Randomized Significance Tests in Machine Translation

Yvette Graham **Nitika Mathur** **Timothy Baldwin**
Department of Computing and Information Systems
The University of Melbourne
ygraham@unimelb.edu.au, nmathur@student.unimelb.edu.au, tb@ldwin.net

Abstract

Randomized methods of significance testing enable estimation of the probability that an increase in score has occurred simply by chance. In this paper, we examine the accuracy of three randomized methods of significance testing in the context of machine translation: paired bootstrap resampling, bootstrap resampling and approximate randomization. We carry out a large-scale human evaluation of shared task systems for two language pairs to provide a gold standard for tests. Results show very little difference in accuracy across the three methods of significance testing. Notably, accuracy of all test/metric combinations for evaluation of English-to-Spanish are so low that there is not enough evidence to conclude they are any better than a random coin toss.

1 Introduction

Automatic metrics, such as BLEU (Papineni et al., 2002), are widely used in machine translation (MT) as a substitute for human evaluation. Such metrics commonly take the form of an automatic comparison of MT output text with one or more human reference translations. Small differences in automatic metric scores can be difficult to interpret, however, and statistical significance testing provides a way of estimating the likelihood that a score difference has occurred simply by chance. For several metrics, such as BLEU, standard significance tests cannot be applied due to scores not comprising the mean of individual sentence scores, justifying the use of randomized methods.

Bootstrap resampling was one of the early randomized methods proposed for statistical significance testing of MT (Germann, 2003; Och, 2003; Kumar and Byrne, 2004; Koehn, 2004), to assess for a pair of systems how likely a difference in BLEU scores occurred by chance. Empirical tests detailed in Koehn (2004) show that even for test sets as small as 300 translations, BLEU confidence intervals can be computed as accurately as if they had been computed on a test set 100 times as large.

Approximate randomization was subsequently proposed as an alternate to bootstrap resampling (Riezler and Maxwell, 2005). Theoretically speaking, approximate randomization has an advantage over bootstrap resampling, in that it does not make the assumption that samples are representative of the populations from which they are drawn. Both methods require some adaptation in order to be used for the purpose of MT evaluation, such as combination with an automatic metric, and therefore it cannot be taken for granted that approximate randomization will be more accurate in practice. Within MT, approximate randomization for the purpose of statistical testing is also less common.

Riezler and Maxwell (2005) provide a comparison of approximate randomization with bootstrap resampling (distinct from *paired* bootstrap resampling), and conclude that since approximate randomization produces higher p-values for a set of apparently equally-performing systems, it more conservatively concludes statistically significant differences, and recommend preference of approximate randomization over bootstrap resampling for MT evaluation. Conclusions drawn from experiments provided in Riezler and Maxwell (2005) are oft-cited, with experiments interpreted as evidence that bootstrap resampling is overly optimistic in reporting significant differences (Riezler and Maxwell, 2006; Koehn and Monz, 2006; Galley and Manning, 2008; Green et al., 2010; Monz, 2011; Clark et al., 2011).

Our contribution in this paper is to revisit statistical significance tests in MT — namely, bootstrap resampling, paired bootstrap resampling and

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 266–274,
Baltimore, Maryland USA, June 26–27, 2014.

approximate randomization — and find problems with the published formulations. We redress these issues, and apply the tests in statistical testing of two language pairs. Using human judgments of translation quality, we find only very minor differences in significance levels across the three tests, challenging claims made in the literature about relative merits of tests.

2 Revisiting Statistical Significance Tests for MT Evaluation

First, we revisit the formulations of bootstrap resampling and approximate randomization algorithms as presented in Riezler and Maxwell (2005). At first glance, both methods appear to be two-tailed tests, with the null hypothesis that the two systems perform equally well. To facilitate a two-tailed test, absolute values of pseudo-statistics are computed before locating the absolute value of the actual statistic (original difference in scores). Using absolute values of pseudo-statistics is not problematic in the approximate randomization algorithm, and results in a reasonable two-tailed significance test. However, the bootstrap algorithm they provide uses an additional shift-to-zero method of simulating the null hypothesis. The way in which this shift-to-zero and absolute values of pseudo-statistics are applied is non-standard. Combining shift-to-zero and absolute values of pseudo-statistics results in all pseudo-statistics that fall below the mean pseudo-statistic to be omitted from computation of counts later used to compute p-values. The version of the bootstrap algorithm, as provided in the pseudo-code, is effectively a one-tailed test, and since this does not happen in the approximate randomization algorithm, experiments appear to compare p-values from a one-tailed bootstrap test directly with those of a two-tailed approximate randomization test. This inconsistency is not recognized, however, and p-values are compared as if both tests are two-tailed.

A better comparison of p-values would first require doubling the values of the one-sided bootstrap, leaving those of the two-sided approximate randomization algorithm as-is. The results of the two tests on this basis are extremely close, and in fact, in two out of the five comparisons, those of the bootstrap would have marginally *higher* p-values than those of approximate randomization. As such, it is conceivable to conclude that the experiments actually show no substantial difference in Type I error between the two tests, which is consistent with results published in other fields of research (Smucker et al., 2007). We also note that the pseudo-code contains an unconventional computation of mean pseudo-statistics, τ_B, for shift-to-zero.

Rather than speculate over whether these issues with the original paper were simply presentational glitches or the actual basis of the experiments reported on in the paper, we present a normalized version of the two-sided bootstrap algorithm in Figure 1, and report on the results of our own experiments in Section 4. We compare this method with approximate randomization and also *paired* bootstrap resampling (Koehn, 2004), which is widely used in MT evaluation. We carry out evaluation over a range of MT systems, not only including pairs of systems that perform equally well, but also pairs of systems for which one system performs marginally better than the other. This enables evaluation of not only Type I error, but the overall accuracy of the tests. We carry out a large-scale human evaluation of all WMT 2012 shared task participating systems for two language pairs, and collect sufficient human judgments to facilitate statistical significance tests. This human evaluation data then provides a gold-standard against which to compare randomized tests. Since all randomized tests only function in combination with an automatic MT evaluation metric, we present results of each randomized test across four different MT metrics.

3 Randomized Significance Tests

3.1 Bootstrap Resampling

Bootstrap resampling provides a way of estimating the population distribution by sampling with replacement from a representative sample (Efron and Tibshirani, 1993). The test statistic is taken as the difference in scores of the two systems, $S_X - S_Y$, which has an expected value of 0 under the null hypothesis that the two systems perform equally well. A bootstrap pseudo-sample consists of the translations by the two systems (X_b, Y_b) of a bootstrapped test set (Koehn, 2004), constructed by sampling with replacement from the original test set translations. The bootstrap distribution S_{boot} of the test statistic is estimated by calculating the value of the pseudo-statistic $S_{X_b} - S_{Y_b}$ for each pseudo-sample.

Set $c = 0$

Compute actual statistic of score differences $S_X - S_Y$ on test data

Calculate sample mean $\tau_B = \frac{1}{B} \sum_{b=1}^{B} S_{X_b} - S_{Y_b}$ over bootstrap samples $b = 1, ..., B$

For bootstrap samples $b = 1, ..., B$

Sample with replacement from variable tuples test sentences for systems X and Y

Compute pseudo-statistic $S_{X_b} - S_{Y_b}$ on bootstrap data

If $|S_{X_b} - S_{Y_b} - \tau_B| \geq |S_X - S_Y|$

$c = c + 1$

If $c/B \leq \alpha$

Reject the null hypothesis

Figure 1: Two-sided bootstrap resampling statistical significance test for automatic MT evaluation

Set $c = 0$

Compute actual statistic of score differences $S_X - S_Y$ on test data

For random shuffles $r = 1, ..., R$

For sentences in test set

Shuffle variable tuples between systems X and Y with probability 0.5

Compute pseudo-statistic $S_{X_r} - S_{Y_r}$ on shuffled data

If $S_{X_r} - S_{Y_r} \geq S_X - S_Y$

$c = c + 1$

If $c/R \leq \alpha$

Reject the null hypothesis

Figure 2: Approximate randomization statistical significance test for automatic MT evaluation

The null hypothesis distribution S_{H_0} can be estimated from S_{boot} by applying the shift method (Noreen, 1989), which assumes that S_{H_0} has the same shape but a different mean than S_{boot}. Thus, S_{boot} is transformed into S_{H_0} by subtracting the mean bootstrap statistic from every value in S_{boot}.

Once this shift-to-zero has taken place, the null hypothesis is rejected if the probability of observing a more extreme value than the actual statistic is lower than a predetermined p-value α, which is typically set to 0.05. In other words, the score difference is significant at level $1 - \alpha$.

Figure 3 provides a one-sided implementation of bootstrap resampling, where H_0 is that the score of System X is less than or equal to the score of

Set $c = 0$

Compute actual statistic of score differences $S_X - S_Y$ on test data

Calculate sample mean $\tau_B = \frac{1}{B} \sum_{b=1}^{B} S_{X_b} - S_{Y_b}$ over bootstrap samples $b = 1, ..., B$

For bootstrap samples $b = 1, ..., B$

Sample with replacement from variable tuples test sentences for systems X and Y

Compute pseudo-statistic $S_{X_b} - S_{Y_b}$ on bootstrap data

If $S_{X_b} - S_{Y_b} - \tau_B \geq S_X - S_Y$

$c = c + 1$

If $c/B \leq \alpha$

Reject the null hypothesis

Figure 3: One-sided Bootstrap resampling statistical significance test for automatic MT evaluation

Set $c = 0$

For bootstrap samples $b = 1, ..., B$

If $S_{X_b} < S_{Y_b}$

$c = c + 1$

If $c/B \leq \alpha$

Reject the null hypothesis

Figure 4: Paired bootstrap resampling randomized significance test

System Y. Figure 5 includes a typical example of bootstrap resampling applied to BLEU, for a pair of systems for which differences in scores are significant, while Figure 6 shows the same for METEOR but for a pair of systems with no significant difference in scores.

3.2 Approximate Randomization

Unlike bootstrap, approximate randomization does not make any assumptions about the population distribution. To simulate a distribution for the null hypothesis that the scores of the two systems are the same, translations are shuffled between the two systems so that 50% of each pseudo-sample is drawn from each system. In the context of machine translation, this can be interpreted as each translation being equally likely to have been produced by one system as the other (Riezler and Maxwell, 2005).

The test statistic is taken as the difference in scores of the two systems, $S_X - S_Y$. If there is

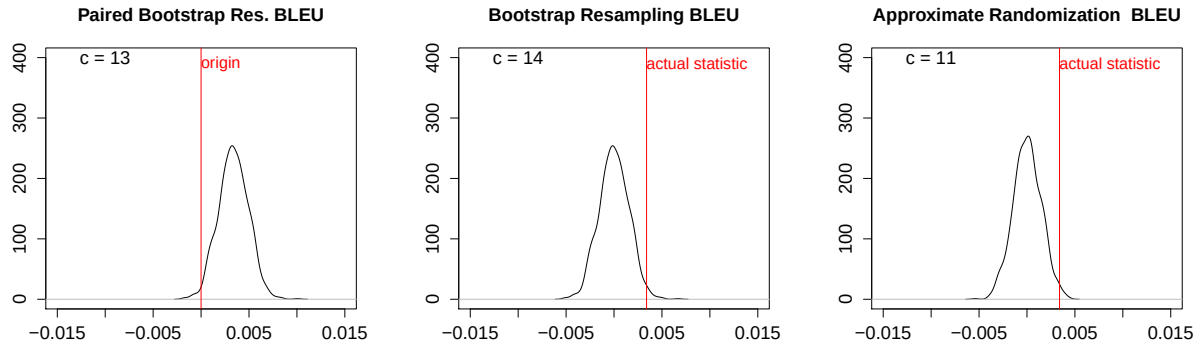

Figure 5: Pseudo-statistic distributions for a typical pair of systems with close BLEU scores for each randomized test (System F vs. System G).

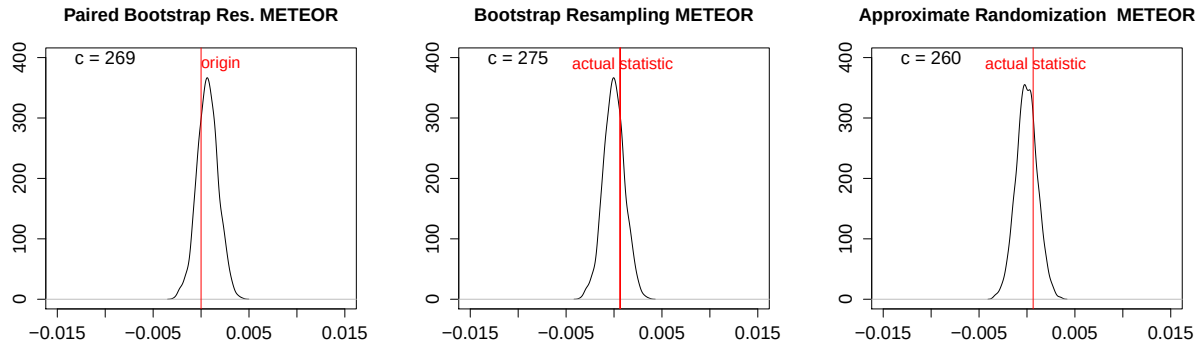

Figure 6: Pseudo-statistic distributions of METEOR with randomized tests (System D vs. System A).

a total of S sentences, then a total of 2^S shuffles is possible. If S is large, instead of generating all 2^S possible combinations, we instead generate samples by randomly permuting translations between the two systems with equal probability. The distribution of the test statistic under the null hypothesis is approximated by calculating the pseudo-statistic, $S_{X_r} - S_{Y_r}$, for each sample. As before, the null hypothesis is rejected if the probability of observing a more extreme value than the actual test statistic is lower than α.

Figure 2 provides a one-sided implementation of approximate randomization for MT evaluation, where the null hypothesis is that the score of System X is less than or equal to the score of System Y. Figure 5 shows a typical example of pseudo-statistic distributions for approximate randomization for a pair of systems with a small but significant score difference according to BLEU, and Figure 6 shows the same for METEOR applied to a pair of systems where no significant difference is concluded.

3.3 Paired Bootstrap Resampling

Paired bootstrap resampling (Koehn, 2004) is shown in Figure 4. Unlike the other two randomized tests, this method makes no attempt to simulate the null hypothesis distribution. Instead, bootstrap samples are used to estimate confidence intervals of score differences, with confidence intervals not containing 0 implying a statistically significant difference.

We compare what takes place with the two other tests, by plotting differences in scores for bootstrapped samples, $S_{X_b} - S_{Y_b}$, as shown in Figure 5 for BLEU and Figure 6 for METEOR. Instead of computing counts with reference to the actual statistic, the line through the origin provides the cut-off for counts.

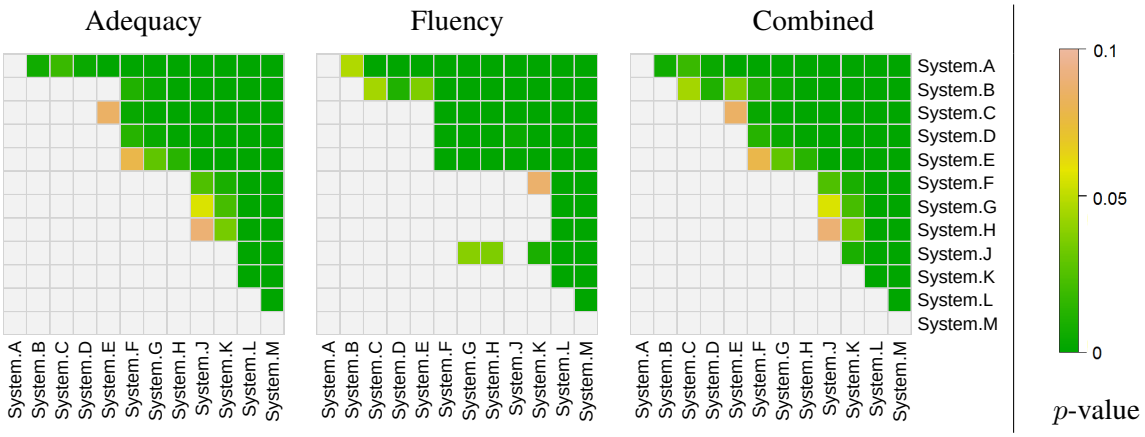

Figure 7: Human evaluation pairwise significance tests for Spanish-to-English systems (colored cells denote scores for System *row* being significantly greater than System *column*.

4 Evaluation

In order to evaluate the accuracy of the three randomized significance significance tests, we compare conclusions reached in a human evaluation of shared task participant systems. We carry out a large-scale human evaluation of all participating systems from WMT 2012 (Callison-Burch et al., 2012) for the Spanish-to-English and English-to-Spanish translation tasks. Large numbers of human assessments of translations were collected using Amazon's Mechanical Turk, with strict quality control filtering (Graham et al., 2013). A total of 82,100 human adequacy assessments and 62,400 human fluency assessments were collected. After the removal of quality control items and filtering of judgments from low-quality workers, this resulted in an average of 1,280 adequacy and 1,013 fluency assessments per system for Spanish-to-English (12 systems), and 1,483 adequacy and 1,534 fluency assessments per system for English-to-Spanish (11 systems). To remove bias with respect to individual human judge preference scoring severity/leniency, scores provided by each human assessor were standardized according to the mean and standard deviation of all scores provided by that individual.

Significance tests were carried out over the scores for each pair of systems separately for adequacy and fluency assessments using the Wilcoxon rank-sum test. Figure 7 shows pairwise significance test results for fluency, adequacy and the combination of the two tests, for all pairs of Spanish-to-English systems. Combined fluency and adequacy significance test results are constructed as follows: if a system's adequacy score is significantly greater than that of another, the combined conclusion is that it is significantly better, at that significance level. Only when a tie in adequacy scores occurs are fluency judgments used to break the tie. In this case, p-values from significance tests applied to fluency scores of that system pair are used. For example, in Figure 7, adequacy scores of System B are not significantly greater than those of Systems C, D and E, while fluency scores for System B are significantly greater than those of the three other systems. The combined result for each pair of systems is therefore taken as the p-value from the corresponding fluency significance test.

We use the combined human evaluation pairwise significant tests as a gold standard against which to evaluate the randomized methods of statistical significance testing. We evaluate paired bootstrap resampling (Koehn, 2004) and bootstrap resampling as shown in Figure 3 and approximate randomization as shown in Figure 2, each in combination with four automatic MT metrics: BLEU (Papineni et al., 2002), NIST (NIST, 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006).

4.1 Results and Discussion

Figure 8 shows the outcome of pairwise randomized significance tests for each metric for Spanish-to-English systems, and Table 1 shows numbers of correct conclusions and accuracy of each test.

When we compare conclusions made by the three randomized tests for Spanish-to-English systems, there is very little difference in p-values for all pairs of systems. For both BLEU and NIST,

α		Paired Bootst. Resamp. Conc.	Paired Bootst. Resamp. Acc.(%)	Bootst. Resamp. Conc.	Bootst. Resamp. Acc. (%)	Approx. Rand. Conc.	Approx. Rand. Acc. (%)
0.05	BLEU	53	80.3 [68.7, 89.1]	53	80.3 [68.7, 89.1]	53	80.3 [68.7, 89.1]
	NIST	**54**	**81.8** [70.4, 90.2]	**54**	**81.8** [70.4, 90.2]	**54**	**81.8** [70.4, 90.2]
	METEOR	52	78.8 [67.0, 87.9]	52	78.8 [67.0, 87.9]	52	78.8 [67.0, 87.9]
	TER	52	78.8 [67.0, 87.9]	52	78.8 [67.0, 87.9]	52	78.8 [67.0, 87.9]
0.01	BLEU	51	77.3 [65.3, 86.7]	51	77.3 [65.3, 86.7]	51	77.3 [65.3, 86.7]
	NIST	51	77.3 [65.3, 86.7]	51	77.3 [65.3, 86.7]	51	77.3 [65.3, 86.7]
	METEOR	53	80.3 [68.7, 89.1]	53	80.3 [68.7, 89.1]	53	80.3 [68.7, 89.1]
	TER	51	77.3 [65.3, 86.7]	51	77.3 [65.3, 86.7]	51	77.3 [65.3, 86.7]
0.001	BLEU	48	72.7 [60.4, 83.0]	48	72.7 [60.4, 83.0]	48	72.7 [60.4, 83.0]
	NIST	48	72.7 [60.4, 83.0]	48	72.7 [60.4, 83.0]	48	72.7 [60.4, 83.0]
	METEOR	53	80.3 [68.7, 89.1]	53	80.3 [68.7, 89.1]	52	78.8 [67.0, 87.9]
	TER	50	75.8 [63.6, 85.5]	51	77.3 [65.3, 86.7]	52	78.8 [67.0, 87.9]

Table 1: Accuracy of randomized significance tests for Spanish-to-English MT with four automatic metrics, based on the WMT 2012 participant systems.

α		Paired Bootst. Resamp. Conc.	Paired Bootst. Resamp. Acc.(%)	Bootst. Resamp. Conc.	Bootst. Resamp. Acc. (%)	Approx. Rand. Conc.	Approx. Rand. Acc. (%)
0.05	BLEU	**34**	**61.8** [47.7, 74.6]	**34**	**61.8** [47.7, 74.6]	**34**	**61.8** [47.7, 74.6]
	NIST	32	58.2 [44.1, 71.3]	32	58.2 [44.1, 71.3]	32	58.2 [44.1, 71.3]
	METEOR	31	56.4 [42.3, 69.7]	31	56.4 [42.3, 69.7]	31	56.4 [42.3, 69.7]
	TER	32	58.2 [44.1, 71.3]	32	58.2 [44.1, 71.3]	32	58.2 [44.1, 71.3]
0.01	BLEU	33	60.0 [45.9, 73.0]	33	60.0 [45.9, 73.0]	33	60.0 [45.9, 73.0]
	NIST	32	58.2 [44.1, 71.3]	32	58.2 [44.1, 71.3]	32	58.2 [44.1, 71.3]
	METEOR	31	56.4 [42.3, 69.7]	32	58.2 [44.1, 71.3]	32	58.2 [44.1, 71.3]
	TER	30	54.5 [40.6, 68.0]	30	54.5 [40.6, 68.0]	30	54.5 [40.6, 68.0]
0.001	BLEU	33	60.0 [45.9, 73.0]	33	60.0 [45.9, 73.0]	33	60.0 [45.9, 73.0]
	NIST	33	60.0 [45.9, 73.0]	32	58.2 [44.1, 71.3]	32	58.2 [44.1, 71.3]
	METEOR	32	58.2 [44.1, 71.3]	32	58.2 [44.1, 71.3]	32	58.2 [44.1, 71.3]
	TER	30	54.5 [40.6, 68.0]	30	54.5 [40.6, 68.0]	31	56.4 [42.3, 69.7]

Table 2: Accuracy of randomized significance tests for English-to-Spanish MT with four automatic metrics, based on the WMT 2012 participant systems.

all three randomized methods produce p-values so similar that when α thresholds are applied, all three tests produce precisely the same set of pairwise conclusions for each metric. When tests are combined with METEOR and TER, similar results are observed: at the α thresholds of 0.05 and 0.01, precisely the same conclusions are drawn for both metrics combined with each of the three tests, and at most a difference of two conclusions at the lowest α level.

Table 2 shows the accuracy of each test on the English-to-Spanish data, showing much the same set of conclusions at all α levels. For BLEU and NIST, all three tests again produce precisely the same conclusions, at $p < 0.01$ there is at most a single different conclusion for METEOR, and only at the lowest p-value level is there a single difference for TER.

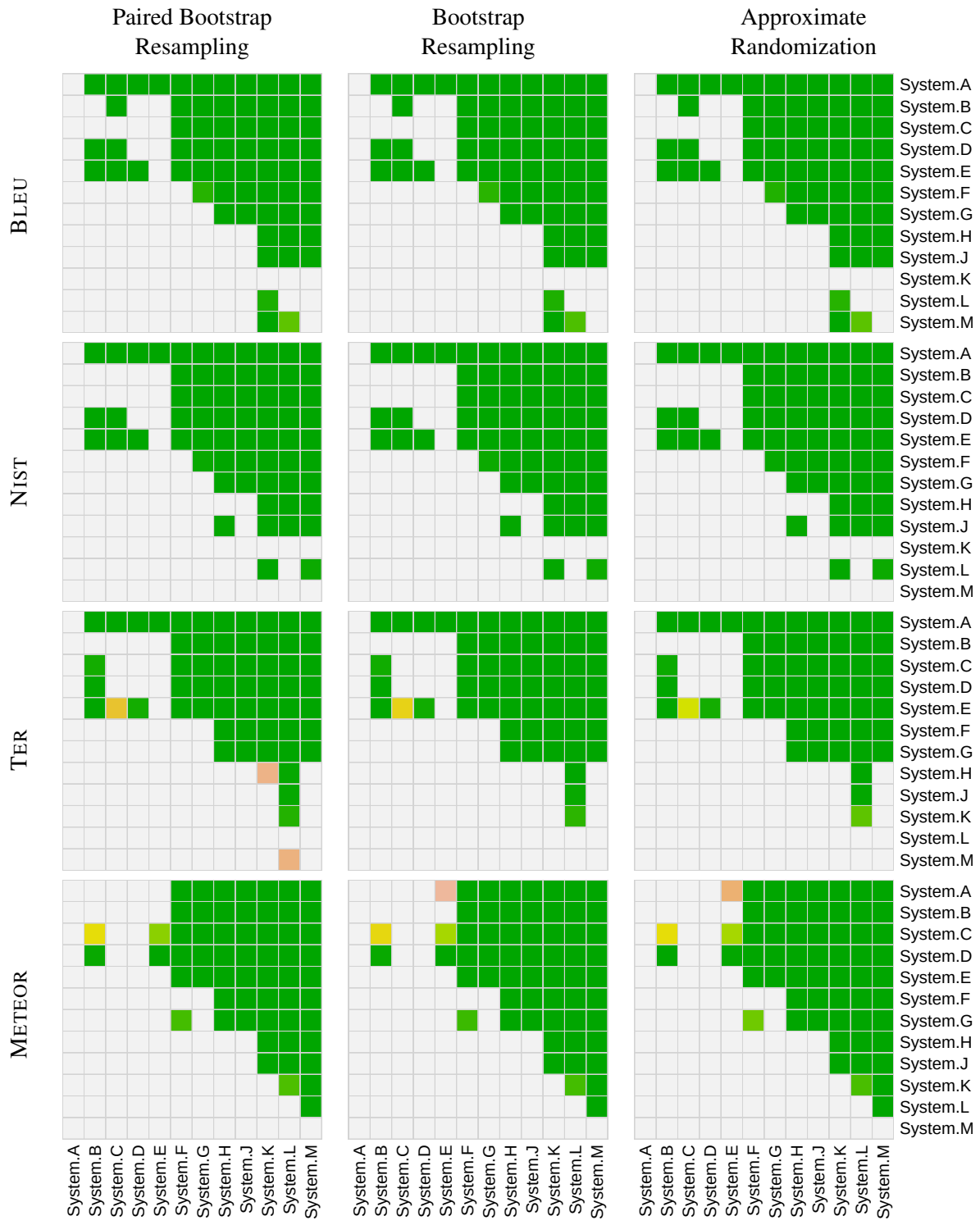

Figure 8: Automatic metric pairwise randomized significance test results for Spanish-to-English systems (colored cells denote scores for System *row* significantly greater than System *column*).

Finally, we examine which combination of metric and test is most accurate for each language pair at the conventional significance level of $p < 0.05$. For Spanish-to-English evaluation, NIST combined with any of the three randomized tests is most accurate, making 54 out of 66 (82%) correct conclusions. For English-to-Spanish, BLEU in combination with any of the three randomized tests, is most accurate at 62%. For both language pairs, however, differences in accuracy for metrics

are not significant (Chi-square test).

For English-to-Spanish evaluation, an accuracy as low as 62% should be a concern. This level of accuracy for significance testing – only making the correct conclusion in 6 out of 10 tests – acts as a reminder that no matter how sophisticated the significance test, it will never make up for flaws in an underlying metric. When we take into account the fact that lower confidence limits all fall below 50%, significance tests based on these metrics for English-to-Spanish are effectively no better than a random coin toss.

5 Conclusions

We provided a comparison of bootstrap resampling and approximate randomization significance tests for a range of automatic machine translation evaluation metrics. To provide a gold-standard against which to evaluate randomized tests, we carried out a large-scale human evaluation of all shared task participating systems for the Spanish-to-English and English-to-Spanish translation tasks from WMT 2012. Results showed for many metrics and significance levels that all three tests produce precisely the same set of conclusions, and when conclusions do differ, it is commonly only by a single contrasting conclusion, which is not significant. For English-to-Spanish MT, the results of the different MT evaluation metric/significance test combinations are not significantly higher than a random baseline.

Acknowledgements

We wish to thank the anonymous reviewers for their valuable comments. This research was supported by funding from the Australian Research Council.

References

S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for mt evaluation with improved correlation with human judgements. In *Proc. Wkshp. Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–73, Ann Arbor, MI. ACL.

C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proc. 7th Wkshp. Statistical Machine Translation*, pages 10–51, Montreal, Canada. ACL.

J. H. Clark, C. Dyer, A. Lavie, and N. A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proc. of the 49th Annual Meeting of the Assoc. Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 176–181, Portland, OR. ACL.

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

M. Galley and C. D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856, Edinburgh, Scotland. ACL.

U. Germann. 2003. Greedy decoding for statistical machine translation in almost linear time. In *Proc. of the 2003 Conference of the North American Chapter of the Assoc. Computational Linguistics on Human Language Technology-Volume 1*, pages 1–8, Edmonton, Canada. ACL.

Y. Graham, T. Baldwin, A. Moffat, and J. Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proc. 7th Linguistic Annotation Wkshp. & Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. ACL.

S. Green, M. Galley, and C. D. Manning. 2010. Improved models of distortion cost for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Assoc. Computational Linguistics*, pages 867–875, Los Angeles, CA. ACL.

P. Koehn and C. Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121, New York City, NY. ACL.

P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. of Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. ACL.

S. Kumar and W. J. Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *HLT-NAACL*, pages 169–176, Boston, MA. ACL.

C. Monz. 2011. Statistical machine translation with local language models. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 869–879, Edniburgh, Scotland. ACL.

NIST. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Technical report.

E. W. Noreen. 1989. *Computer intensive methods for testing hypotheses*. Wiley, New York City, NY.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. 41st Ann. Meeting of the Assoc. Computational Linguistics*, pages 160–167, Sapporo, Japan. ACL.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. A method for automatic evaluation of machine translation. In *Proc. 40th Ann. Meeting of the Assoc. Computational Linguistics*, pages 311–318, Philadelphia, PA. ACL.

S. Riezler and J. T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for mt. In *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, MI. ACL.

S. Riezler and J. T. Maxwell. 2006. Grammatical machine translation. In *Proc. of the Main Conference on Human Language Technology Conference of the North American Chapter of the Assoc. Computational Linguistics*, pages 248–255, New York City, NY. ACL.

M. Smucker, J. Allan, and B. Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM 2007)*, pages 623–632, Lisbon, Portugal. ACM.

M. Snover, B. Dorr, R. Scwartz, J. Makhoul, and L. Micciula. 2006. A study of translation error rate with targeted human annotation. In *Proc. 7th Biennial Conf. of the Assoc. Machine Translaiton in the Americas*, pages 223–231, Boston, MA. ACL.