Ranking Human and Machine Summarization Systems

Peter Rankel
University of Maryland
College Park, Maryland
rankel@math.umd.edu

John M. Conroy
IDA/Center for Computing Sciences
Bowie, Maryland
conroyjohnm@gmail.com

Eric V. Slud
University of Maryland
College Park, Maryland
evs@math.umd.edu

Dianne P. O'Leary
University of Maryland
College Park, Maryland
oleary@cs.umd.edu

Abstract

The Text Analysis Conference (TAC) ranks summarization systems by their average score over a collection of document sets. We investigate the statistical appropriateness of this score and propose an alternative that better distinguishes between human and machine evaluation systems.

1 Introduction

For the past several years, the National Institute of Standards and Technology (NIST) has hosted the Text Analysis Conference (TAC) (previously called the Document Understanding Conference (DUC)) (Nat, 2010). A major theme of this conference is multi-document summarization: machine summarization of sets of related documents, sometimes query-focused and sometimes generic. The summarizers are judged by how well the summaries match human-generated summaries in either automatic metrics such as ROUGE (Lin and Hovy, 2003) or manual metrics such as responsiveness or pyramid evaluation (Nenkova et al., 2007). Typically the systems are ranked by their average score over all document sets.

Ranking by average score is quite appropriate under certain statistical hypotheses, for example, when each sample is drawn from a distribution which differs from the distribution of other samples only through a location shift (Randles and Wolfe, 1979). However, a non-parametric (rank-based) analysis of variance on the summarizers' scores on each document set revealed an impossibly small p-value (less

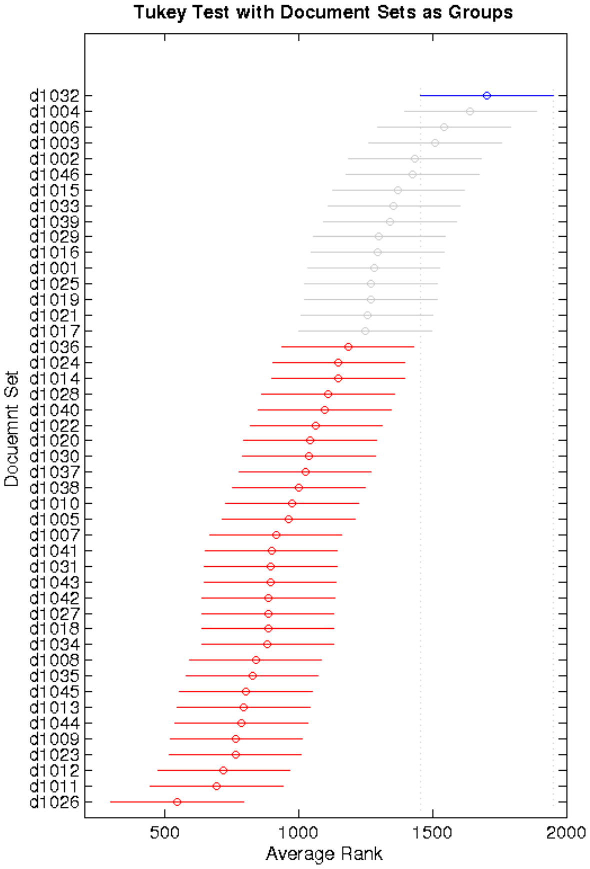

Figure 1: Confidence Intervals from a non-parametric Tukey's honestly significant difference test for 46 TAC 2010 update document sets. The blue confidence interval (for document set d1032) does not overlap any of the 30 red intervals. Hence, the test concludes that 30 document sets have mean significantly different from the mean of d1032.

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 467–473,
Edinburgh, Scotland, UK, July 27–31, 2011. ©2011 Association for Computational Linguistics

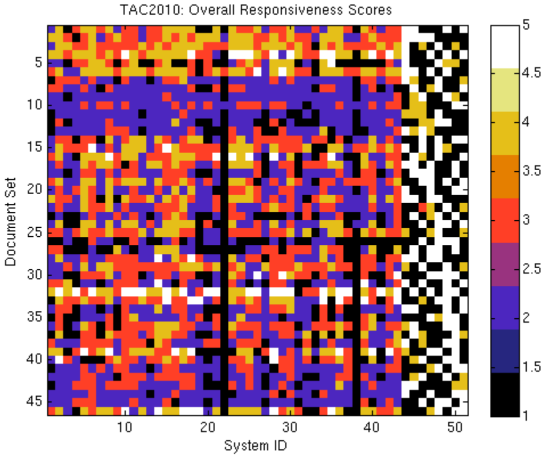

Figure 2: Overall Responsiveness scores.

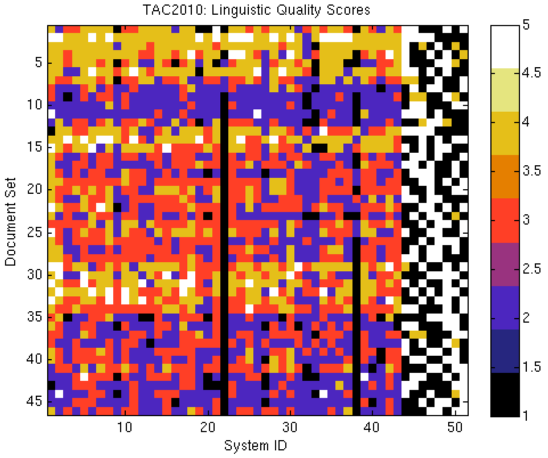

Figure 3: Linguistic scores.

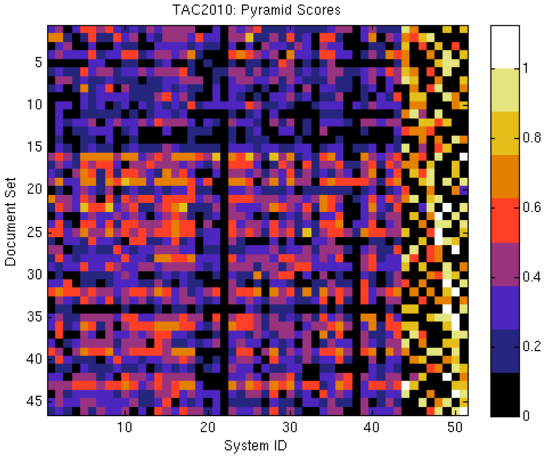

Figure 4: Pyramid scores.

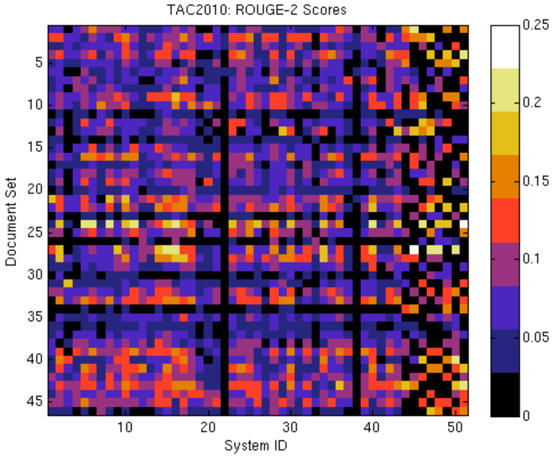

Figure 5: ROUGE-2 scores for the TAC 2010 update summary task, organized by document set (y-axis) and summarizer (x-axis). The 51 summarizers fall into two distinct groups: machine systems (first 43 columns) and humans (last 8 columns). Note that each human only summarized half of the document sets, thus creating 23 missing values in each of the last 8 columns. Black is used to indicate missing values in the last 8 columns and low scores in the first 43 columns.

than 10^{-12} using Matlab's `kruskalwallis` [1]), providing evidence that a summary's score is not independent of the document set. This effect can be seen in Figure 1, showing the confidence bands, as computed by a Tukey honestly significant difference test for each document set's difficulty as measured by the mean rank responsiveness score for TAC 2010. The test clearly shows that the summarizer performances on different document sets have different averages.

We further illustrate this in Figures 2 – 5, which show the scores of various summarizers on various document sets using standard human and automatic evaluation methods (Dang and Owczarzak, 2008) of overall responsiveness, linguistic quality, pyramid scores, and ROUGE-2 using color to indicate the value of the score. Some rows are clearly darker, indicating overall lower scores for the sum-

[1] The Kruskal-Wallis test performs a one-way analysis of variance of document-set differences after first converting the summary scores for each sample to their ranks within the pooled sample. Computed from the converted scores, the Kruskal-Wallis test statistic is essentially the ratio of the between-group sum of squares to the combined within-group sum of squares.

maries of these documents, and the variances of the scores differ row-by-row. These plots show qualitatively what the non-parametric analysis of variance demonstrates statistically. While the data presented was for the TAC 2010 update document sets, similar results hold for all the TAC 2008, 2009, and 2010 data. Hence, it may be advantageous to measure summarizer quality by accounting for heterogeneity of documents within each test set. A non-parametric paired test like the Wilcoxon signed-rank is one way to do this. Another way would be paired t-tests.

In the paper (Conroy and Dang, 2008) the authors noted that while there is a significant gap in performance between machine systems and human summarizers when measured by average manual metrics, this gap is not present when measured by the averages of the best automatic metric (ROUGE). In particular, in the DUC 2005-2007 data some systems have ROUGE performance within the 95% confidence intervals of several human summarizers, but their pyramid, linguistic, and responsiveness scores do not achieve this level of performance. Thus, the inexpensive automatic metrics, as currently employed, do not predict well how machine summaries compare to human summaries.

In this work we explore the use of document-paired testing for summarizer comparison. Our main approach is to consider each pair of two summarizers' sets of scores (over all documents) as a balanced two-sample dataset, and to assess that pair's mean difference in scores through a two-sample T or Wilcoxon test, paired or unpaired. Our goal has been to confirm that human summarizer scores are uniformly different and better on average than machine summarizer scores, and to rate the quality of the statistical method (T or W, paired or unpaired) by the consistency with which the human versus machine scores show superior human performance. Our hope is that paired testing, using either the standard paired two-sample t-test or the distribution-free Wilcoxon signed-rank test, can provide greater power in the statistical analysis of automatic metrics such as ROUGE.

2 Size and Power of Tests

Statistical tests are generally compared by choosing rejection thresholds to achieve a certain small probability of Type I error (usually as $\alpha = .05$). Given multiple tests with the same Type I error, one prefers the test with the smallest probability of Type II error. Since power is defined to be one minus the Type II error probability, we prefer the test with the most power. Recall that a *test-statistic* S depending on available data-samples gives rise to a *rejection region* by defining rejection of the null hypothesis H_0 as the event $\{S \geq c\}$ for a *cutoff* or *rejection threshold* c chosen so that

$$P(S \geq c) \leq \alpha$$

for all probability laws compatible with the null hypothesis where the (nominal) *significance level* α is chosen in advance by the statistician, usually as $\alpha = .05$. However, in many settings, the null hypothesis comprises many possible probability laws, as here where the null hypothesis is that the underlying probability laws for the score-samples of two separate summarizers are equal, without specifying exactly what that probability distribution is. In this case, the significance level is an upper bound for the attained *size* of the test, defined as $\sup_{P \in H_0} P(S \geq c)$, the largest rejection probability $P(S \geq c)$ achieved by any probability law compatible with the null hypothesis. The power of the test then depends on the specific probability law Q from the considered alternatives in H_A. For each such Q, and given a threshold c, the power for the test at Q is the rejection probability $Q(S \geq c)$. These definitions reflect the fact that the null and alternative hypotheses are *composite*, that is, each consists of multiple probability laws for the data. One of the advantages of considering a *distribution-free* two-sample test statistic such as the Wilcoxon is that the probability distribution for the statistic S is then the same for all (continuous, or non-discrete) probability laws $P \in H_0$, so that one cutoff c serves for all of H_0 with all rejection probabilities equal to α. [2]

Two test statistics, say S and $\tilde{S}$, are generally compared in terms of their powers at fixed alternatives Q in the alternative hypothesis H_A, when their respective thresholds c, c^* have been defined so that the sizes of the respective tests, $\sup_{P \in H_0} P(S \geq$

[2] The Wilcoxon test is not distribution-free for discrete data. However, the discrete TAC data can be thought of as rounded continuous data, rather than as truly discrete data.

c) and $\sup_{P \in H_0} P(\tilde{S} \geq c^*)$, are approximately equal. In this paper, the test statistics under consideration are – in one-sided testing — the (unpaired) two-sample t test with pooled sample variance (T), the paired two-sample t test (T^p), and the (paired) signed-rank Wilcoxon test (W); and for two-sided testing, S is defined by the absolute value of one of these statistics. The thresholds c for the tests can be defined either by theoretical distributions, by large-sample approximations, or by data-resampling (*bootstrap*) techniques, and (only) in the last case are these thresholds data-dependent, or random. We explain these notions with respect to the two-sample data-structure in which the scores from the first summarizer are denoted $X_1, \ldots, X_n$, where n is the number of documents with non-missing scores for both summarizers, and the scores from the second summarizer are $Y_1, \ldots, Y_n$. Let $Z_k = X_k - Y_k$ denote the document-wise differences between the summarizers' scores, and $\bar{Z} = n^{-1} \sum_{k=1}^{n} Z_k$ be their average. Then the paired statistics are defined as

$$T^p = \sqrt{n(n-1)}\, \bar{Z} / (\sum_{k=1}^{n} (Z_k - \bar{Z})^2)^{1/2}$$

and

$$W = \sum_{k=1}^{n} \mathrm{sgn}(Z_k)\, R_k^+$$

where R_k^+ is the rank of $|Z_k|$ among $|Z_1|, \ldots, |Z_n|$. Note that under both null and alternative hypotheses, the variates Z_k are assumed independent identically distributed (*iid*), while under H_0, the random variables Z_k are symmetric about 0.

The t-statistic T^p is 'parametric' in the sense that exact theoretical calculations of probabilities $P(a < T^p < b)$ depend on the assumption of normality of the differences Z_k, and when that holds, the two-sided cutoff $c = c(T^p)$ is defined as the $1 - \alpha/2$ quantile of the t_{n-1} distribution with $n-1$ degrees of freedom. However, when n is moderately or very large, the cutoff is well approximated by the standard-normal $1 - \alpha/2$ quantile $z_{\alpha/2}$, and T^p becomes approximately nonparametrically valid with this cutoff, by the Central Limit Theorem. The Wilcoxon signed-rank statistic W has theoretical cutoff $c = c(W)$ which depends only on n, whenever the data Z_k are continuously distributed; but for large n, the cutoff is given simply as $\sqrt{n^3/12} \cdot z_{\alpha/2}$. When there are ties (as might be common in discrete data), the calculation of cutoffs and p-values for Wilcoxon becomes slightly more complicated and is no longer fully nonparametric except in a large-sample approximate sense.

The situation for the two-sample unpaired t-statistic T currently used in TAC evaluation is not so neat. Even when the two samples $\mathbf{X} = \{X_k\}_{k=1}^n$ and $\mathbf{Y} = \{Y_k\}_{k=1}^n$ are independent, exact theoretical distribution of cutoffs is known only under the parametric assumption that the scores are normally distributed (and in the case of the pooled-sample-variance statistic, that $\mathrm{Var}(X_k) = \mathrm{Var}(Y_k)$.) However, an essential element of the summarization data is the heterogeneity of documents. This means that while $\{X_k\}_{k=1}^n$ can be viewed as *iid* scores when documents are selected randomly – and not necessarily equiprobably – from the ensemble of all possible documents, the Y_k and X_k samples are *dependent*. Still, the pairs $\{(X_k, Y_k)\}_{k=1}^n$, and therefore the differences $\{Z_k\}_{k=1}^n$, are *iid* which is what makes paired testing valid. However, there is no theoretical distribution for T from which to calculate valid quantiles c for cutoffs, and therefore the use of the unpaired t-statistic cannot be recommended for TAC evaluation.

What can be done in a particular dataset, like the TAC summarization score datsets we consider, to ascertain the approximate validity of theoretically derived large-sample cutoffs for test statistics? In the age of plentiful and fast computers, quite a lot, through the powerful computational machinery of the *bootstrap* (Efron and Tibshirani, 1993).

The idea of bootstrap hypothesis testing (Efron and Tibshirani, 1993), (Bickel and Ren, 2001) is to randomly sample with replacement (the rows with non-missing data in) the dataset $\{(X_k, Y_k)\}_{k=1}^n$ in such a way as to generate representative data that plausibly *would* have been seen if two-sample score data had been generated from two equally effective summarizers with score distributional characteristics like the pooled scores from the two observed summarizers. We have done this in two distinct ways, each creating 2000 datasets with n paired scores:

MC *Monte Carlo Method.* For each of many iterations (in our case 2000), define a new dataset $\{(X'_k, Y'_k)\}_{k=1}^n$ by independently swapping X_k and Y_k with probability $1/2$. Hence, $(X'_k, Y'_k) = (X_k, Y_k)$ with probability $1/2$ and (Y_k, X_k) with probability $1/2$.

HB *Hybrid MC/Bootstrap.* For each of 2000 iterations, create a re-sampled dataset $\{(X''_k, Y''_k)\}_{k=1}^n$ in the following way. First, sample n pairs (X_k, Y_k) with replacement from the original dataset. Then, as above, randomly swap the components of each pair, each with $1/2$ probability.

Both of these two methods can be seen to generate two-sample data satisfying H_0, with each score-sample's distribution obtained as a mixture of the distributions actually generating the $\mathbf{X}$ and $\mathbf{Y}$ samples. The *empirical q^{th} quantiles* for a statistic $S = S(\mathbf{X}, \mathbf{Y})$ such as $|W|$ or $|T^p|$ are estimated from the resampled data as $\hat{F}_S^{-1}(q)$, where $\hat{F}_S(t)$ is simply the fraction of times (out of 2000) that the statistic S applied to the constructed dataset had a value less than or equal to t. The upshot is that the $1-\alpha$ empirical quantile for S based on either of these simulation methods serves as a data-dependent cutoff c attaining approximate size α for all H_0-generated data. The MC and HB methods will be employed in Section 4 to check the theoretical p-values.

3 Relative Efficiency of W versus T^p

Statistical theory does have something to say about the comparative powers of paired W versus T^p statistics. These statistics have been studied (Randles and Wolfe, 1979), in terms of their *asymptotic relative efficiency* for location-shift alternatives based on symmetric densities ($f(z-\vartheta)$ is a location-shift of $f(z)$). For many pairs of parametric and rank-based statistics $S, \tilde{S}$, including W and T^p, the following assertion has been proved for testing H_0 at significance level α.

First assume the Z_k are distributed according to some density $f(z-\vartheta)$, where $f(z)$ is a symmetric function ($f(-z) = f(z)$). Next assume $\vartheta = 0$ under H_0. When n gets large the powers at any alternatives with very small $\vartheta = \gamma/\sqrt{n},\ \gamma \neq 0$, can be made asymptotically equal by using samples of size n with statistic S and of size $\rho \cdot n$ with statistic $\tilde{S}$. Here $\rho = ARE(S, \tilde{S})$ is a constant not depending on n or γ but definitely depending on f, called *asymptotic relative efficiency* of S with respect to $\tilde{S}$. (The smaller $\rho < 1$ is, the more statistic $\tilde{S}$ is preferred among the two.)

Using this definition, it is known (Randles and Wolfe 1979, Sec. 5.4 leading up to Table 5.4.7 on p. 167) that the Wilcoxon signed-rank statistic W provides greater robustness and often much greater efficiency than the paired T, with ARE which is 0.95 with f a standard normal density, and which is never less than 0.864 for any symmmetric density f. However, in our context, continuous scores such as pyramid exhibit document-specific score differences between summarizers which often have approximately normal-looking histograms, and although the alternatives perhaps cannot be viewed as pure location shifts, it is unsurprising in view of the ARE theory cited above that the W and T paired tests have very similar performance. Nevertheless, as we found by statistical analysis of the TAC data, both are far superior to the unpaired T-statistic, with either theoretical or empirical bootstrapped p-values.

4 Testing Setup and Results

To evaluate our ideas, we used the TAC data from 2008-2010 and focused on three manual metrics (overall responsiveness, pyramid score, and linguistic quality score) and two automatic metrics (ROUGE-2 and ROUGE-SU4). We make the assumption, backed by both the scores given and comments made by NIST summary assessors [3], that automatic summarization systems do not perform at the human level of performance. As such, if a statistic based on an automatic metric, such as ROUGE-2, were to show fewer systems performing at human level of performance than the statistic of averaging scores, such a statistic would be preferable because

[3] Assessors have commented privately at the Text Analysis Conference 2008, that while the origin of the summary is hidden from them, "we know which ones are machine generated." Thus, automatic summarization fails the Turing test of machine intelligence (Turing, 1950). This belief is also supported by (Conroy and Dang, 2008) and (Dang and Owczarzak, 2008). Finally, our own results show no matter how you compare human and machine scores all machines systems score significantly worse than humans.

	2008: $2145 = \binom{66}{2}$ pairs			2009: $1830 = \binom{61}{2}$ pairs			2010: $1275 = \binom{51}{2}$ pairs		
Metric	**Unpair-T**	**Pair-T**	**Wilc.**	**Unpair-T**	**Pair-T**	**Wilc.**	**Unpair-T**	**Pair-T**	**Wilc.**
Linguistic	1234	**1416**	1410	1000	**1182**	1173	841	**939**	934
Overall	1202	**1353**	1342	982	**1149**	1146	845	**894**	889
Pyramid	1263	1417	**1418**	1075	**1238**	1216	875	**933**	926
ROUGE-2	1243	1453	**1459**	1016	1182	**1193**	812	938	**939**
ROUGE-SU4	1333	1493	**1507**	1059	1241	**1254**	894	**983**	976

Table 1: Number of significant differences found when testing for the difference of all pairs of summarization systems (including humans).

	2008: $464 = 58 \times 8$ pairs			2009: $424 = 53 \times 8$ pairs			2010: $344 = 43 \times 8$ pairs		
Metric	**Unpair-T**	**Pair-T**	**Wilc.**	**Unpair-T**	**Pair-T**	**Wilc.**	**Unpair-T**	**Pair-T**	**Wilc.**
Linguistic	464	464	464	424	424	424	344	344	344
Overall	464	464	464	424	424	424	344	344	344
Pyramid	464	464	464	424	424	424	344	344	344
ROUGE-2	375	**409**	402	323	**350**	341	275	**309**	305
ROUGE-SU4	391	**418**	414	354	**378**	373	324	**331**	328

Table 2: Number of significant differences resulting from $8 \times (N - 8)$ tests for human-machine system means or signed-rank comparisons.

of its greater power in the machine vs. human summarization domain.

For each of these metrics, we first created a score matrix whose (i, j)-entry represents the score for summarizer j on document set i (these matrices generated the colorplots in Figures 2 – 5). We then performed a Wilcoxon signed-rank test on certain pairs of columns of this matrix (any pair consisting of one machine system and one human summarizer). As a baseline, we did the same testing with a paired and an unpaired t-test. Each of these tests resulted in a p-value, and we counted how many were less than .05 and called these the significant differences.

The results of these tests (shown in Table 2), were somewhat surprising. Although we expected the nonparametric signed-rank test to perform better than an unpaired t-test, we were surprised to see that a paired t-test performed even better. All three tests always reject the null hypotheses when human metrics are used. This is what we'd like to happen with automatic metrics as well. As seen from the table, the paired t-test and Wilcoxon signed-rank test offer a good improvement over the unpaired t-test.

The results in Table 1 are less clear, but still positive. In this case, we are comparing pairs of machine summarization systems. In contrast to the human vs. machine case, we do not know the truth here. However, since the number of significant differences increases with paired testing here as well, we believe this also reflects the greater discriminatory power of paired testing.

We now apply the Monte Carlo and Hybrid Monte Carlo to check the theoretical p-values reported in Tables 1 and 2. The empirical quantiles found by these methods generally confirm the theoretical p-value test results reported there, especially in Table 2. In the overall tallies of all comparisons (Table 1), it seems that the bootstrap results (comparing only W and the un-paired T) make W look still stronger for linguistic and overall responsiveness versus the T; but for the pyramid and ROUGE scores, the bootstrap p-values bring T slightly closer to W although it still remains clearly inferior, achieving roughly 10% fewer rejections.

5 Conclusions and Future Work

In this paper we observed that summarization systems' performance varied significantly across document sets on the Text Analysis Conference (TAC) data. This variance in performance suggested that paired testing may be more appropriate than the t-test currently employed at TAC to compare the performance of summarization systems. We proposed a non-parametric test, the Wilcoxon signed-rank test, as a robust more powerful alternative to the t-test. We estimated the statistical power of the t-test and the Wilcoxon signed-rank test by calculating the number of machine systems whose performance was significantly different than that of human summarizers. As human assessors score machine systems as not achieving human performance in either content or responsiveness, automatic metrics such as ROUGE should ideally indicate this distinction. We found that the paired Wilcoxon test significantly increases the number of machine systems that score significantly different than humans when the pairwise test is performed on ROUGE-2 and ROUGE-SU4 scores. Thus, we demonstrated that the Wilcoxon paired test shows more statistical power than the t-test for comparing summarization systems.

Consequently, the use of paired testing should not only be used in formal evaluations such as TAC, but also should be employed by summarization developers to more accurately assess whether changes to an automatic system give rise to improved performance.

Further study needs to analyze more summarization metrics such as those proposed at the recent NIST evaluation of automatic metrics, Automatically Evaluating Summaries of Peers (AESOP) (Nat, 2010). As metrics become more sophisticated and aim to more accurately predict human judgements such as overall responsiveness and linguistic quality, paired testing seems likely to be a more powerful statistical procedure than the unpaired t-test for head-to-head summarizer comparisons.

Throughout our research in this paper, we treated each separate kind of scores on a document set as data for one summarizer to be compared with the same kind of scores for other summarizers. However, it might be more fruitful to treat *all* the scores as multivariate data and compare the summarizers that way. Multivariate statistical techniques such as Principal Component Analysis may play a constructive role in suggesting highly discriminating new composite scores, perhaps leading to statistics with even more power to measure a summary's quality.

ROUGE was inspired by the success of the BLEU (BiLingual Evaluation Understudy), an n-gram based evaluation for machine translation (Papineni et al., 2002). It is likely that paired testing may also be appropriate for BLEU as well and will give additional discriminating power between machine translations and human translations.

References

Peter J. Bickel and Jian-Jian Ren. 2001. The Bootstrap in Hypothesis Testing. In *State of the Art in Statistics and Probability Theory, Festschrift for Willem R. van Zwet*, volume 36 of *Lecture Notes– Monograph Series*, pages 91–112. Institute of Mathematical Statistics.

John M. Conroy and Hoa Trang Dang. 2008. Mind the Gap: Dangers of Divorcing Evaluations of Summary Content from Linguistic Quality. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 145–152, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hoa T. Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task. In *Proceedings of the 1st Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

B. Efron and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-Occurrences Statistics. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Alberta.

National Institute of Standards and Technology. 2010. *Text Analysis Conference, http://www.nist.gov/tac*.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Transactions on Speech and Language Processing*, 4(2).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

R.H. Randles and D.A. Wolfe. 1979. *Introduction to the Theory of Nonparametric Statistics*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley.

Alan Turing. 1950. Computing Machinery and Intelligence. *Mind*, 59(236):433–460.