (15) (described in section 2.2.6) providing another aspect of human assessment available for analysis, and second, the Progress test set will be reused over time providing the opportunity to update the MetricsMATR data set with more recent and presumably better translations of the same source data.

With MT08 being an open evaluation it was possible to select translations from several systems (10). And the MT08 data had 4 high-quality reference translations, allowing the MT08 system translations to be used in both the single reference and the multiple reference evaluation tracks of MetricsMATR, both of which are described in section 4.2.

[2] See http://www.nist.gov/speech/tests/mt/2008 for details of the MT08 evaluation and the Progress test set.

MT08 data represents translations of text-to-text MT technology. The source language was either Arabic or Chinese, with the target language being English. OpenMT evaluations traditionally use data with a news focus, but in recent years, more informal types of data, such as data extracted from weblogs and on-line forums have been included. Both types of data were in the MetricsMATR test set.

MT08 data represents 48% of the MetricsMATR test data. Translations from several available systems contribute to the high percentage.

2.1.2. GALE data set

The GALE data was drawn from translations produced during phase 2 (P2) of the Global Autonomous Language Exploitation (GALE) evaluation.[3]

The inclusion of data from GALE provided added benefit to MetricsMATR since the GALE program's official metric was a form of human assessment, HTER (16), providing another form of analysis in the determination of metric usefulness.

The GALE P2 data represents translations from text-to-text MT systems, while P2.5 represents translations from speech-to-text MT evaluations. Note that only the resulting translations (text data) were used in MetricsMATR tests.

As with the MT08 data, the target language was English, which was translated from one of the two source languages, Arabic or Chinese. The P2 data contained the same type of text data as MT08, newswire and web data.

GALE data represents 17% of the MetricsMATR test set; relatively few systems were available.

2.1.3. TRANSTAC data set

The TRANSTAC data was drawn from systems submitted for the January and July 2007 Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) evaluations.[4]

TRANSTAC data is unique to the MetricsMATR test data in that it represents free form dialogs. These dialogs were collected between English speaking military personnel and native speakers of Iraqi Arabic and Farsi. Although these collections were bi-directional, only the translations into English were included in MetricsMATR 2008. In addition, only the text translations (input to speech synthesis systems in the TRANSTAC evaluation) were used in MetricsMATR.

TRANSTAC data represents 35% of the MetricsMATR test set.

2.2. Human Assessments

A key component of the MetricsMATR evaluation was the production of meaningful human assessments to serve as the evaluation's reference key. The automatic metrics were evaluated by how well they model these human assessments. Two types of assessment were implemented specifically for MetricsMATR, while several others that were created for the individual technology evaluations were also available.

NIST designed and the Linguistic Data Consortium (LDC) implemented the two types of human assessments that were judged over the entire MetricsMATR test set. We refer to these assessments as segment level *Adequacy* and pair-wise *Preferences*.

[3] See http://www.nist.gov/speech/tests/GALE for details related to NIST coordinated GALE MT evaluations.
[4] See http://www.darpa.mil/IPTO/programs/transtac/transtac.asp for details of the TRANSTAC program.

In addition to these assessments, MT08 data was previously scored using a MT comprehension test, DLPT-star (15); the GALE data set was previously assessed using HTER; and the TRANSTAC portion of the evaluation set was annotated for segment level adequacy on a 4-point scale and for adjusted probability that a concept was correctly relayed. The assessments are summarized in Table 2 and described in more detail below.

Human Assessment Type	Short Name	MT08	GALE	TRANSTAC
Adequacy (7-pt, Y/N question)	**Adequacy7, AdequacyYN**	**Yes**	**Yes**	**Yes**
Preferences	**Preferences**	**Yes**	**Yes**	**Yes**
HTER	**HTER**	No	**Yes**	No
Adequacy (4-pt)	**Adequacy4**	No	No	**Yes**
Adjusted probability that a concept is correct	**AdjProbCorr**	No	No	**Yes**
DLPT-star	**DLPT***	**Yes**	No	No

Table 2: Human Assessment types available for the MetricsMATR 2008 evaluation data sets

2.2.1. Adequacy (7-pt scale, Y/N question)

For this assessment, judges were presented with the human reference translation and one candidate translation to evaluate. The first question asked was a "quantitative" adequacy question:

"How much of the meaning expressed in the Reference translation is also expressed in the System translation?"

The answer was recorded using a 7-point scale with the extremes and mid-point labeled as "None" (1), "Half" (4), and "All" (7). Then, a second, a more "qualitative" question was asked:

"Does the Machine translation mean essentially the same as the Reference translation?"

The answer was recorded using a binary Yes/No choice. When the answer to the first question ranged from "None" to "Half", the second question was not asked and the answer was set to "No".

Judges assessed the quality of the translations one segment at a time. All segments were presented in the order of appearance of a given document, and each segment received decisions from two judges.

Adequacy7 was the average of all judges' scores for a given segment. For document and system level values, an average of the segment level scores weighted by the number of reference tokens was used.

AdequacyYN was the ratio of the number of 'Yes' judgments for a given segment, to the total number of judgments, across all judges. Counts were aggregated to obtain document and system level ratios.

On average a judge spent 23 seconds to assess a single segment for Adequacy. This included both the **Adequacy7** and the **AdequacyYN** decision. Inter-judge agreement is shown in Figure 1. We calculate the distance in categories between two judges that assessed the same segment. When both judges assign the same score we use the term "exact match". When the judge's scores differ by one category, we use the term "1-off, by 2 "2-off", and so on. As seen in Figure 1 the majority of the judgments were very close, with approximate 78% falling into either the exact-match or 1-Off category. Less than 1.5% of all the judgments differed by more than 3 categories.

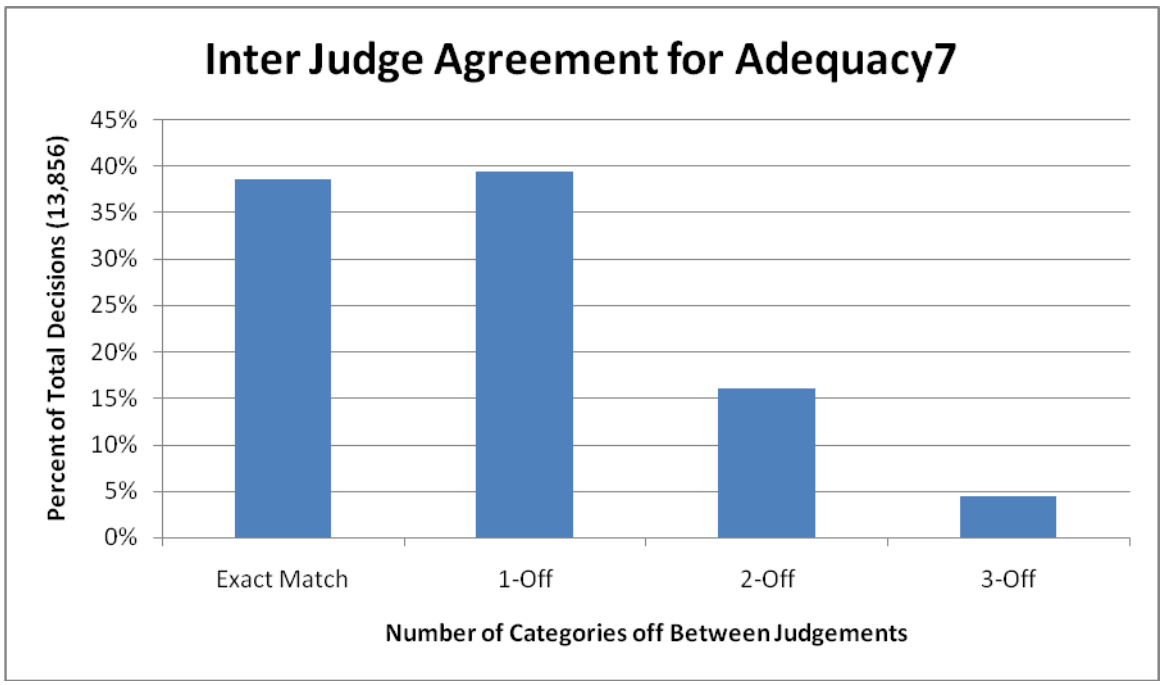


Figure 1: Inter-judge agreements for the Adequacy7 assessment task. A linear weighted Kappa value was calculated for each category. Requiring exact match Kappa = .25, allowing for 1-off Kappa =.58, and allowing for 2-off Kappa = .80.

2.2.2. Preference

For this task, judges were asked to express their preference between two candidate translations when compared to a human reference translation. Judges were presented with all possible pair-wise comparisons for a given segment, with segments being presented in the order in which they appeared in the document and with the order of system comparisons appearing randomly. Due to the large number of comparisons, this task was limited to the first four segments of each document.

The question asked of the judge was: *"Which translation do you prefer?"*

Judges could select a preference for one of the two system translations, or they could choose no preference when the translations were equally good or equally bad – a choice made about once in every four decisions. At least one judgment per segment was collect for each possible system-to-system comparison and for 5,826 comparisons two decisions were collected. In these instances the judges disagreed 10% of the time, including when one judge selected a system as preferred and the other did not have a preference.

Preferences was the number of times a given system segment was preferred, divided by the total number of comparisons involving the system. Counts were aggregated to obtain document and system level ratios.

2.2.3. Adjusted probability that a concept is correct

This measure was used in the TRANSTAC evaluations, thus assessments were limited to the TRANSTAC portion of the evaluation set. For this measure, experts identified concepts in the source data. Bilingual judges then match these against the system translation as correct, substituted, or deleted; they also marked inserted concepts. The adjusted probability that a concept is correct was computed from counts of correctly translated concepts and errors (17).

2.2.4. Adequacy (4-pt scale)

Another measure used in the TRANSTAC evaluations was adequacy judgments made on a four point scale in which bilingual judges went through a two-step decision process comparing the source data to a system translation. First, judges decided whether the translation was more adequate or more inadequate, then whether it was completely or tending adequate, or tending inadequate or completely inadequate (17).

Adequacy4 was the weighted average (by reference segment length) of all judges' scores, per segment.

2.2.5. HTER

This measure was available for the GALE portion of the data. HTER stands for "Human Targeted Translation Edit Rate" (16). For the HTER annotations, a human assessor compared a system translation to a reference translation, and edited the system translation such that it would have the same meaning as the reference translation. This was emphasized to be done with as few edits as possible. The number of needed edits (insertions, deletions, substitutions, and shifts) was then measured automatically using the TER (Translation Edit Rate) (16) measure.

2.2.6. DLPT-star MT Comprehension Test

This measure was available for the MT08 portion of the evaluation data. DLPT-star (15) is a MT comprehension test that uses questions developed from the source data. Subjects answered the test questions based translated output and their answers were graded by test experts. Correlations of automatic metrics with the DLPT-star test are not analyzed in this paper.

3. MetricsMATR 2008 Metrics

There were 39 metrics evaluated in the MetricsMATR 2008 evaluation, seven of which were preexisting baseline metrics, that is, metrics that have been used prominently in past evaluations. The remaining 32 metrics were submitted by the participants. Each metric is referred to by a unique identifier (in bold).

3.1. Baseline metrics

3.1.1. Variants of the BLEU metric

BLEU (1) is a precision-based metric that counts the number of *n*-grams (sequences of *n* consecutive tokens) that a candidate translation and a corresponding reference translation have in common. The different precision scores (one per *n*-gram length) are combined using the geometric mean. Once the overall precision score is computed, a brevity penalty is computed over the entire corpus. The purpose of this brevity penalty is to penalize candidate translations that are shorter (overall) than the reference translations.

MetricsMATR evaluated four baseline variants of the BLUE metric using case-sensitive scoring:

- **BLEU-1**: (IBM version 1.04) limited to unigram precision
- **BLEU-4**: (IBM version 1.04) precision scores for *n*-grams of size between 1 and 4 tokens
- **BLEU-v11b**: (NIST mteval-v11b) similar to BLEU-4, but with a modified brevity penalty
- **BLEU-v12**: (NIST mtevale-v12) similar to BLEU-v11b, but with a modified tokenization scheme

3.1.2. NIST score

The NIST score (3) was the official metric in early DARPA TIDES MT evaluations. It is based on information weighted *n*-gram co-occurrences. Some of the differences between BLEU and the NIST score include the method of co-occurrence measures (arithmetic mean replacing geometric mean), a modified brevity penalty, and a modified weighting of n-grams, depending on the frequency of specific n-grams.

- **NIST-v11b**: (NIST mteval-v11b) scores case-sensitive *n*-grams of size varying between 1 and 5

3.1.3. TER

TER (16) is a measure of edit distance which captures the number of edits required to make a candidate translation *identical* to a reference translation, counting block moves as a single error. Scoring was case sensitive and uses similar text normalization as the variants of BLEU.

- **TER-v0.7.25**: (BBN/UMD version 0.7.25) TERCOM scoring software[5]

3.1.4. METEOR

METEOR (18) was one of the first metrics developed to use additional lexical information (synonymy information from WordNet (19)), and additional syntactic information (stemming using the Snowball stemmers (20)) to enhance word matching. METEOR can use different mapping modules to find the optimal word-to-word matches. The final score is computed as a combination of precision and recall over the unigram matches.

- **METEOR-v0.6**: (CMU version 0.6[6]) modules used: exact, porter, wn_stem and wn_synonymy

Version 0.6 was not configured to output document level scores. As a substitute, document level scores were computed as the un-weighted average of all segment scores of a given document. The authors of METEOR entered an updated version in the MetricsMATR evaluation.

3.2. Submitted metrics

Table 3 lists the metrics submitted for evaluation. The affiliations submitting the metrics and the metric names are identified. Several of these metrics have corresponding papers in this issue of MT Journal.

Affiliation	Metric name(s)
BabbleQuest	Badger, BadgerLite
Carnegie Mellon University	METEOR-v0.7, METEOR-ranking, mBLEU, mTER
City University of Hong Kong, Department of Chinese, Translation and Linguistics	ATEC1, ATEC2, ATEC3, ATEC4
Columbia University	SEPIA1, SEPIA2
Harbin Institute of Technology, School of Computer Science and Technology	SVM-Rank, SNR, LET
National University of Singapore	MaxSim
RWTH Aachen University	BleuSP, invWer, CDer
Stanford University	RTE, RTE-MT

[5] http://www.cs.umd.edu/~snover/tercom
[6] http://www.cs.cmu.edu/~alavie/METEOR

University of Maryland / BBN Technologies	TERp
Universitat Politècnica de Catalunya, LSI	ULCh, ULCopt, DP-Or, SR-Or, DR-Or, DP-Orp
USC, Information Sciences Institute (Team 1)	BEwT-E
USC, Information Sciences Institute (Team 2)	Bleu-sbp, 4-GRR
University of Washington	EDPM

Table 3: List of submitted metrics and the affiliation who created them

4. MetricsMATR 2008 Evaluation Protocols

The evaluation specification document (13) created for the MetricsMATR 2008 evaluation thoroughly described the data, tasks, any rules and restrictions that applied, and the evaluation protocols to be followed. Such a document is necessary to ensure a smooth implementation of the evaluation. In this section, we review the key elements described in the evaluation specification document.

4.1. Metric Development

NIST MT technology evaluations require that a set of rules and restrictions be obeyed during system development such that systems may be directly compared. This scenario does not apply to the development of new metrology techniques. In fact, for MetricsMATR, NIST encouraged the use of innovative techniques, welcoming submissions from a wide range of disciplines not limited to those with an interest in MT technology development. The requirements that were enforced for MetricsMATR were developed for ease of metric implementation, not to restrict metric development.

4.1.1. Requirements

There were four basic restrictions that NIST required for each submitted metric.

Rule #1: Metrics were required to accept as input NIST OpenMT XML formatted files.

Since metrics were installed at NIST to be run locally over the MetricsMATR test set, it was important that they accept the standard file format allowing for trouble-free invocation of each metric.

Rule #2: Metrics were to output *segment*, *document*, and *system* level scores in a prescribed format.

The three sets of outputs were required for various levels of analysis. In a few cases, a particular metric was not designed to calculate scores at each of the three levels. In such cases, the metric score was meaningless for the particular level and it was noted in the system description. The requirement that the output files follow a prescribed format was to aid batch comparisons. Table 4 lists the tab-separated information required in an output record for each of the three levels of scores.

Level	Field #1	Field #2	Field #3	Field #4	Field #5	Field #6
System	Test_ID	System_ID	Score	<optional>		
Document	Test_ID	System_ID	Doc_ID	Score	<optional>	
Segment	Test_ID	System_ID	Doc_ID	Seg_ID	Score	<optional>

Table 4: Record contents for the three levels of score reports

Rule #3: Metric software was to run on one of three predetermined operating systems.

NIST identified three common system architectures that were available for the running of submitted metrics. The first, and most commonly used, was a Linux CENT OS 5 (or newer) system. The second, which was required for two metrics, was a Windows XP system. The third available system was a MAC OS X system. NIST did not test the same metric on each of the three system architectures.

Rule #4: Reasonable installation effort

This rule was included to guard against the NIST installation process being used as a debugging service. The rule stated that a metric should be installed, compiled, and tested within approximately four hours, and all required support programs and scripts would be compatible with the most recent releases.

4.1.2. Properties

To assist potential metric developers, NIST identified properties that useful MT metrics have in common:

- *Automaticity*: Metrics that do not require human intervention outside the creation of the reference translation are useful to evaluate systems over large test sets and can be used in training by certain MT algorithmic approaches.
- *Repeatability (Reliability)*: Metrics that produce the same score every time they process the same set of data are useful in determining progress.
- *Portability*: Metric software should be universally available.
- *Speed*: To the extent possible, metrics should be quick to run.
- *Limited Annotation of the Reference Data*: Metrics had at their disposal up to four high quality reference translations for the MetricsMATR evaluation. Metrics that required additional annotation on the reference (or source) data in order to provide insights into quality were to use an automated mark-up scheme (which if proven useful, NIST would consider for full manual mark-up in later MetricsMATR tests).

4.1.3. Objectives

Our search for new and improved automated MT metrics was motivated by what we found missing from current approaches.

High Correlation with Human Assessments of MT Quality: Assessment by human judges is the most widely accepted standard for the definitive evaluation of MT quality. Automated metrics that correlate very highly with qualified human assessments of translation quality are useful as surrogate measures for the slow and expensive process of obtaining assessments from humans, preferably bilingual, judges.

Ability to Differentiate between Systems of Varying Quality: To the extent possible, metrics should be able to differentiate the quality of two different systems. Metrics should be sensitive enough, and the scores that they report should be fine-grained enough, to rank-order systems that are fairly close in quality. This property should apply across the range of poor to high quality translations.

Intuitive Interpretation: Ideally, the information that metrics report (the scores) would be meaningful standing on its own. A complaint levied against most current automatic MT metrics is that a particular score value reported does not give insights into quality, nor is it easy to understand the practical significance of a difference in scores. For most metrics, all one can say is that higher is better for accuracy metrics and lower is better for error metrics.[7]

Applicable to Multiple Target Languages: Current automated metrics, especially those that were designed to exploit linguistically-motivated data, have been developed primarily for the evaluation of translations into English. These metrics may have linguistic characteristics of English (and, more generally, Indo-European languages) in their underlying assumptions. For example, n-grams have little relevance for languages with free word order. Similarly, metrics that count individual words as correct or wrong will score harshly translations into languages that aggregate multiple elements of meaning into

[7] Some metrics have better intuitive interpretations; these include the Word Error Rate (WER) metric for automatic speech recognition and the TER and HTER metrics for machine translation.

one word.[8] Further, there is a lack of equality in the linguistic resources a metric might use for evaluation of the many possible target languages.

4.2. Evaluation Tracks

One of the difficulties in measuring translation quality is in determining what the best translation should be, given that there are many acceptable variations of a translation. Several approaches have been pursued to assist automatic metrics in comparing a system translation to the full space of possible correct translations. The most common method has been to compare the system translation against multiple, independently created, reference translations. Other methods include providing alternatives for ambiguous text or idioms. Some metrics attempt to generate this large space of translation possibilities themselves by making use of stemmers, synonymy, and phrasal reordering techniques.

4.2.1. Single Reference Track

A benefit could be gained for large scale MT technology evaluations if a single high-quality reference translation was all that was required to assess MT quality. To this end, MetricsMATR tested the submitted metrics when limiting the reference set to a single high quality translation.

4.2.2. Multiple Reference Track

Traditionally, MT technology evaluations have placed the burden of alternative translations on the reference set, providing multiple, independently created translations to account for translation variability. Metrics were developed to exploit a set of multiple translations when calculating the quality of a system translation.

MetricsMATR included a track that tested the submitted metrics when four independent reference translations were available. Results are contrasted with those from the single reference track.

4.3. Processing the Evaluation Data

We acknowledge that the performance of a metric might be tied to the method the evaluator uses to score the evaluation set. This section reviews the method that NIST used in MetricsMATR 2008 to score the evaluation test set with each metric.

The approach taken mimicked what is typically done for evaluation where a specific metric is run over a well-defined set of translations for a single system. Scores are produced for that system over that set of data. Using the MT08 evaluation test set as an example, a system produced translations for Arabic newswire data and Arabic web data. While the two sets of translations could be combined and scored as one set, for MetricsMATR, first the references and system translations were limited to the MT08 Arabic newswire data and scores for each metric were produced. Then the references and system translations were limited to the MT08 Arabic web data and scores for each metric were produced.

This is important because some metrics may make use of the reference data to define weights for the internal language model scoring. For these metrics, the more data to be scored, the smoother the weighting will be.

[8] For example, the Arabic word للبرنامج (pronounced lilbernamij) means “to the program” and consists of three elements (li al bernamij) with those three pieces of meaning. An error on any one of the three would cause typical MT metrics to score the whole word as wrong, thus in effect scoring all three pieces of meaning as incorrect.

4.4. Schedule

MetricsMATR was a fourteen month endeavor. The key milestone dates were:

- SEP-2008, planning began
- JAN-2008, call for participation issued
- MAR-2008, development data released
- AUG-2008, developer's declaration of plans to participate deadline
- AUG-SEP-2008, metric installation
- OCT-2008, analysis and report generation
- OCT-2008, evaluation workshop

4.5. Types of Analysis

Metrics were evaluated separately at the segment level (useful for error analysis), document level (most natural unit of data), and system level (as is done in a formal evaluation).

Metrics were evaluated as to the level to which they agree with human judgments of quality, using several correlation statistics. A metric's ability to distinguish quality between systems was determined using significance tests at the document and segment level.

In addition to the evaluation tracks, single vs. multiple references, results are reported for each well defined subset of data, including by evaluation test set, genre, and technology type. Metric correlations are reported[9] for many of the human assessments available for MetricsMATR 2008.

4.6. Workshop

MetricsMATR 2008 concluded with a workshop, held in conjunction with the AMTA 2008 Conference.[10] The workshop brought together the organizers and participants of MetricsMATR and those interested in MT metrology development. Results were reviewed and metric talks were given. Much of the workshop discussion was geared toward improvements and planning for the next MetricsMATR evaluation, currently planned for the spring of 2010.

5. MetricsMATR 2008 Evaluation Results

Thirty-nine metrics were evaluated in the single reference track, 32 newly submitted metrics and 7 baseline metrics. All but two[11] of the metrics were also evaluated in the multiple reference track. In the following sections, we examine the correlation results, significance test results, and other findings.

5.1.Correlation results

We report correlations for the Spearman's rho statistic (Spearman's correlation coefficient for ranked data) as our primary correlation measure. Although we found that the correlations statistics for Spearman's Rho, Kendall's Tau, and Pearson's R to closely track each other, Spearman's provides the benefit of not showing sensitivity to outliers (as does Pearson's R*)*, and being based on ranks, Spearman's does not assume samples from a bivariate normal distribution (21).

Table 5 and Table 6 show the correlation results for the single and multiple reference tracks, respectively, when comparing the metrics scores to the *Adequacy7* judgments over the entire

[9] http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008/results
[10] http://www.amtaweb.org/AMTA2008.html
[11] Metrics **RTE** and **RTE-MT** were designated to be run in the single reference track only.

MetricsMATR evaluation set. **Bolding** is used to identify the highest correlation at each level of analysis.

Rank	Metric	Spearman's rho (95% confidence interval)		
		Segment level (25473 data points)	**Document level (2179 data points)**	**System level (89 data points)**
1	TERp	**-0.68** (-0.69, -0.68)	-0.81 (-0.83, -0.80)	-0.87 (-0.91, -0.81)
2	METEOR-v0.6	**0.68** (0.67, 0.69)	0.81 (0.80, 0.83)	0.89 (0.83, 0.93)
3	METEOR-ranking	0.67 (0.66, 0.68)	**0.84** (0.83, 0.85)	0.89 (0.84, 0.93)
4	Meteor-v0.7	0.67 (0.66, 0.67)	**0.84** (0.83, 0.85)	**0.90** (0.85, 0.93)
5	CDer	-0.65 (-0.66, -0.65)	**-0.84** (-0.85, -0.82)	**-0.90** (-0.94, -0.86)
6	EDPM	0.64 (0.63, 0.64)	0.81 (0.80, 0.83)	0.88 (0.82, 0.92)
7	SEPIA1	0.63 (0.62, 0.64)	0.81 (0.80, 0.83)	0.87 (0.81, 0.91)
8	LET	0.63 (0.62, 0.64)	0.80 (0.78, 0.81)	0.88 (0.82, 0.92)
9	SEPIA2	0.63 (0.62, 0.64)	0.81 (0.79, 0.82)	0.86 (0.80, 0.91)
10	BleuSP	0.62 (0.61, 0.63)	0.79 (0.77, 0.81)	0.85 (0.78, 0.90)
11	BLEU-1	0.62 (0.61, 0.63)	0.80 (0.79, 0.82)	0.86 (0.80, 0.91)
12	NIST-v11b	0.62 (0.61, 0.63)	0.81 (0.80, 0.83)	0.88 (0.82, 0.92)
13	SVM-Rank	0.61 (0.60, 0.62)	0.79 (0.78, 0.81)	0.88 (0.83, 0.92)
14	RTE-MT	0.61 (0.60, 0.61)	0.69 (0.67, 0.71)	0.70 (0.57, 0.79)
15	invWer	-0.60 (-0.61, -0.59)	-0.81 (-0.82, -0.79)	-0.89 (-0.93, -0.84)
16	ATEC1	0.58 (0.58, 0.59)	0.67 (0.64, 0.69)	0.84 (0.76, 0.89)
17	4-GRR	0.58 (0.57, 0.59)	0.78 (0.76, 0.79)	0.86 (0.79, 0.90)
18	ATEC2	0.58 (0.57, 0.59)	0.67 (0.64, 0.69)	0.84 (0.76, 0.89)
19	BLEU-4	0.58 (0.57, 0.59)	0.77 (0.75, 0.79)	0.84 (0.77, 0.89)
20	ATEC4	0.58 (0.57, 0.59)	0.66 (0.64, 0.69)	0.83 (0.75, 0.89)
21	TER-v0.7.25	-0.58 (-0.59, -0.57)	-0.79 (-0.81, -0.78)	-0.89 (-0.93, -0.83)
22	ATEC3	0.57 (0.57, 0.58)	0.67 (0.64, 0.69)	0.87 (0.81, 0.92)
23	RTE	0.57 (0.56, 0.57)	0.66 (0.63, 0.68)	0.62 (0.48, 0.74)
24	BEwT-E	0.49 (0.48, 0.50)	0.65 (0.63, 0.68)	0.78 (0.68, 0.85)
25	MaxSim	0.46 (0.45, 0.47)	0.53 (0.49, 0.56)	0.59 (0.43, 0.71)
26	ULCopt	0.46 (0.45, 0.47)	0.48 (0.45, 0.51)	0.56 (0.40, 0.69)
27	ULCh	0.45 (0.44, 0.46)	0.46 (0.42, 0.49)	0.54 (0.38, 0.67)
28	SNR	0.45 (0.44, 0.46)	0.50 (0.47, 0.53)	0.56 (0.40, 0.69)
29	DP-Or	0.45 (0.44, 0.46)	0.49 (0.46, 0.52)	0.58 (0.43, 0.71)
30	BadgerLite	0.44 (0.43, 0.45)	0.55 (0.52, 0.58)	0.69 (0.56, 0.79)
31	BLEU-v12	0.44 (0.43, 0.45)	0.78 (0.76, 0.79)	0.86 (0.79, 0.90)
32	BLEU-v11b	0.43 (0.42, 0.44)	0.77 (0.76, 0.79)	0.85 (0.78, 0.90)
33	Bleu-sbp	0.43 (0.42, 0.44)	0.78 (0.76, 0.79)	0.87 (0.81, 0.91)
34	SR-Or	0.40 (0.39, 0.41)	0.45 (0.42, 0.49)	0.51 (0.34, 0.65)
35	DR-Or	0.39 (0.38, 0.41)	0.41 (0.37, 0.44)	0.50 (0.33, 0.64)
36	mBLEU	0.39 (0.38, 0.40)	0.52 (0.49, 0.55)	0.69 (0.56, 0.79)
37	Badger	0.39 (0.38, 0.40)	0.53 (0.50, 0.56)	0.66 (0.53, 0.77)
38	DP-Orp	0.33 (0.32, 0.34)	0.33 (0.30, 0.37)	0.47 (0.29, 0.62)
39	mTER	-0.33 (-0.34, -0.32)	-0.50 (-0.53, -0.47)	-0.68 (-0.78, -0.56

Table 5: Overall correlation results against Adequacy7 judgments, for the single reference track. Metrics are ordered by segment level correlation. (Negative correlations represent error metrics.)

Rank	Metric	Spearman's rho (95% confidence interval)		
		Segment level (16450 data points)	Document level (1375 data points)	System level (55 data points)
1	METEOR-v0.6	**0.72** (0.71, 0.73)	0.77 (0.75, 0.79)	0.85 (0.75, 0.91)
2	SVM-Rank	**0.72** (0.71, 0.73)	0.79 (0.77, 0.81)	0.83 (0.72, 0.89)
3	Meteor-v0.7	**0.72** (0.71, 0.72)	0.84 (0.82, 0.85)	0.88 (0.81, 0.93)
4	CDer	-0.71 (-0.72, -0.71)	**-0.85** (-0.87, -0.84)	-0.92 (-0.95, -0.86)
5	TERp	-0.71 (-0.72, -0.71)	-0.81 (-0.83, -0.79)	-0.87 (-0.92, -0.78)
6	METEOR-ranking	0.71 (0.70, 0.72)	0.82 (0.80, 0.84)	0.84 (0.75, 0.91)
7	BleuSP	0.69 (0.68, 0.69)	0.80 (0.79, 0.82)	0.84 (0.73, 0.90)
8	SEPIA1	0.67 (0.67, 0.68)	0.83 (0.81, 0.85)	0.90 (0.84, 0.94)
9	LET	0.67 (0.67, 0.68)	0.80 (0.78, 0.82)	**0.93** (0.89, 0.96)
10	EDPM	0.67 (0.66, 0.68)	0.83 (0.81, 0.84)	0.92 (0.87, 0.95)
11	SEPIA2	0.67 (0.66, 0.67)	0.83 (0.82, 0.85)	**0.93** (0.88, 0.96)
12	invWer	-0.66 (-0.67, -0.66)	-0.83 (-0.84, -0.81)	-0.91 (-0.95, -0.85)
13	NIST-v11b	0.65 (0.64, 0.66)	**0.85** (0.83, 0.86)	**0.93** (0.89, 0.96)
14	ATEC4	0.65 (0.64, 0.66)	0.67 (0.64, 0.70)	0.90 (0.83, 0.94)
15	ATEC1	0.65 (0.64, 0.65)	0.68 (0.65, 0.70)	0.90 (0.84, 0.94)
16	ATEC2	0.64 (0.63, 0.65)	0.67 (0.64, 0.70)	0.90 (0.84, 0.94)
17	ATEC3	0.64 (0.63, 0.65)	0.68 (0.66, 0.71)	**0.93** (0.89, 0.96)
18	BLEU-1	0.63 (0.62, 0.64)	0.82 (0.80, 0.84)	0.91 (0.85, 0.95)
19	BLEU-4	0.62 (0.61, 0.63)	0.78 (0.76, 0.80)	0.89 (0.82, 0.94)
20	4-GRR	0.62 (0.61, 0.63)	0.74 (0.71, 0.76)	0.84 (0.73, 0.90)
21	TER-v0.7.25	-0.59 (-0.60, -0.58)	-0.76 (-0.78, -0.74)	-0.90 (-0.94, -0.83)
22	BEwT-E	0.57 (0.56, 0.58)	0.77 (0.75, 0.79)	0.92 (0.87, 0.96)
23	BLEU-v12	0.51 (0.49, 0.52)	0.78 (0.76, 0.80)	0.90 (0.84, 0.94)
24	BLEU-v11b	0.50 (0.49, 0.51)	0.78 (0.76, 0.80)	0.90 (0.83, 0.94)
25	Bleu-sbp	0.50 (0.49, 0.51)	0.79 (0.77, 0.81)	0.91 (0.85, 0.95)
26	mBLEU	0.45 (0.44, 0.46)	0.69 (0.66, 0.71)	0.85 (0.76, 0.91)
27	mTER	-0.39 (-0.40, -0.38)	-0.65 (-0.68, -0.61)	-0.86 (-0.92, -0.77)
28	BadgerLite	0.33 (0.32, 0.35)	0.18 (0.13, 0.23)	0.25 (-0.02, 0.48)
29	Badger	0.29 (0.28, 0.31)	0.16 (0.11, 0.21)	0.16 (-0.12, 0.40)
30	MaxSim	0.25 (0.23, 0.26)	0.12 (0.07, 0.17)	0.16 (-0.11, 0.41)
31	DP-Or	0.23 (0.22, 0.25)	0.02 (-0.03, 0.08)	0.12 (-0.15, 0.38)
32	SNR	0.21 (0.20, 0.23)	0.05 (-0.01, 0.10)	0.06 (-0.21, 0.32)
33	SR-Or	0.20 (0.19, 0.22)	0.02 (-0.04, 0.07)	0.12 (-0.15, 0.38)
34	ULCopt	0.20 (0.19, 0.22)	0.02 (-0.03, 0.08)	0.12 (-0.16, 0.37)
35	ULCh	0.20 (0.18, 0.21)	-0.01 (-0.06, 0.05)	0.13 (-0.14, 0.38)
36	DR-Or	0.20 (0.18, 0.21)	0.02 (-0.04, 0.07)	0.12 (-0.15, 0.37)
37	DP-Orp	0.14 (0.13, 0.16)	-0.06 (-0.11, 0.00)	0.09 (-0.19, 0.34)

Table 6: Overall correlation results against Adequacy7 judgments, for the multiple reference track. Metrics are ordered by segment level correlation.

These results indicate that metrics correlated higher to the human assessments of adequacy at lower (system) levels of granularity than they do at the higher (segment) granularity. This was always the case in the single reference track, and was true for the top 27 metrics in the multiple reference track.

Another observed trend is that the correlations for *most* metrics increased when using multiple references, which occurred at the segment level in 27 of the 37 metrics evaluated. This may not be surprising since many metrics were designed for use with multiple references. The interesting case is when a metric's correlation was stable between the two tracks. In MetricsMATR, 12 metrics had less than 5% difference in Spearman's correlations at the segment level between the single and multiple reference tracks.

We found that metrics were not robust across conditions; no single metric outperformed all other metrics for a given track across all correlation levels. Table 7 summarizes the three highest correlating metrics for each evaluation track, with regard to the analysis level.

Level of Analysis	Single Reference Track		Multiple reference track	
Segment level	TERp	-0.68	METEOR-v0.6	0.72
	METEOR-v0.6	0.68	SVM-Rank	0.72
	METEOR-ranking	0.67	Meteor-v0.7	0.72
Document level	Meteor-v0.7	0.84	CDer	-0.85
	METEOR-ranking	0.84	NIST-v11b	0.85
	CDer	-0.84	Meteor-v0.7	0.84
System level	CDer	-0.90	ATEC3	0.93
	Meteor-v0.7	0.90	LET	0.93
	invWer	-0.89	NIST-v11b	0.93

Table 7: Spearman's rho for the three highest correlating metrics, per evaluation track and analysis level Table 8: Spearman's rho for the 10 highest correlating metrics, for each evaluation track and level of analysis, conditioned on the evaluation data set

Table 8 displays the ten highest correlating metrics for each evaluation track conditioned on the evaluation data set: MT08, each phase of GALE, and each TRANSTAC evaluation. Correlations for the three levels of analysis are shown. Correlations for the GALE P2.5 data (speech-to-text) consistently lag the other data sets evaluated, across all levels of analysis in the single reference track.

Table 9 is a similar table, but with a different set of highest correlating metrics when conditioned on the type of human assessment. This is another example of how the metrics were not robust across conditions. Metric correlations with the *Adequacy7* judgments were consistently high. Metric correlations with the *Preference* judgments were found to be consistently lowest.

Table 10 conditions the results on the source language, either Arabic, Chinese, or Farsi, where we found notably less variation in correlation rates regardless of the original source language.

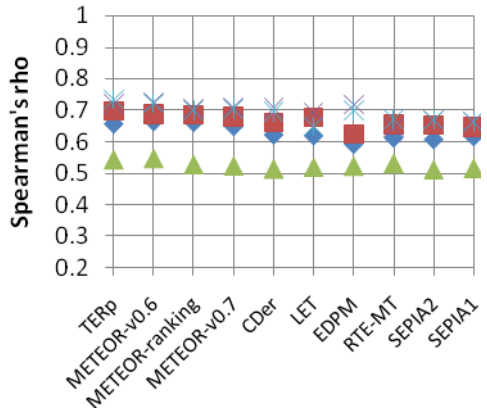
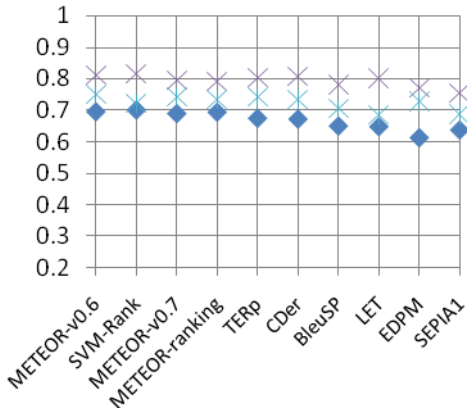
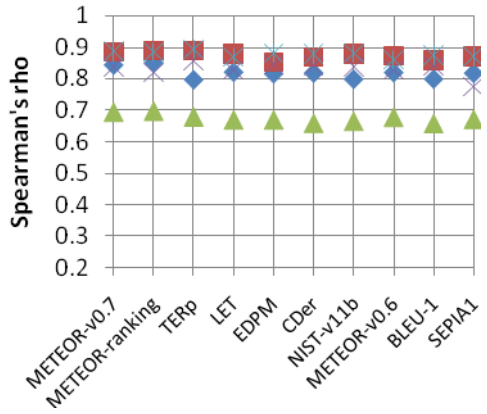
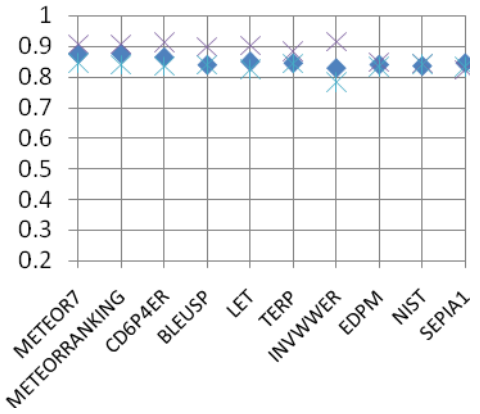
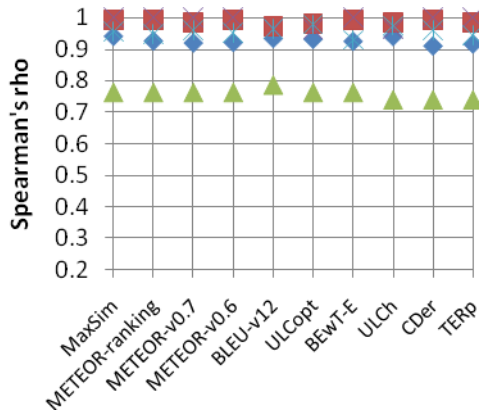
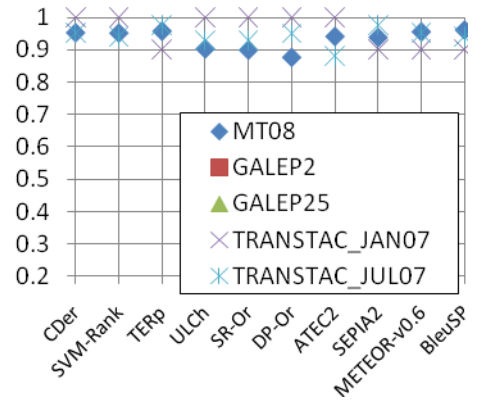
Level of Analysis	Single Reference Track	Multiple reference track
Segment level		
Document level		
System level		

Table 8: Spearman's rho for the 10 highest correlating metrics, for each evaluation track and level of analysis, conditioned on the evaluation data set

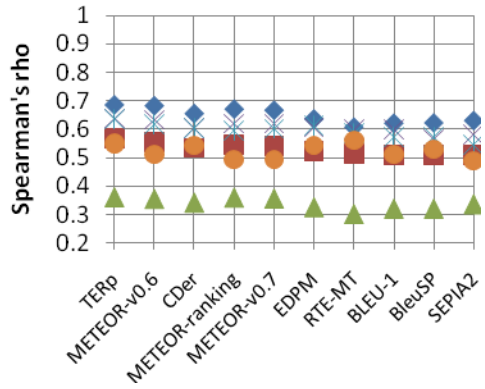
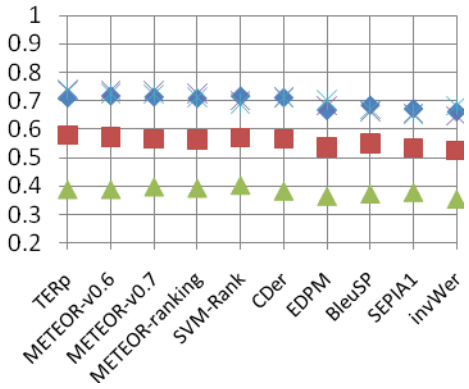
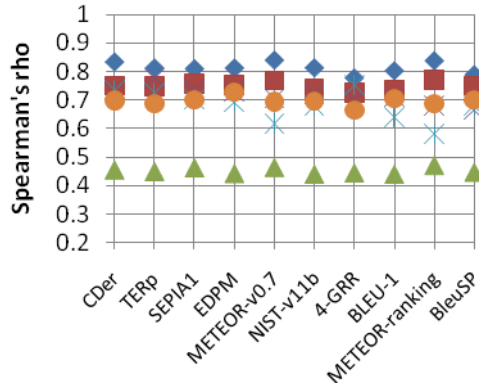
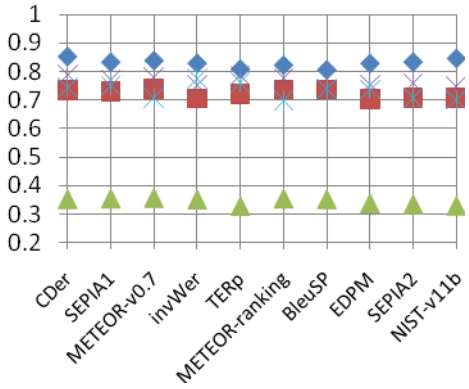
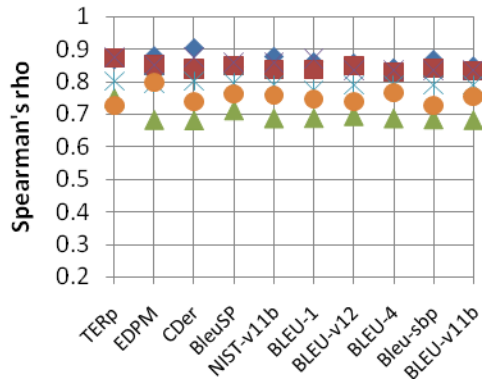
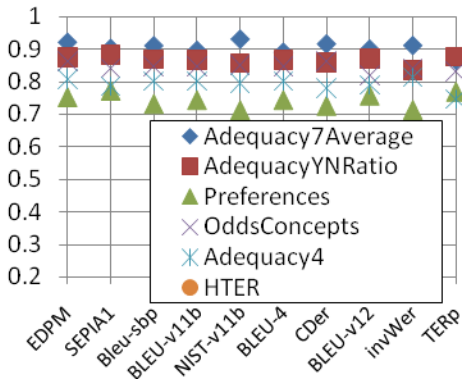
Level of Analysis	Single Reference Track	Multiple reference track
Segment level		
Document level		
System level		

Table 9: Spearman's rho for the 10 highest correlating metrics for each evaluation track and analysis level, conditioned on the human assessment type

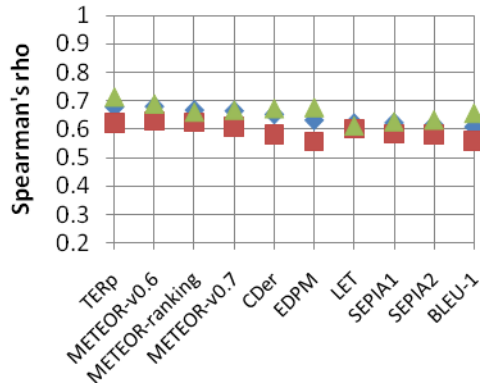
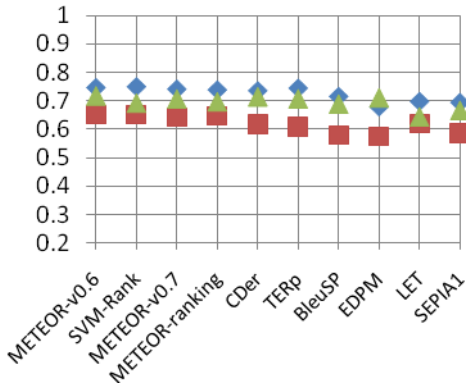
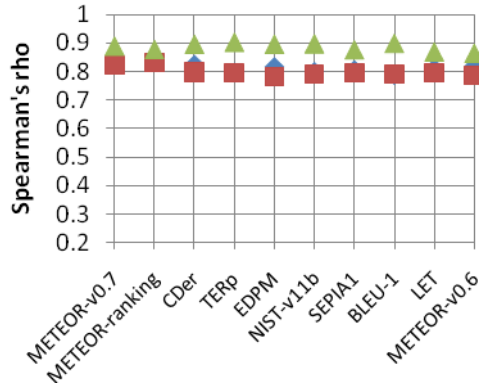
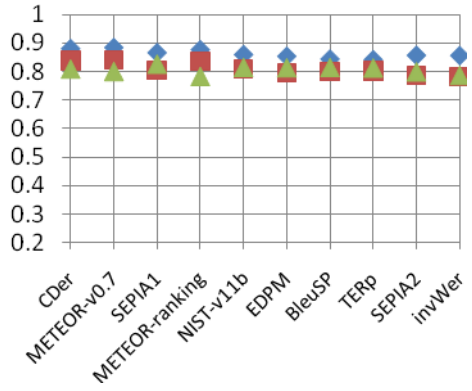
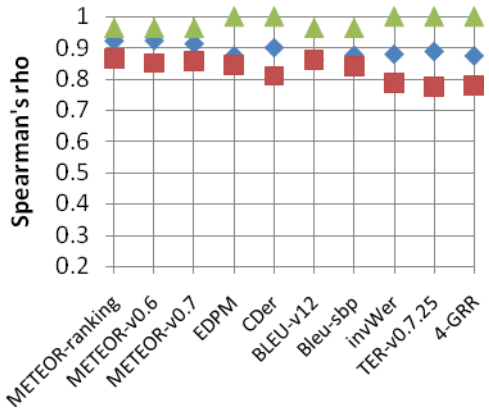
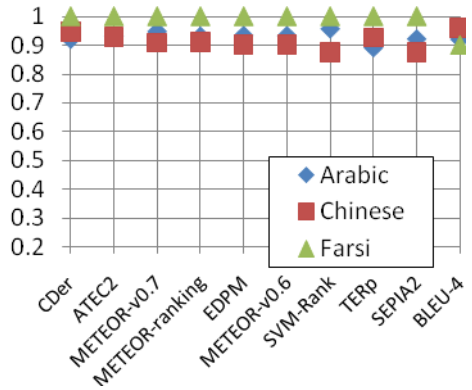
Level of Analysis	Single Reference Track	Multiple reference track
Segment level		
Document level		
System level		

Table 10: Spearman's rho for the 10 highest correlating metrics for each evaluation track and analysis level, conditioned on the source language

The different evaluation test sets represented translations from different technology types. Next, we examined how the metrics scores correlate to human assessments conditioned on three technology types; text-to-text from MT08 and GALE P2 data; speech-to-text from the GALE P2.5 data; and speech-to-speech from the TRANSTAC data.

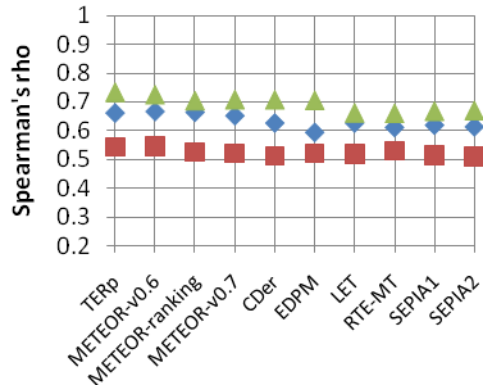
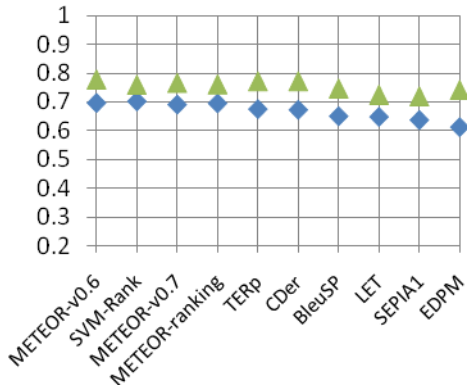
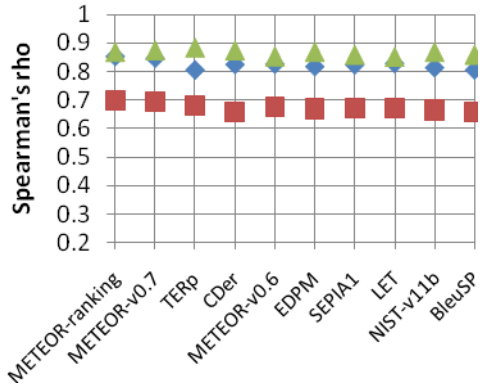
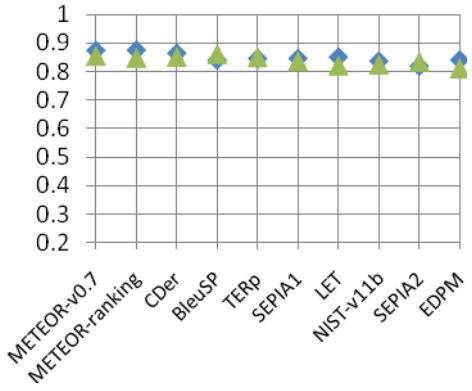
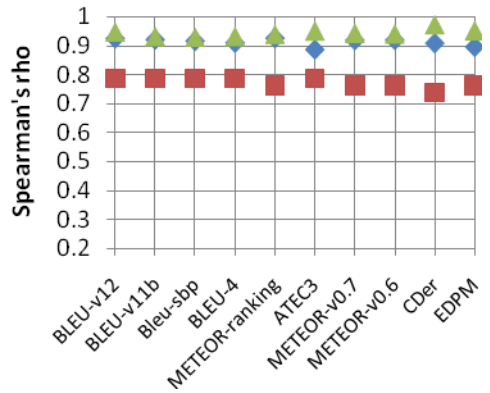
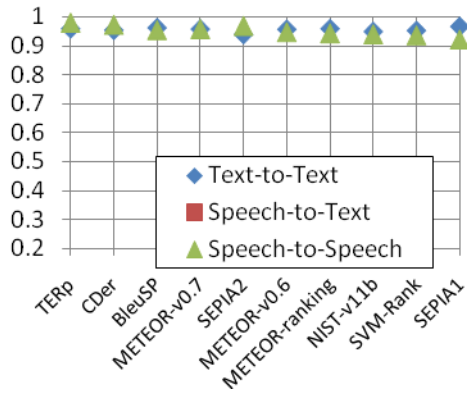
Level of Analysis	Single Reference Track	Multiple reference track
Segment level		
Document level		
System level		

Table 11: Spearman's rho for the 10 highest correlating metrics for each evaluation track and level of analysis, conditioned on technology type

We plot the Spearman's rho value for the 10 highest correlating metrics, for each evaluation track, and for each level of analysis (see Table 11). The metric correlations are higher for speech-to-speech, and lower for speech-to-text, while the performances on text-to-text are somewhat in-between. Differences between speech-to-speech and text-to-text attenuate at the system level. However, regardless of the level of analysis, the correlation values for speech-to-text remain well below that of speech-to-speech

and text-to-text. Multiple references help raise the correlation value, but this is much more noticeable at the segment level, and barely noticeable at the system level.

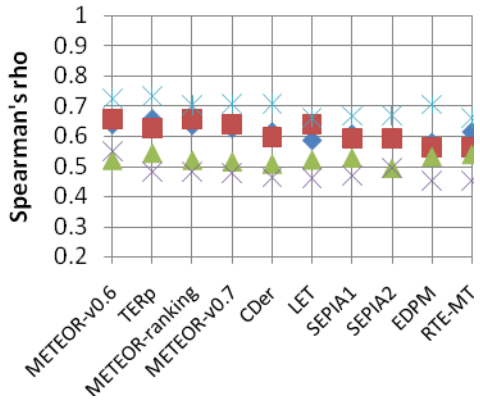
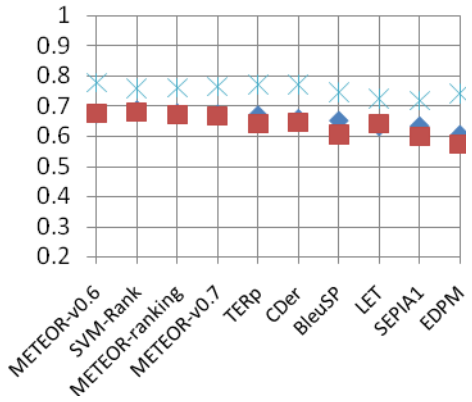
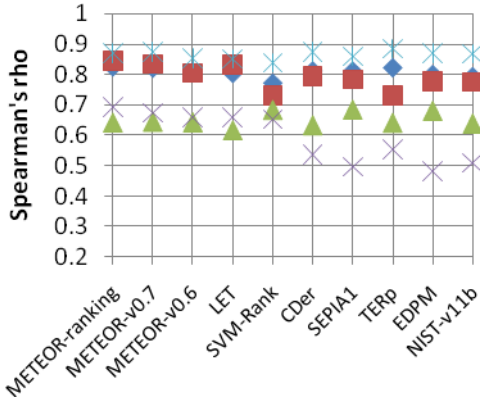
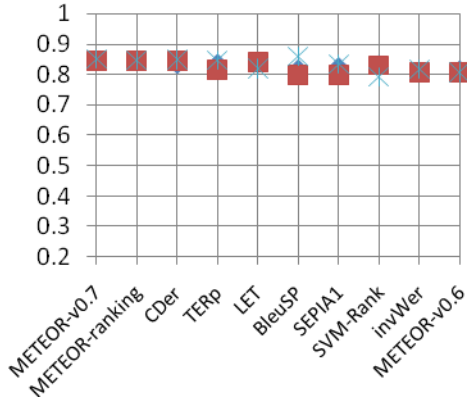
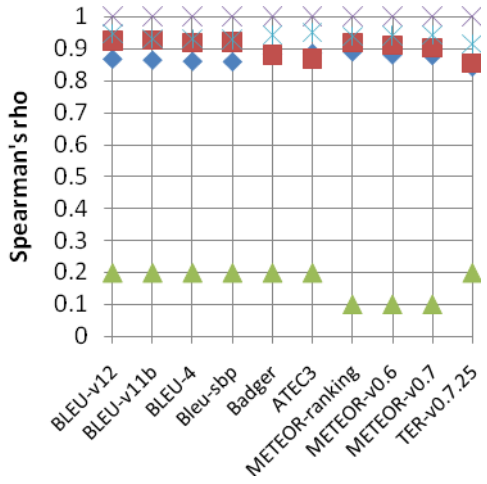
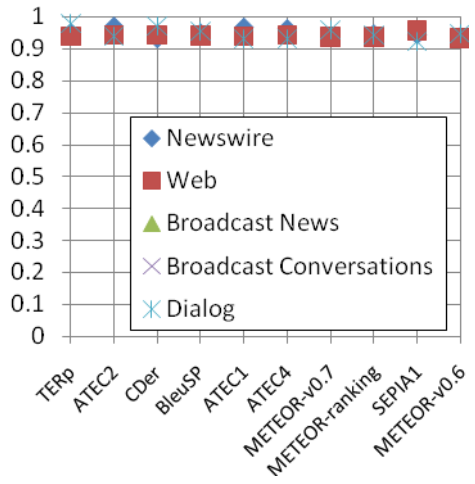
Level of Analysis	Single Reference Track	Multiple reference track
Segment level		
Document level		
System level		

Table 12: Spearman's rho for the 10 highest correlating metrics, for each evaluation track and level of analysis, conditioned on the data genre

Table 12 conditions the correlations on the 5 genres of data included in the MetricsMATR evaluation test set: newswire, web text, broadcast news transcripts, broadcast conversation transcripts, and dialogs.

Overall, metrics correlated higher with the structured styles of data, broadcast news and newswire, than they did with the web text, which represents a more freeform style of writing.

Broadcast news and broadcast conversation correlations are included for completeness, but is should be noted that the poor system level correlations are due to the sparseness of data points (only 5!); the GALE P2.5 data contains data from only very few GALE systems.

5.2. Significance Tests for performance differences

Three kinds of performance differences are of most interest here. First, we may want to compare metrics and ask which yields the closest to the same information as human judgments. Second, we may want to compare different versions of the same system, which will all typically be quite similar to each other. Third, we may want to compare different systems to each other, which will often be very different from each other (for example, comparing a rule-based MT system to a statistical MT system to a human translation).

Frank Wilcoxon proposed the Wilcoxon Signed-Rank Test (22) in 1945. The Wilcoxon Signed-Rank test is used to analyze matched-pair numeric data, looking at the difference between the two values in each matched pair. NIST has employed this significant test in various speech-to-text transcription, speaker recognition, and machine translation evaluations.

5.2.1. Comparison of similar systems

In a first step, similar system pairs are identified by performing the Wilcoxon Signed-Rank Test on the Adequacy7 judgments of two distinct systems. This is done for all possible system pairs. We create two lists of similar system pairs, i.e. pairs deemed by the test to not have a statistically significant difference in their Adequacy7 scores: one list computed using segment level Adequacy7 judgments, and one list using document level Adequacy7 judgments. The first (segment level) list contains 54 pairs of systems, and the second, 82 pairs.

The second step is to compute the same test using metric scores. For each pair of systems identified in the previous step, we perform the Wilcoxon test and count the number of times the test does not reject the null hypothesis, i.e. agrees with the human assessments in not finding a statistically significant difference between the two systems of a pair.

Results are reported in Table 13 for the test computed at the segment level and at the document level.

Segment level Scores		Document level Scores	
Metric Name	System pairs correctly identified	Metric Name	System pairs correctly identified
ULCSR	85.70%	**ATEC1**	81.70%
BADGER	75.00%	**ATEC2**	80.50%
ULCDP	75.00%	**ATEC4**	80.50%
MTEVALV12	67.90%	**BADGER**	80.50%
ULCDRP	67.90%	**BEWTE**	79.30%
BLEUSBP	66.10%	**ULCDRP**	79.30%
INVWWER	66.10%	**ULCSR**	79.30%
MTEVALV11B	66.10%	**MBLEU**	78.00%
SNR	66.10%	**ULCDP**	76.80%
ULCH	66.10%	**BLEU**	75.60%
ATEC3	64.30%	**ULCDR**	75.60%
ULC	64.30%	**MTEVALV12**	74.40%
BLEUSP	62.50%	**SVMRanking**	74.40%
MAXSIM	62.50%	**BLEUSP**	73.20%
TER	62.50%	**SNR**	73.20%

Table 13: Number of correctly identified similar systems pairs, for the 15 top-performing metrics, using the Wilcoxon test on segment level and document level scores

5.2.2. Comparison of different systems

We perform the same tests as done previously, this time identifying the system pairs that are considered *different*, by using Adequacy7 as a baseline. We identify 123 pairs for the segment level test, and 244 pairs for the document level test.

Then we compute the same test using metric scores, and award a point to a metric whenever the outcome is the same as for the baseline test: the null hypothesis must be rejected, and the Hodges-Lehmann statistic must point to the same system yielding higher data values than the other, at a 95% confidence interval. Table 14 (resp. Table 15) shows the segment level (resp. document level) results.

Metric Name	Correct system identified	Wrong system chosen
METEOR7	95.90%	0.00%
METEOR	95.10%	0.00%
METEORRANKING	95.10%	0.00%
ATEC4	93.50%	0.00%
CD6P4ER	92.70%	0.80%
LET	92.70%	0.00%
SEPIA1	92.70%	0.00%
TERP	92.70%	0.80%
ATEC1	91.90%	0.00%
ATEC2	91.90%	0.00%
BLEUSP	91.90%	0.80%
RTEMT	90.20%	1.60%
ATEC3	89.40%	0.00%
BADGERLITE	88.60%	0.80%
INVWWER	87.80%	0.80%

Table 14: Number of correctly identified different system pairs, for the 15 top-performing metrics, using the Wilcoxon test on segment level scores

Metric Name	Correct system identified	Wrong system chosen
METEOR7	88.90%	0.00%
CD6P4ER	87.70%	0.00%
METEORRANKING	86.90%	0.00%
TERP	86.50%	0.00%
BLEUSP	84.00%	0.00%
METEOR	84.00%	0.00%
RTEMT	84.00%	0.00%
EDPM	83.20%	0.40%
INVWWER	83.20%	0.40%
GRR	82.80%	0.00%
SEPIA1	82.40%	0.00%
TER	81.60%	0.80%
LET	81.10%	1.60%
RTE	81.10%	0.40%
BLEU	80.30%	0.00%

Table 15: Number of correctly identified different system pairs, for the 15 top-performing metrics, using the Wilcoxon test on document level scores

5.2.3. Approximate Randomization test

In 1989, Eric Noreen (23) described approximate randomization in detail, a technique used to analyze results of the 2001 MUC-3 conference (24). In 2005, Riezler and Maxwell (25) suggested its use for testing differences in sentence-level scores from two MT systems. For this technique, one compares the actual test statistic to a great many pseudo-statistics generated via a randomization process. The *p* value is based on what fraction of the pseudo-statistics unexpectedly exceeds the actual test statistic.

Tables 16 shows the results of the randomization test for analysis at the segment level (using 89 pairs) and at the document level (using 88 pairs).

Segment level scores		Document level scores	
Metric Name	Pairs correctly identified	Metric Name	Pairs correctly identified
ULCDRP	82.00%	**BADGER**	81.80%
ULCSR	79.80%	**ATEC1**	78.40%
ATEC2	77.50%	**ATEC3**	78.40%
MBLEU	77.50%	**ULCDP**	78.40%
ATEC1	76.40%	**ATEC2**	77.30%
ATEC4	76.40%	**ATEC4**	77.30%
BADGER	76.40%	**BEWTE**	77.30%
ULCDR	75.30%	**ULCSR**	77.30%
SEPIA1	74.20%	**ULCDR**	76.10%
ATEC3	73.00%	**BLEU**	75.00%
METEOR7	73.00%	**MBLEU**	75.00%
ULCDP	73.00%	**METEOR**	75.00%
TERP	71.90%	**MTEVAL12**	75.00%
METEORRANKING	69.70%	**SNR**	75.00%
NIST	69.70%	**MTEVALV11B**	73.90%

Table 16: Number of correctly identified similar systems pairs, for the 15 top-performing metrics, using the paired randomization test on segment level and document level scores

5.3.Other MetricsMATR 2008 Findings

The following sub-sections examine other interesting findings from MetricsMATR 2008.

5.3.1. Ability to differentiate between Machine and Human Translations

For this analysis, we included two human reference translations in the MetricsMATR pool of evaluation data. We looked at the metrics' ability to distinguish between machine translations (MT) and the human translation (HT). The available HT data was limited to the MT08 and TRANSTAC partitions, since these data sets contained four reference translations.

Using the scores produced by the metrics in the single reference track, we compared the corresponding segment scores or document scores of the MT and HT. Each time the metric scored the MT better than the HT system, an error was counted. The number of errors was then divided by the total number of comparisons, to obtain a percentage of errors. Table 17 reports the 15 metrics yielding the least amount of errors at the segment level and document level.

Segment level scores		Document level scores	
Metric Name	**Percentage of errors**	**Metric Name**	**Percentage of errors**
ULCH	9.1%	MAXSIM	1.9%
ULC	9.5%	ULCH	2.0%
MAXSIM	9.8%	ULC	2.1%
ULCDP	10.2%	ULCDP	2.3%
SNR	10.5%	SNR	2.3%
ULCSR	11.4%	ULCDR	3.3%
ULCDR	12.2%	ULCSR	3.7%
ULCDRP	15.9%	ULCDRP	4.1%
MTEVALV12	16.6%	TERP	6.1%
BLEUSBP	16.6%	METEORRANKING	6.4%
MTEVALV11B	16.6%	EDPM	6.5%
BEWTE	19.2%	METEOR7	6.9%
METEORRANKING	20.2%	LET	7.3%
TERP	20.3%	METEOR	7.4%
METEOR	20.7%	CD6P4ER	7.6%

Table 17: Percentage of times a MT was deemed better than a RT, for the 15 top-performing metrics, at the segment level and document level

5.3.2. Timing Information

The time to run a metric affects its usefulness for evaluation and also system development; in particular, training approaches that involve tuning to metrics, such as Och's (2003) Minimum Error Rate Training (MERT) (26) as well as Chiang's (2009) approach (27) benefit from fast-running metrics. For each metric, NIST recorded the time (wall clock) required to score the entire MetricsMATR evaluation set. For the single reference track, this required scoring approximately 25K segments for 89 systems, and for the multiple reference track, it required scoring approximately 16K segments for 69 systems. The majority of the metrics were run on our Linux 64-bit machine. We do not distinguish between metrics that take advantage of multi-core CPUs and those that do not.

Table 18 shows the total time spent per metric to process the MetricsMATR evaluation set, for the single reference track. Table 19 shows the same timing statistics for the multiple reference track.

Time range	Metrics
Less than 1 minute	BLEU-1, BLEU-4, ATEC1, BLEU-SBP
Between 1 minute and 10 minutes	BleuSP, BLEU-v11b, NIST-v11b, CDer, BLEU-v12, SNR[12], TER-v0.7.25, 4-GRR, BadgerLite, ATEC4, ATEC2, ATEC3
Between 10 minutes and 40 minutes	METEOR-v0.6, LET[12], METEOR-v0.7, METEOR-ranking, mBLEU, mTER, MaxSim, DP-Or
Between 1 hour and 6 hours	SVM-Rank[12], DP-Orp, DR-Or, SEPIA1, SEPIA2, ULCopt, TERp
Between 13 hours and 14 hours	invWer, Badger, EDPM
Between 1 day and 2 days	BEwT-E, SR-Or, ULCh
5 days, 14 hours	(RTE + RTE-MT)[13]

Table 18: Time spent per metric to process the primary evaluation set, for the single reference track

[12] Ran on a Windows XP machine; everything else ran on a Linux box.
[13] Both scores computed at the same time.

Time range	Metrics
Less than 1 minute	BLEU-1, ATEC1, BLEU-4
Between 1 minute and 10 minutes	BLEU-SBP, BleuSP, CDer, BLEU-v11b, NIST-v11b, BLEU-v12, SNR[12]
Between 10 minutes and 40 minutes	ATEC4, ATEC3, ATEC2, TER-v0.7.25, BadgerLite, mBLEU, 4-GRR, LET[12], METEOR-ranking, METEOR-v0.7, METEOR-v0.6, mTER, MaxSim
Between 2 hours and 5 hours	DP-Or, SVM-Rank[12], DP-Orp, DR-Or, SEPIA1, SEPIA2
Between 6 hours and 16 hours	ULCopt, TERp, Badger
Between 1 day and 3 days	EDPM, invWer, SR-Or, BEwT-E
4 days, 8 hours	ULCh

Table 19: Time spent per metric to process the primary evaluation set, for the multiple reference track

6. MetricsMATR 2008 Summary, Lessons Learned, Future Directions

NIST's first edition of MetricsMATR was a success. The framework for the evaluation of automated MT metrics is in place, and we are quickly building critical mass. There were a great number of metrics submitted for evaluation, some combining techniques from previously existing metrics. Figure 1 illustrates the landscape of automated metrics at NIST's disposal for the evaluation of MT technology, before and after MetricsMATR 2008.

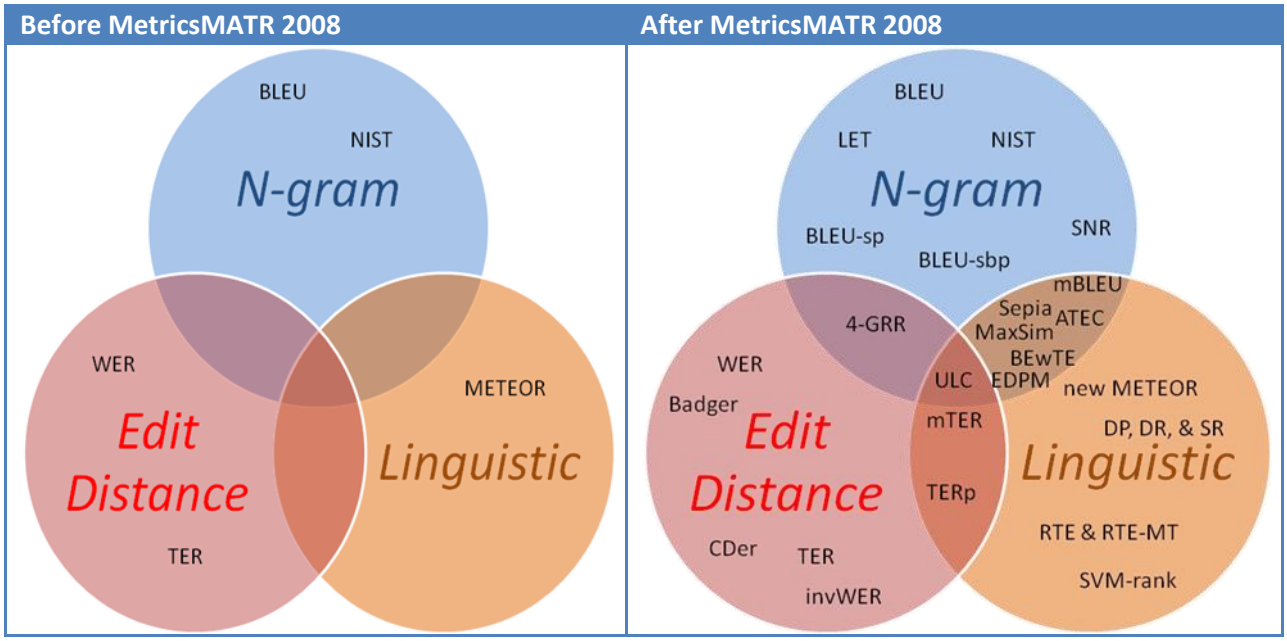


Figure 2: Classes of automated metrics available to NIST for MT technology evaluations, before and after MetricsMATR 2008

The evaluation of automated metrics will continue, seeking the development of innovative metrics that provide insights into the quality of a translation. The current set of metrics provides an excellent test bed for analysis. We demonstrated that metrics are not robust across conditions, but determining each metric's strengths and weaknesses will enable improvements in future instantiations.

The next MetricsMATR evaluation is being planned for 2010. The focus of the preparations for the next challenge will likely be on the human assessments. It is imperative that the human assessments themselves provide detailed levels of quality, since we are asking the automated metrics to do the same.

Disclaimer

These results are not to be construed, or represented as endorsements of any participant's system, methods, or commercial product, or as official findings on the part of NIST or the U.S. Government.

Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

There is ongoing discussion within the MT research community regarding the most informative metrics for machine translation. The design and implementation of these metrics are themselves very much part of the research. At the present time, there is no single metric that has been deemed to be completely indicative of all aspects of system performance.

The data and protocols employed in this evaluation were chosen to support MT metric development and should not be construed as indicating how well these metrics would perform in applications.

Bibliography

1. **Papineni, Kishore, et al.** *BLEU: a Method for Automatic Evaluation of Machine Translation.* Yorktown Heights, NY : IBM Research Division, September 17, 2001. Technical Report. RC22176 (W0190–022).

2. **Coughlin, Deborah.** *Correlating Automated and Human Assessments of Machine Translation Quality.* New Orleans, LA : Association for Machine Translation in the Americas, 2003. Proceedings of MT Summit IX.

3. **Doddington, George.** *Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrence Statistics.* San Francisco : Morgan Kaufmann, 2002, Proceedings of the Second International Conference on Human Language Technology Research (San Diego, CA).

4. **Babych, Bogdan and Hartley, Anthony.** *Extending the BLEU MT Evaluation Method with Frequency Weightings.* Barcelona, Spain : Association for Computational Linguistics, 2004. Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04).

5. **Callison-Burch, Chris, Osborne, Miles and Koehn, Philipp.** *Re-evaluating the Role of BLEU in Machine Translation Research.* Trento, Italy : Association for Computational Linguistics, 2006. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics.

6. **Chiang, David, et al.** *Decomposability of Translation Metrics for Improved Evaluation and Efficient Algorithms.* Honolulu, Hawaii : Association for Computational Linguistics, 2008. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.

7. **Condon, Sherri, et al.** *Normalization for Automated Metrics: English and Arabic Speech Translation.* Ottawa, ON, Canada : Association for Machine Translation in the Americas, 2009. Proceedings of MT Summit XII.

8. **Cohen, Jacob.** *A Coefficient of Agreement for Nominal Scales.* 1960, Educational and Psychological Measurement, pp. 37–46.

9. **Fleiss, Joseph L., Cohen, Jacob and Everitt, B.S.** *Large Sample Standard Errors of Kappa and Weighted Kappa.* 1969, Psychological Bulletin, Vol. 72.

10. **Fleiss, Joseph L.** *Measuring Nominal Scale Agreement among Many Raters.* 1971, Psychological Bulletin, Vol. 76.

11. **Callison-Burch, Chris, et al.** *(Meta-) Evaluation of Machine Translation.* Prague, Czech Republic : Association for Computational Linguistics, 2007. Proceedings of the Second Workshop on Statistical Machine Translation.

12. **Callison-Burch, Chris, et al.** *Further Meta-Evaluation of Machine Translation.* Columbus, OH : Association for Computational Linguistics, 2008. Proceedings of the Third Workshop on Statistical Machine Translation (WMT08).

13. **Przybocki, Mark, Peterson, Kay and Bronsart, Sébastien.** NIST Metrics for Machine Translation Challenge (MetricsMATR). *NIST Multimodal Information Group.* [Online] April 4, 2008. http://www.nist.gov/speech/tests/metricsmatr/2008/doc/mm08_evalplan_v1.1.pdf.

14. **Paul, Michael.***Overview of the IWSLT 2006 Evaluation Campaign.* Kyoto, Japan : s.n., 2006. Proceedings of the International Workshop on Spoken Language Translation.

15. **Jones, Douglas, et al.** *ILR-Based MT Comprehension Test with Multi-Level Questions.* Rochester, NY : Association for Computational Linguistics, 2007. Proceedings of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007).

16. **Snover, Matthew, et al.** *A Study of Translation Edit Rate with Targeted Human Annotation.* Cambridge, MA : s.n., 2006. Proceedings of Association for Machine Translation in the Americas.

17. **Sanders, Gregory A., et al.** *Odds of Successful Transfer of Low-level Concepts: A Key Metric for Bidirectional Speech-to-speech Machine Translation in DARPA's TRANSTAC Program.* Marrakech, Morocco : European Language Resources Association (ELRA), 2008. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08).

18. **Lavie, Alon and Agarwal, Abhaya.** *METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments.* Prague : s.n., 2007. Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association of Computational Linguistics (ACL-2007).

19. **Fellbaum, Christiane.** *Wordnet: An Electronic Lexical Database.* s.l. : Bradford Books, 1998.

20. **Porter, Martin.** Snowball. *Tartarus.org.* [Online] [Cited: July 24, 2009.] http://snowball.tartarus.org.

21. **Neter, John, et al.** *Applied Linear Statistical Models.* 1996 : McGraw-Hill/Irwin.

22. **Wilcoxon, Frank.** *Individual Comparisons by Ranking Methods.* 1945, Biometrics, Vol. 1.

23. **Noreen, Eric W.** *Computer Intensive Methods for Testing Hypotheses. An Introduction.* New York : Wiley.

24. **Chinchor, Nancy, Hirschman, Lynette and Lewis, David D.** *Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3).* 3, 1993, Computational Linguistics, Vol. 19.

25. **Riezler, John and Maxwell, John T., III.** *On Some Pitfalls in Automatic Evaluation and Significance Testing for MT.* Ann Arbor, MI : Association for Computational Linguistics, 2005. ACL-05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.

26. **Och, Franz Josef.** *Minimum Error Rate Training in Statistical Machine Translation.* Sapporo, Japan : Association for Computational Linguistics, 2003.

27. **Chiang, David, Knight, Kevin and Wang, Wei.** *11,001 New Features for Statistical Machine Translation.* Boulder, CO : Association for Computational Linguistics, 2009. Proceedings of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2009).