

A Statistical Analysis of Automated MT Evaluation Metrics for Assessments in Task-Based MT Evaluation

Calandra R. Tate

US Army Research Laboratory
2800 Powder Mill Rd.
Adelphi, MD 20783, USA
ctate@arl.army.mil

Abstract

This paper applies nonparametric statistical techniques to Machine Translation (MT) Evaluation using data from a large scale task-based study. In particular, the relationship between human task performance on an information extraction task with translated documents and well-known automated translation evaluation metric scores for those documents is studied. Findings from a correlation analysis of this connection are presented and contrasted with current strategies for evaluating translations. An extended analysis that involves a novel idea for assessing partial rank correlation within the presence of grouping factors is also discussed. This work exposes the limitations of descriptive statistics generally used in this area, mainly correlation analysis, when using automated metrics for assessments in task handling purposes.

1 Introduction

Automated machine translation evaluation metrics (autometrics) have been justified as an evaluation tool based on how well they correlate across an entire document testbed with human judgments of translation quality on the same data set. With system ranking as the primary objective for such translation evaluations, an important part has remained left out—the practical assessment of documents from the user’s perspective. As a result, there has been the assumption among MT developers that MT engines are “good enough” to support people performing certain applications in the real world (Church and Hovy, 1993).

One of the main drawbacks of this assumption is that there really is no solid understanding of what a specific automatic score means. How far off is a metric score of .35 from a score of .50 when you are dealing with translated outputs? Is the .15 score difference really that significant? Likewise, is a translation quality score of .20 twice as bad as one with a score of .40? There has been no validation of this linearity of scores, and similarly, there have been no empirical results indicating how useful a translation is based on these scores. Because of questions like these, although autometrics have become standard and have offered much insight in the evaluation community, their limits, stability, and interpretation still remain questionable. A natural question to ask is, *Will these metrics correlate with other evaluation methods?*

A limited number of studies that approach document quality evaluation, based not on an intrinsic question of what is actually in the document, but more extrinsically on what one can do with the documents, have been performed.¹ Although this method has not been explored as much as other approaches, acquiring a connection between intrinsic and extrinsic metrics for MT evaluation would give users, as well as researchers, a threshold for task performance relative to machine performance. Jones et. al (2005) have begun to discuss this issue by addressing how remarkable gains in autometric scores in technology-center evaluations over the past few years are reflected in measures of effectiveness such as human readability of machine generated texts.

¹See (Tomita, 1992); (Taylor and White, 1998); (Fuji et. al, 2001)and (Tate et. al., 2003) for a few examples.

Given the lack of connection in the field of MT Evaluation between task utility and the interpretation of automated evaluation metrics, there is a need to bring innovative and interdisciplinary analytical techniques to this problem. In this work, we explore the relationship between quality (as judged by autometrics) and utility (as judged by correct performance on an information extraction task) in an attempt to leverage existing automated MT Evaluation metrics to assess task performance on machine translated documents. In the next section, we begin with a description of the task-based data and automated metrics used in our analysis.

2 Data Description

2.1 Task-Based MT Evaluation Metrics

We use responses collected from the extraction experiment conducted by (Voss and Tate, 2006) as our task-based metrics in this study. That experiment was designed to assess the translation output of three different Arabic MT systems and the performance of multi-level translation analysts on a Who, Where, When (referred to as Wh's) information extraction task using translated documents produced by these machines.

The experiment consisted of 59 subjects who each analyzed 18 translated documents. For each translated document that they viewed, subjects identified all occurrences of words or phrases that met the pre-defined criteria for the particular Wh-type they sought in the translated document.

The authors collected various types of metrics from the subject responses which are described in more detail in (Tate and Voss, 2006). In this paper, we utilize the collection of *correct responses*, those subject responses that fully matched the reference truth (RT) answer item. Moreover, we are interested in the proportion of correct items out of the RT items extracted by subjects from the translated documents called the *hit rate*. Document and collection level task response rates are computed for the roughly 354 documents across each of the three MT systems.

2.2 Automated MT Evaluation Metrics

We chose four pre-existing autometrics—BLEU, GTM, METEOR, and TER—that have been fairly standard for MT Evaluation to compare to task re-

sponses in our study. We briefly review the individual characteristics of the four automated metrics.

IBM researchers Papineni et al. (2002) proposed the Bilingual Language Evaluation Understudy (BLEU) metric which is the most widely used of the four autometrics. This measure of “precision” scores candidate translations against a user-selected number of stored reference translations by counting the number of consecutive word groups of size n , or n -grams, that overlap between the candidate and reference. The final BLEU score is a combination of these matches (n -gram precision) using the geometric mean across different values of n and a brevity penalty for shorter machine translations.

Turian et al. (2003) established the General Text Matcher (GTM), that builds on an earlier idea by (Melamed et. al., 2003). Candidate translations are scored with respect to a reference translation by computing similarity through the number of matching words. Unlike BLEU, this metric abandons the “precision only” idea altogether by using the *F-score*, a scoring function that takes the harmonic mean of precision *and* recall.

Researchers at Carnegie Mellon University introduced the Metric for Evaluation of Translation with Explicit ORDERing (METEOR) for MT evaluation in (Banerjee, 2005; Lavie and Agarwal, 2007). METEOR heavily relies on an algorithm for finding an optimal word-to-word matching between a candidate system translation and a human-produced reference translation for the same input sentence. Recall is a major contributor to the score.

Translation Error Rate (TER) (Snover et. al., 2005) measures the minimum number of edits required to change a candidate output into one of the available human references. The score is normalized by the average length of the references and only uses edits recorded from the closest reference. TER uses an edit distance measure similar to word error rate to find the translation/reference pair that has the minimal number of edits and assigns this score as the translation quality metric for the particular translation.

2.3 Data Summary

Table 1 shows a portion of the data set for ten of the experimental cases. Each row represents one subject's analysis of one of the translated documents.

Subj	MT	WH	Rep	RTMTot	Hit Rate	BLEU	METEOR	TER	GTM
S43	3	WHERE	4	7	.5714	.040	.421	.214	.432
S9	3	WHO	6	10	.4000	.126	.582	.333	.584
S57	2	WHEN	4	12	.0000	.099	.383	.297	.540
S59	1	WHERE	6	8	.3750	.084	.445	.302	.601
S1	2	WHO	3	9	.2222	.211	.503	.423	.611
S46	1	WHEN	5	5	.4000	.043	.198	.220	.401
S55	2	WHERE	6	8	.5000	.155	.494	.355	.669
S34	3	WHO	2	7	.2857	.046	.223	.245	.373
S52	2	WHERE	6	8	.8750	.155	.494	.355	.669
S14	2	WHEN	3	8	.3750	.283	.567	.376	.692

Table 1: Random sample of 10 cases from data collected. Column headings are described in the text.

Variable	Min	Max	Mean	Median	Std. Dev		MT-1	MT-2	MT-3
Hit Rate	0	1	.436	.429	.226	# of Correct			
BLEU	.016	.283	.106	.080	.075	Extractions (Hits)	1181	1506	1370
GTM	.264	.709	.528	.540	.097	Total # of Possible			
METEOR	.198	.621	.425	.413	.100	Correct Responses	3091	3066	3086
TER	.105	.437	.272	.269	.091	Hit Rate	.382	.491	.444

Table 2: Summary statistics for study variables.

The first column denotes the subject who viewed the particular document. The second through fourth columns represent the document identifier detailing the machine system which produced the translation, the (WH) type of information that was being extracted, and the replicate number (1-6) of the category, respectively. The fifth column is the total number of possible items to extract from the document (RTMTot). The sixth column is the proportion of correct items the subject extracted (Hits) out of the total possible items. The remaining columns correspond to the various autometric scores computed for the document.

Summary statistics of each variable can be found in Table 2. METEOR and GTM scores are slightly more dispersed than the other two metrics. GTM and TER scores classify more documents as above average translations, with 51% (550) and 50% (531) of the responses, respectively, having higher scores than the mean score. The BLEU scores of 35% (373) responses were higher than the mean BLEU score while this held true for 38% of METEOR scores.

The collection level metric scores for each MT system are presented in Table 3 and Table 4. MT-1 yielded significantly lower rates of correct answers

Table 3: Hit rates by MT engine, aggregated over all WH-types, subjects and documents.

Automated Metric	MT-1	MT-2	MT-3
BLEU	.088	.187	.055
GTM	.529	.617	.453
METEOR	.385	.524	.397
TER	.221	.370	.233

Table 4: Automated metric scores by MT engine, aggregated over all WH-types, subjects and documents.

from subjects. This pattern does not hold true for the autometric scores calculated by MT system. MT-1 has the lowest METEOR and TER scores while MT-3 has the lowest BLEU and GTM scores. Both subject performance and autometrics indicate MT-2 is the best translating engine in terms of utility and translation quality.

3 Correlation Analysis

3.1 Spearman Rank Correlation

The Spearman rank correlation(ρ) is a distribution-free rank statistic that tests the direction and strength of the relationship between two variables (Lehmann, 1998). Both sets of data are ranked from the highest to the lowest with the smallest

observation having rank 1 and the largest having rank n . Ranks are averaged in the case of ties. The statistic ρ is defined using the formula:

$$\frac{\sum_{i=1}^n \left(R(X_i, \underline{X}) - \frac{n+1}{2} \right) \left(R(Y_i, \underline{Y}) - \frac{n+1}{2} \right)}{\sqrt{\sum_{i=1}^n \left(R(X_i, \underline{X}) - \frac{n+1}{2} \right)^2 \sum_{i=1}^n \left(R(Y_i, \underline{Y}) - \frac{n+1}{2} \right)^2}} \quad (1)$$

where $R(A_i, \underline{A})$ represents the rank of A_i in the subset \underline{A} and is equal to the number of elements of \underline{A} less than or equal to A_i .

Unlike Pearson correlation, this method based on ranking the two variables, makes fewer assumptions about the distribution of the values and measures monotone rather than only linear covariation. For this reason, we chose to use Spearman rank correlation for the basis of our results.

3.2 Correlation in Aggregated Evaluation Datasets

In general, autometrics are computed on a document collection translated by a system, and the collection is then given a system score for each autometric. Separately, humans are solicited to make quality judgments on the documents according to some pre-assigned numeric scale. Note that human judgments are generally made at the individual segment level and then averaged across document and the collection to produce a ‘system specific’ human judgment score, as well.

Following this method, for any comparison, the number of total possible data points in a test set is $S \times (M + 1)$ where S is the number of systems and $M + 1$ is the number of metrics plus the human judgment of the system. For example, if there are 3 systems under consideration ($S1$, $S2$, and $S3$) and 2 different autometrics ($M1$ and $M2$) to compare to human judgment scores, there are only 9 total data points. Moreover, the pairwise autometric versus judgment correlation is computed from only 3 data points and in any case, using this aggregated approach, one would only have as many data points to correlate as systems under study.

Consider the system level correlation in Table 5 between the autometrics studied in this work and task responses.² For most metrics, our results show

²For the sake of comparison with previous results, Pearson correlation is used in this section for correlating aggregate

BLEU	GTM	METEOR	TER
.6634	.4676	.8654	.8626

Table 5: Autometrics correlated with hit rate for aggregate scores by MT using Pearson correlation

high correlations in the ‘aggregate’ sense between autometrics and task performance similar to those observed between autometrics and human judgments (Papineni et. al., 2002; Turian et. al., 2003; Lavie and Agarwal, 2007; Snover et. al., 2005). At this level of analysis, the METEOR and TER autometrics correlate highly with human performance on this task. However, as is true of previous work, the calculated ‘correlation’ simply represents the normalized inner product of only 3-dimensional vectors (3 metric values, one for each MT system) for the whole collection.

While aggregating scores across collections has given system developers insight into the development of their systems on average for particular collections of documents, users have not been able to make the connections between individual document scores and autometrics. The conclusions generated from aggregate-level correlations are not useful for further interpretation because they are a very coarse summary of group differences. However, more definitive conclusions about the relationship are possible when more system-level data points are available.

Some attempts have been made to utilize autometrics at other levels (Melamed et. al., 2003; Banerjee, 2005; Snover et. al., 2005). As one might expect, lower levels of aggregation have not achieved the high correlations that were observed previously in the aggregated case with autometrics, but recently some metrics have done well in comparison with others. For instance, METEOR and CDER were purposely designed to improve correlation at the *sentence* level, and the authors of both have shown results of higher correlations at the sentence level with human judgments than other metrics (Lavie and Agarwal, 2007; Leusch et. al., 2006).

Since our goal is to find a relationship between autometrics and subject task performance, and eventually to calibrate **document scores** with some degree scores.

BLEU	GTM	METEOR	TER
.211	.193	.242	.231

Table 6: Autometric Spearman correlation with hit rate on non-aggregate individual document scores

of utility for a specific task, it is relevant that the data is analyzed at a level more useful for task-based comparisons. Next, the differences in correlation results between task responses and autometrics at the individual document level of aggregation are shown.

3.3 Correlation at Unit Level for Task Performance Evaluation

Reeder and White (2003) mention that for many reasons, the evaluation issue is not solved since finer-grained metrics for smaller units of data (i.e., sentences, documents, etc) are needed. This is true especially in this work because eventually we want to use the metrics to **predict** task performance at the document level. To test the question *what happens to the relationship between autometrics and response rates when scores are compared at the document level?*, the entire set of non-aggregated individual document scores is compared. It is found that, even though the system rankings are the same, the degree of correlation between each autometric and task performance drastically changes [see Table 6] from the aggregate-level results. This suggests that useful relationships between autometrics and task performance at document level may not exist. However, scatterplot smoothing in the next section indicates quite the contrary.

The remainder of this paper shows that although a weak correlation exists at this stage of granularity, that does not necessarily indicate that there is no relationship between the two variables. One metric may prove to be a better *predictor* of task performance although its correlation may be weak. We proceed by using data smoothing techniques as well as the patterns for correlation, scatter, and linearity in the relationship of variables cross-classified into finer groups to identify which of the given set of metrics may be better in predicting our extraction task responses.

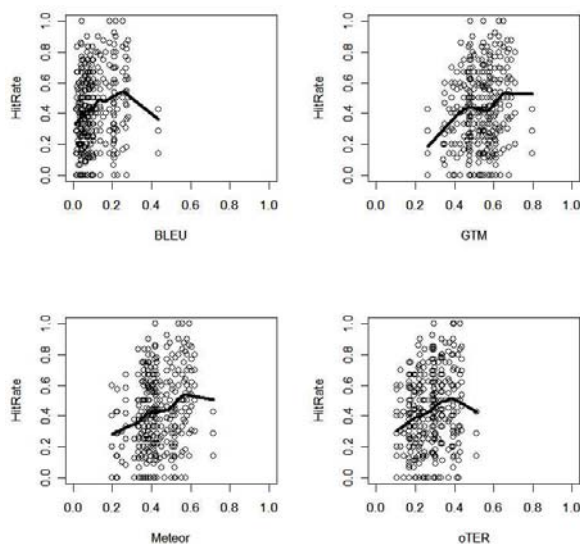


Figure 1: Scatterplot of the relationship between autometric scores and hit rate with smoothed lines denoting the lowest scatterplot smoother

4 Visualizing the Relationship Between Task-Based Metrics and Autometrics

There are several nonparametric regression techniques for smoothing data including: local averaging, kernel estimation, and smoothing splines. We focus on a widely used method in statistics for smoothing scatterplots of noisy data, originally introduced by (Cleveland, 1979), called *locally weighted scatterplot smoothing* (lowess). This method builds on classical approaches, such as linear and nonlinear least squares regression, as well as kernel estimation, providing weighted combinations of simple models fitted to localized subsets of the data.

Scatterplots of each metric plotted against the proportion of correct answers (hit rate) are displayed in Figure 1. Initially, the noisy raw data indicate that there is no clear picture of a possible association between metrics and task responses. The lowess smoothing technique described in the previous paragraph enhances the interpretation of the plot. The bold line shows that in each case, there is a generally increasing pattern of scores with an increase in hit rate.

It is interesting to point out in Figure 1 that there

BLEU	GTM	METEOR	TER
.249	.230	.281	.270

Table 7: Autometric Spearman correlation with Hit rate on individual document scores with outlier document removed.

are three outlying points in the data across all autometrics. These three points represent the same WHEN document from MT-2 that was viewed by a total of 20 subjects. Each of the autometrics achieve extremely higher scores for this document versus other documents in the collection while its hit rates are slightly lower than the average hit rates. We inspected this particular document suspecting that perhaps its characteristics would provide more insight. However, upon further analysis, we found no glaring evidence for this peculiarity and considered it an anomaly since the scores are so extreme compared to other documents. Entries of it were omitted from further study to prevent results from being adversely affected by this phenomenon.

Table 7 shows that the correlation between metrics and task performance from Table 6 slightly increases once this outlier is removed. This increase is more apparent for the correlations involving BLEU and GTM that were much lower than the others. Now all metrics appear to be on equal footing in relation to hit rate.

While scatterplot smoothers are a good tool for lots of data, they are not as good for smaller sets cross-classified into finer categories. For instance, there is less scatter in the relationship cross-classified by MT system 2, but Figure 2 shows that the lowess curves are more nonlinear and are non-monotonic. When the data is observed by WH-type only and further cross-classified into the 9 MT \times WH groupings, similar patterns can be found. Overall, lowess lines for the plots with BLEU and GTM metrics versus hit rate are nonlinear but for METEOR and TER, there is a mostly linear pattern. The latter two metrics have more readily visible increasing tendencies with hit rate than the former two.

The noisy and mixed patterns of the scatterplots when the data are cross-classified by MT-2 indicate that the relationship seen in Figure 1 may be due to the MT variable effect on the autometric scores. This finding is more confirmatory than surprising

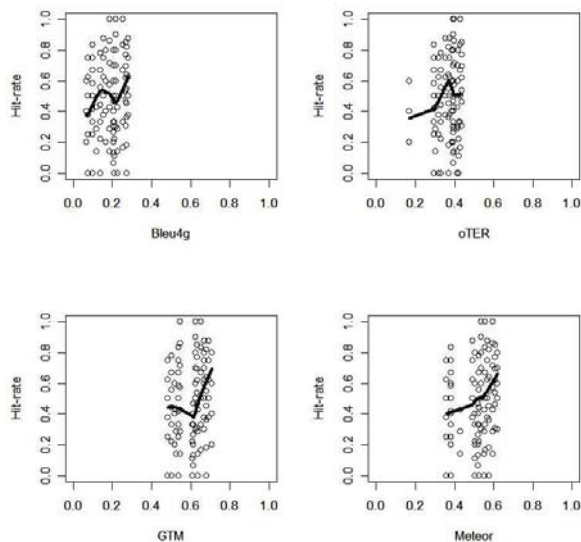


Figure 2: Scatterplot of autometric scores versus Hit rate with lowess lines for MT system 2

because previous evaluation work has shown that autometrics are useful in distinguishing between MT systems of varying quality. It was shown in the data summary results in Section 2.3 that MT systems can also be distinguished based on utility via the task response rates. Yet it is not evident whether autometrics contribute anything beyond being able to distinguish between systems when it comes to task response rates. Further analysis will help refine these distinctions.

5 Further Correlation Analysis

It has been established in this work that there is a positive and generally monotonic relationship between autometric and task performance variables in our data. However, the evidence of a strong relationship in the presence of other effects—such as MT and WH-type—is less apparent. This leads to further correlation analysis, in which we extend beyond the study of the strict bivariate relationships by using other categorical variables in the cross-classified data set to determine the extent of residual correlation between the two variables once the third variable is held constant.

In this section, we want to know: *to what extent does the autometric score still account for the cor-*

rect task response rate after adjusting for the MT effect and is this different for different MT systems? We also go further to explore: are there any groups other than MT that show interesting document to document variation that will help quantify the response rate? Studying the within-group and partial relationship answers such questions better than does population-wide correlation. Permutation tests, discussed in the next section, are used to determine whether relationships are significant.

5.1 Permutation Testing for Significance

Permutation tests provide a robust nonparametric alternative to using traditional, model dependent significance testing methods. The main idea of permutational testing is to estimate the empirical distribution of statistic values over the ensemble of randomly permuted datasets. Sampling of the permuted data provides a null hypothesis benchmark and “exact” significance level. Permutation tests of significance are conducted in this paper according to the following procedure (Good, 1999):

1. Compute the correlation of the original observations.
2. Resample the autometric scores, based only on permutations that preserve groups, and recompute the correlation for these permuted values.
3. Calculate the exact significance level (p-value) of the test from the formula

$$\frac{(\#\text{recomputed statistics} \geq \text{original statistic})}{n}$$

where $n = 5000$ is the number of permutations.

5.2 Within-Group Correlation

The results of the within MT-group correlation between hit rate and autometrics can be found in Table 8. Documents translated by all MT systems showed monotonic and significantly positive associations between hit rate and evaluation metric scores across all autometrics. Thus, there are real within MT-group relationships in our data. Response rates for extraction within WH-group also showed a monotonic and statistically significant positive association with evaluation metric scores across all metrics.

MT	BLEU	GTM	METEOR	TER
1	.140	.147	.109	.235
2	.134	.303	.298	.111
3	.298	.323	.311	.182

Table 8: Autometric correlation with hit rate on individual document scores cross-classified by MT system. All scores are shown to be permutationally significant with p-values less than .01

WH	BLEU	GTM	METEOR	TER
WHO	.253	.292	.320	.198
WHERE	.229	.185	.252	.251
WHEN	.250	.158	.237	.331

Table 9: Autometric correlation with hit rate on individual document scores cross-classified by WH type. All scores are shown to be permutationally significant with p-values less than .01

Thus similarly, Table 9 shows that there are significant relationships after grouping the data by WH-type.

When documents are further classified into the 3×3 (WH \times MT) grouping, it appears in several of the cases in Table 10, with the exception of WHO documents, the relationship between metrics and hit rate is negative thus, inconsistent.³ Also, more relationships are found to be non-significant at this finer classification. Yet in some cases, there are very strong relationships between hit rate and autometric as demonstrated by the Spearman correlation value of GTM (.754) for WHO documents from MT-3.

This section showed that autometrics reflect task performance rates even within different cross-classifications of our data. However, this relationship is not always consistent. In the next section, we use this finding to take into consideration the grouping effects that are reflected in the original correlations we found in Section 3.3. We introduce a method for partial correlation that reveal the true population-wide association after removing the MT \times WH group effects.

³There are about 115 data points within each MT \times WH group in this table.

WH	MT	BLEU	GTM	METEOR	TER
WHO	1	.571	.322	.217	.123(.18)
	2	.161(.09)	.285	.467	.172(.07)
	3	.472	.754	.694	.478
WHERE	1	.195	.274	.116(.20)	.322
	2	-.227	.094(.29)	-.020(.83)	-.012(.89)
	3	.006(.95)	.030(.75)	-.128(.17)	-.111(.24)
WHEN	1	-.220	-.161(.08)	-.098(.28)	.074(.41)
	2	.528	.456	.502	.366
	3	.492	.311	.184	.196

Table 10: Autometric correlation with Hit rate on non-aggregate individual document scores cross-classified by WH \times MT type. Permutational significance values for non-significance at the .05 level are shown in parentheses.

5.3 Partial Correlation for Task Performance Evaluation

If X_1 , X_2 , and X_3 are three random variables, the partial correlation coefficient of the variables can be calculated by the formula

$$r_{x_1x_2.x_3} = \frac{r_{x_1x_2} - r_{x_2x_3}r_{x_1x_3}}{\sqrt{(1 - r_{x_2x_3}^2)(1 - r_{x_1x_3}^2)}} \quad (2)$$

where $r_{x_1x_2}$, $r_{x_2x_3}$, and $r_{x_1x_3}$ are the ordinary Pearson r correlation coefficients obtained between the indicated pairs of variables (Conover, 1980). We are interested in the Spearman version of this partial correlation as introduced by (Kendall, 1942), but expanded to the case where variable x_3 is a discrete or categorical grouping effect, as is the case with our task data. This correction after adjusting for a grouping effect does not seem to have been studied in the literature.

Such partial correlations are sought for the categorical variable Z representing the MT, WH, and MT \times WH cross-classified groups. Partial correlations can be thought of as a way to reveal the population-wide correlation after removing the MT \times WH group effects. Our use of partial correlation in this work offers a more general methodological perspective of the partial rank correlation statistic since when considering Z , there is a decision that can be made as to how to actually rank the data. In general, this within Z stratum detection of relevant variable relationships is appropriate for many applications such as ecological and social science studies.

We derived two distinct expressions yielding two different statistics for partial rank correlation after

adjusting for categorical grouping effects.

Method 1: ρ computed from X and Y , linearly corrected for Z by the group mean. This statistic is called S_1 . Let $I_Z(z)$ denote the indicator function such that

$$|x| = \begin{cases} 1 & \text{if } z \in Z; \\ 0 & \text{if } z \notin Z. \end{cases}$$

then $X_i^* = X_i - c_{Z_i}$, $Y_i^* = Y_i - d_{Z_i}$ and

$$S_1 = \frac{Cov[R(X_i^*, \underline{X}^*)R(Y_i^*, \underline{Y}^*)]}{\sqrt{Var(R(X_i^*, \underline{X}^*))Var(R(Y_i^*, \underline{Y}^*))}} \quad (3)$$

for $j = 1, \dots, L$ indexing the different levels of Z , $c_j = \sum_{i=1}^n I_{[Z_i=j]} X_i / \sum_{i=1}^n I_{[Z_i=j]}$ and similarly $d_j = \sum_{i=1}^n I_{[Z_i=j]} Y_i / \sum_{i=1}^n I_{[Z_i=j]}$. Recall that $R(X_i^*, \underline{X}^*)$ and $R(Y_i^*, \underline{Y}^*)$ are the ranks of the corresponding X and Y values. This formula can be equivalently written as

$$\frac{\sum_{j=1}^L \left(R(X_i^*, \underline{X}^*) - \frac{n+1}{2} \right) \left(R(Y_i^*, \underline{Y}^*) - \frac{n+1}{2} \right)}{\sqrt{\sum_{j=1}^L \left(R(X_i^*, \underline{X}^*) - \frac{n+1}{2} \right)^2 \sum_{j=1}^L \left(R(Y_i^*, \underline{Y}^*) - \frac{n+1}{2} \right)^2}} \quad (4)$$

Method 2: Compute weighted combination of within Z -group ρ . This statistic is called S_2 . Let w_z be the weight given for group z and $\underline{X}^{(z)} = (x_i : i \in J_z)$ where J_z is the subset of indices $i = 1, \dots, n$ for which $Z_i = z$. Then,

$$S_2 = \frac{\sum_{z=1}^L w_z \rho_{y,x|Z=z}}{\sum_{z=1}^L w_z} \quad (5)$$

where ρ in this case is defined as:

$$\frac{\sum_{i \in J_z} \left(R(X_i, \underline{X}^{(z)}) - \frac{n+1}{2} \right) \left(R(Y_i, \underline{Y}^{(z)}) - \frac{n+1}{2} \right)}{\sqrt{\sum_{i \in J_z} \left(R(X_i, \underline{X}^{(z)}) - \frac{n+1}{2} \right)^2 \sum_{i \in J_z} \left(R(Y_i, \underline{Y}^{(z)}) - \frac{n+1}{2} \right)^2}} \quad (6)$$

The results of the partial correlation between hit rate and autometrics in the presence of the MT, WH, WH \times MT grouping effects, respectively, for statistics S_1 and S_2 can be found in Table 11. The partial correlations between task performance and autometrics after accounting for groups are slight but all partial relationships are significant permutationally with p-values equal to .001.

Group	Statistic	B	G	M	T
MT	S_1	.156	.251	.232	.178
	S_2	.193	.255	.237	.179
WH	S_1	.243	.241	.273	.270
	S_2	.244	.214	.271	.257
WH/MT	S_1	.200	.300	.243	.183
	S_2	.200	.256	.204	.174

Table 11: Autometric partial rank correlation with hit rate on non-aggregate individual document scores by grouping effect. All values are permutational significant with p-value equal to .001. B-BLEU, G-GTM, B-METEOR, T-TER

6 Conclusion and Future Work

Current evaluation methods have considerably furthered the development of translation engines based on the system developer’s ability to obtain a numerical estimate of the system’s current capabilities. A universally good metric would not only be a metric that aids in system tuning but gives an assessment of the utility of the document for purposes because “there are no absolute standards of translation quality but only more or less appropriate translations for the purpose for which they are intended” (Sager, 1989).

We have begun to tackle the autometric/task-based metric relationship problem through the use of statistical methods. This provides a practical way

of evaluating translation by enabling users to identify what types of tasks can be performed using the output, while still utilizing fast, reusable automated evaluation metrics. Most of the statistical tools discussed in this paper are standard except for the partial rank correlation statistics. Our application of correlational and permutational tools here were customized for this particular MT evaluation application and serve as a case study without any precursor in the MT literature.

It was shown that autometric sensitivity to granularity can be exposed when trying to assess task performance. This study calls for document-level autometrics. We found that even though correlations are quite low at this level, there is a slight relationship when autometrics are considered within the cross-classifications of other variables in the study, namely the MT system that translated a particular document or the actual WH-task at hand. There is certainly variety in these relationships, group by group, and even though they are hard to see by eye and may be weak, permutational testing shows that the relationship is one that cannot be ascribed to chance.

Our findings have motivated several avenues for follow on work that extend beyond simple correlation for determining the effectiveness of autometrics, particularly for assessing document utility. First, an additional approach that we have pursued to determine whether autometrics can play a role in task based evaluation is to recode the metrics to remove their dependence on the MT system and to use the new variable to test if the metrics have additional information concerning the relationship with task response rate. The method proposed is to take each metric and find a non-MT dependent variant. This is achieved by averaging across MT systems the ratios of autometric scores divided by the within-MT average over documents. Promising results have shown that in some instances, this recoded variant of the autometrics perform better at predicting task performance than the original metric. Secondly, the magnitude of the partial rank correlation for each group (MT, WH, or MT \times WH) and autometric (BLEU, METEOR, GTM, and TER) gives us a basis for using more advance statistical regression models by serving as indicators of factors that make good predictors in modeling this relationship. Current results from modeling show that models with a combination

of recodes of both BLEU and METEOR together are the most important document level difficulty variables that describe task performance.

Acknowledgments

The author would like to acknowledge professors Eric Slud and Bonnie Dorr at the University of Maryland, College Park and Clare Voss of the US Army Research laboratory for their guidance on the PhD thesis work from which the results of this paper are derived.

References

- S. Banerjee and A. Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at ACL*. Ann Arbor, Michigan. (2005).
- K. Church and E. Hovy. Good Applications for Crummy Machine Translation. *Machine Translation*, **8**. (1993).
- William S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, **74**, 368. (1979).
- W. J. Conover. *Practical Nonparametric Statistics*. Wiley, New York. (1980).
- M. Fuji, N. Hatanaka, E. Ito, S. Kamei, H. Kumai, T. Sukehiro and H. Isahara. Evaluation Method for Determining Groups of Users Who Find MT Useful. In *MT Summit VIII: Machine Translation in the Information Age*. Santiago de Compostela, Spain, pp. 103108. (2001).
- Phillip I. Good. *Resampling Methods: A Practical Guide to Data Analysis*. Birkhäuser, Boston. (1999).
- Doug Jones, Wade Shen, Neil Granoien, Martha Herzog, and Clifford Weinstein. Measuring Translation Quality by Testing English Speakers with a New Defense Language Proficiency Test for Arabic. *International Conference on Intelligence Analysis*. McLean, VA. (2005).
- M. Kendall. Partial Rank Correlation. *Biometrika*, **30**, 81-93. (1942).
- A. Lavie and A. Agarwal. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of Workshop on Statistical Machine Translation at ACL*. Prague. (2007).
- E. L. Lehmann and H. J. M. D’Abrera. *Nonparametrics: Statistical Methods Based on Ranks*. Prentice-Hall, Englewood Cliffs, NJ. (1998).
- Gregor Leusch, Nicola Ueffing and Hermann Ney. CDER: Efficient MT Evaluation Using Block Movements. In *Proceedings of the 11th Conference of EACL*. (2006).
- I. Dan Melamed, Ryan Green and Joseph P. Turian. Precision and Recall of Machine Translation. In *Proceedings of HLT-NAACL*. Edmonton, Canada. (2003).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation, In *Proceedings of the 40th Annual Meeting of ACL*. Philadelphia, Pennsylvania. (2002).
- Florence Reeder and John White. Granularity in MT Evaluation. In *Proceedings of Towards Systematizing MT Evaluation: A Workshop on Machine Translation Evaluation at the MT Summit IX*. New Orleans, LA. (2003).
- J. C. Sager. Quality and Standards: The Evaluation of Translations. In *The Translators Handbook* C. Picken (Eds.). London. (1989).
- M. Snover, B. J. Dorr, R. Schwartz, J. Makhoul, L. Micciulla, and R. Weischedel. A Study of Translation Error Rate with Targeted Human Annotation. Technical Report LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58. University of Maryland, College Park. (2005).
- Calandra R. Tate and Clare R. Voss. Combining Evaluation Metrics Via Loss Functions. In *Proceedings of AMTA*. Boston, MA. (2006).
- Calandra Tate, Sooyon Lee, Clare Voss. Task-based MT Evaluation: Tackling Software, Experimental Design, & Statistical Models. In *Proceedings of Workshop on Machine Translation Evaluation Towards Systematizing MT Evaluation at MT Summit IX*. New Orleans, LA. (2003).
- Kathryn Taylor and John White. Predicting What MT is Good for: User Judgments and Task Performance. In *Proceedings of the Third Conference of AMTA*. Langhorne, PA. (1998).
- Masaru Tomita. Application of the TOEFL Test to the Evaluation of English-Japanese MT. In *Proceedings of a Workshop sponsored by the NSF*. San Diego, CA. (1992).
- Joseph P. Turian, Luke Shen and I. Dan Melamed. Evaluation of Machine Translation and its Evaluation, In *Proceedings of MT Summit IX*, New Orleans, LA. (2003).
- M. Vanni and K. Miller. Scaling the ISLE Framework: Validating Tests of Machine Translation Quality for Multi-Dimensional Measurement, In *Proceedings of the Workshop on MT Evaluation at the MT Summit VIII*, Santiago de Compostela, Spain. (2001).
- Clare R. Voss and Calandra R. Tate. Task-based Evaluation of Machine translation (MT) Engines: Measuring How Well People Extract Who, When, Where-Type Elements in MT Output”, In *Proceedings of EAMT*. Oslo, Norway. (2006).