BOOTSTRAP ESTIMATES FOR CONFIDENCE INTERVALS IN ASR PERFORMANCE EVALUATION

M. Bisani and H. Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology, D-52056 Aachen, Germany
{bisani,ney}@informatik.rwth-aachen.de

ABSTRACT

The field of speech recognition has clearly benefited from precisely defined testing conditions and objective performance measures such as word error rate. In the development and evaluations of new methods, the question arises whether the empirically observed difference in performance is due to a genuine advantage of one system over the other, or just an effect of chance. However still many publications do not concern themselves with the statistical significance of the results reported. In this paper we present a *bootstrap* method for significance analysis which is at the same time intuitive, precise and and easy to use. Unlike some methods, we make no (possibly ill-founded) approximations and the results are immediately interpretable in terms of word error rate.

1. INTRODUCTION

The most popular performance measure in automatic speech recognition is the word error rate

$$W := \frac{\sum_i e_i}{\sum_i n_i} \tag{1}$$

where n_i is the number of word in sentence i and e_i is the edit distance between the recognizer output and the reference transcription of sentence i. The edit distance, or Levenshtein distance, is the minimum number of insert, substitute and delete operations necessary to transform one sentence into the other. It can be efficiently calculated using dynamic programming algorithms. The word error rate is an attractive metric, because it is intuitive, it corresponds well with application scenarios and (unlike sentence error rate) it is sensitive to small changes. On the downside it is not very amenable to statistical analysis. W is really a rate (number of errors per spoken word), and not a probability (chance of misrecognizing a word). (It can exceed 100% due to insertions.) Moreover, error events do not occur independently.

The need for significance tests in ASR evaluations has been recognized long ago [1]. During DARPA evaluations not less than four different significance tests have been routinely conducted [2]. Nevertheless hardly any other publication reports figures on these significance tests. Instead it is common to report the absolute and relative change in word error rate only.

2. MOTIVATION

The question of how certain we can be of an observed word error rate arises in two contexts:

- What error rate do we have to expect when changing to a different test set?
- How reliable is a observed improvement of a system?

We will consider these problem in sections 3 and 5 respectively.

Since it is not possible (or at least hard) to apply classical methods of statistics to the word error rate, it has been proposed to resort to the sentence level [1][3]. We think that this is not a good alternative, since any serious large vocabulary continuous speech recognition task, will show sentence error rates very close to 100% (at least with present day technology). Using the number of errors per sentence (NES) as proposed in [3] does not seem very attractive either, since this measure depends on the distribution of sentence lengths. When moving to test corpus with (on average) longer sentence, NES increases although general recognition accuracy stays the same.

The method we explain in the remainder of this paper, provides accurate confidence intervals for the word error rate as it is widely established, at the expense of a little computational effort.

3. BOOTSTRAP

The bootstrap is a computer-based method for assigning measures of accuracy to statistical estimates [4]. The core idea is to create replications of a statistic by random sampling from the data set with replacement (so-called Monte Carlo estimates).

We assume that the test corpus can be divided into s segments for which the recognition result is independent and the number of errors can thus be evaluated independently. In continuous, speaker-dependent, speech recognition these will typically be sentences or dialog turns: Each of them is presented to the recognition system individually and the result is independent from all other segments. It is well known, however, that recognition performance varies strongly across speakers. Moreover, when speaker adaptation methods are used, it is not possible to evaluate sentences individually, since the recognition result also depends on all (previous) sentences. Thus, for speaker-independent CSR it seems appropriate to to choose the set of all utterances of one speaker as a segment. Our example results shown below support this view. Nev-

ertheless, for simplicity we use the term sentence interchangeably with segment in the following.

For each sentence i we record the number of words n_i and the number errors e_i:

$$X = (n_1, e_1), \ldots, (n_s, e_s) \tag{2}$$

Evaluating equation (1) now yields the maximum-likelihood estimate of the WER.

The following procedure is repeated B times (typically $B = 10^3 \ldots 10^4$): For $b = 1 \ldots B$ we randomly select with replacement s pairs from X, to generate a *bootstrap sample*

$$X^{*b} = (n_1^{*b}, e_1^{*b}), \ldots, (n_s^{*b}, e_s^{*b}) \tag{3}$$

The sample will contain several of the original sentences multiple times, while others are missing. Then we calculate the word error rate on this sample

$$W^{*b} := \frac{\sum_{i=1}^{s} e_i^{*b}}{\sum_{i=1}^{s} n_i^{*b}} \tag{4}$$

The W^{*b} are called *bootstrap replications* of W. They can be thought of as samples of the word error rate from an ensemble of virtual test sets. This can be visualized in the form of a histogram (see fig. 1). The bootstrap estimate of word error rate is

$$W_{\text{boot}} := \langle W^* \rangle \approx \frac{1}{B} \sum_{b=1}^{B} W^{*b} \tag{5}$$

One may be tempted to think that necessarily $W_{\text{boot}} = W$, but this is not the case.[1] Reassuringly, in practice we have found the difference ("bias") to be negligibly small.

The uncertainty of W_{boot} can be quantified by the standard error, which has the following bootstrap estimate:

$$\begin{aligned} \text{se}_{\text{boot}}(W) &:= \text{se}(W^*) \\ &= \sqrt{\left\langle (W^* - \langle W^* \rangle)^2 \right\rangle} \\ &\approx \sqrt{\frac{\sum_{b=1}^{B} (W^{*b} - W_{\text{boot}})^2}{B-1}} \end{aligned} \tag{6}$$

For large s the distribution of W^* is approximately Gaussian. In this case the true word error rate lies with 90% probability in the interval $W_{\text{boot}} \pm 1.64\, \text{se}_{\text{boot}}(W)$.

Even when s is small, we can use the table of replications W^{*b} to determine percentiles which in turn can serve as confidence intervals. For a chosen error threshold α, let $W_{\text{boot}}^{-\alpha}$ be the αB-th smallest value in the list $W^{*1} \ldots W^{*B}$, and $W_{\text{boot}}^{+\alpha}$ be the αB-th largest. [2] The interval

$$C_{\text{boot}}(W, \alpha) := (W_{\text{boot}}^{-\alpha}, W_{\text{boot}}^{+\alpha}) \tag{7}$$

contains the true value of W with probability $1 - 2\alpha$. This is the *bootstrap-t* confidence interval. More sophisticated confidence intervals are discussed in [4].

[1] It is easy to come up with an artificial counter-example.

[2] For example: With $B = 1000$ and $\alpha = 0.05$, we sort the list of W^{*b} and use the values at position 50 and 950.

4. EXAMPLE: NAB ERROR RATES

We used our baseline "Wall Street Journal" speech recognizer [5] to process three different test sets from the 1994 and 1995 ARPA North American Business News (NAB) CSR benchmarks. The recognition vocabulary contained about 20.000 words. All result presented in table 1 were obtained with identical settings. The dev'94 results are also shown in figure 1. The results from all three test sets agree on the 90% level. The accuracy of the recognition system varies greatly among different speakers. (It is not uncommon for the "best" speaker to achieve an error rate several times lower than the "worst" one's.) Therefore we have applied the bootstrap method on the speaker level as well as on the sentence level: the sentence-wise bootstrap produces considerably smaller standard errors and narrower confidence intervals. Due their false assumption of independence the sentence-level confidence intervals are too optimistic and should not be trusted.

Table 1. Word error rates of a "Wall Street Journal" dictation system on three different test sets of the NAB task with confidence estimates: 90% confidence intervals based on standard error and on bootstrap-t with $B = 10^4$.

Test Set		dev'94	eval'94	dev'95
words		7397	8347	7361
W	[%]	11.56	12.75	11.98
		sentence-wise bootstrap		
sentences		310	316	309
W_{boot}	[%]	11.56	12.76	11.98
$1.64\ \text{se}_{\text{boot}}(W)$	[%]	1.27	1.13	1.09
$W_{\text{boot}}^{-0.05}$	[%]	10.33	11.66	10.90
$W_{\text{boot}}^{+0.05}$	[%]	12.88	13.90	13.09
		speaker-wise bootstrap		
speakers		20	20	20
W_{boot}	[%]	11.53	12.75	11.98
$1.64\ \text{se}_{\text{boot}}(W)$	[%]	2.86	2.44	2.13
$W_{\text{boot}}^{-0.05}$	[%]	9.08	10.45	9.97
$W_{\text{boot}}^{+0.05}$	[%]	14.71	15.31	14.19

5. COMPARING SYSTEMS

We note that confidence intervals found in the examples shown above are quite wide. The large majority of system improvements reported in the literature are not much bigger than $\text{se}(W)$. Nevertheless it often turns out that they are consistent over different scenarios and are more or less additive in combination with other improvements. As has already been pointed out in [1], it is important to take into account, that competing algorithms are usually tested on the *same* data. Estimates of the error of W, such as (6), however, are appropriate for *independent* test sets.

While the result on each sentence is independent from the others, the results of two systems on the *same* sentence is strongly correlated. Typically a small modification to a system will alter the recognition results in a few sentences only. Intuitively, we would be quite confident that an improvement is genuine if the number

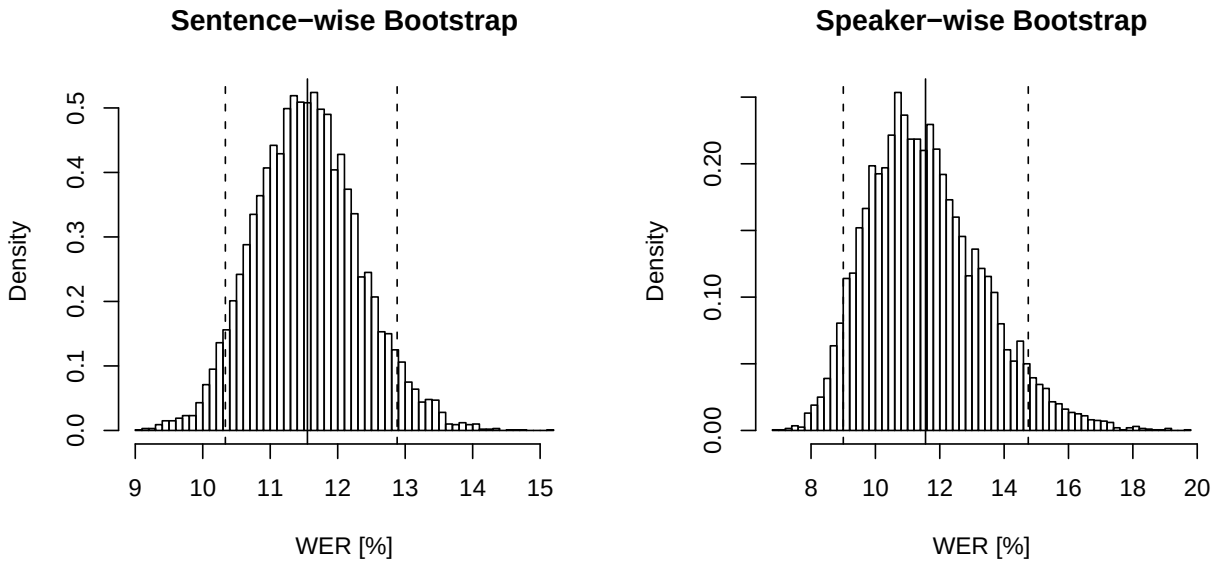

Fig. 1. Histogram of bootstrap replications of the word error rate on the NAB dev'94 test set). The solid line is the bootstrap estimate of the word error rate W_{boot}, the dashed lines mark the 90% confidence interval $C_{\mathrm{boot}}(0.05)$.

of errors drops on 5% of the sentences while the others remain unchanged, but if 50% of the sentences improved while 45% degraded, the overall improvement could very well be random. We can easily extend the proposed bootstrap procedure to this situation and obtain relatively tight confidence intervals for observed word error rate changes.

Given two recognition systems A and B with word error counts e_i^A and e_i^B, the (absolute) difference in word error rate is

$$\Delta W := W^A - W^B = \frac{\sum_i (e_i^A - e_i^B)}{\sum_i n_i} \tag{8}$$

We can apply the same bootstrap technique to the quantity ΔW as we did to W. The crucial point is that we calculate the difference in the number of errors of the two systems on *identical* bootstrap samples. Given the aforementioned correlation between the results of the two systems, this has the important consequence that ΔW^* has much lower variance than W^* of either system. (See figure 2 for an illustration.) In addition to the two-tailed confidence interval $C_{\mathrm{boot}}(\Delta W)$, we may be more interested in whether system B is a real improvement over system A. We propose to use the bootstrap estimate of the probability of error reduction for this purpose:

$$\begin{aligned} \mathrm{poi}_{\mathrm{boot}} &:= Pr(\Delta W^* < 0) \\ &= \langle \Theta(-\Delta W^*) \rangle \\ &\approx \frac{1}{B} \sum_{b=1}^{B} \Theta(-\Delta W^{*b}) \end{aligned} \tag{9}$$

where $\Theta(x)$ is the step function, which is one for $x > 0$ and zero otherwise. So (9) is the relative number of bootstrap samples which favor system B. We call this measure "probability of improvement" (poi).

6. EXAMPLE: SYSTEM COMPARISON

The system used for the examples described in section 4 now plays the role of system B, while a second system with slightly different acoustic models is system A. The results for this scenario are given in table 2. System B is apparently better by 0.3% to 0.4% absolute in terms of word error rate. The probability of improvement ranging between 82% and 95%, indicates that we can be be moderately confident that this reflects a real superiority of system B, but we should not be too surprised if a fourth test set would be favorable to system A. We also note that, unlike in the case of absolute error rates, the speaker-level standard errors are comparable to the sentence-level ones. This indicates that the improvement tends to be consistent across speakers.

The notable advantage of this differential analysis is that the standard error of ΔW is approximately one third of the standard error of W. Considering that one has to use the root sum of the standard errors, when independent random variables are concerned, we see that a four times higher difference in word error rate would be necessary to achieve a similar level of significance, if the tests were done on independent test sets.

7. CONCLUSION

We have demonstrated how the bootstrap method can be applied to finding confidence intervals on word error rate in ASR evaluations. We would like to emphasize that what we propose is *not* a new metric for performance evaluation, but a refined analysis of an established metric (word error rate). The method presented in this article is by no means limited to speech recognition or word error rate. It can be applied to other metrics and other NLP or pattern classification tasks where individual decisions are not independent (e.g. machine translation). The proposed method seems attractive, because it is easy to use, it makes no assumption about the dis-

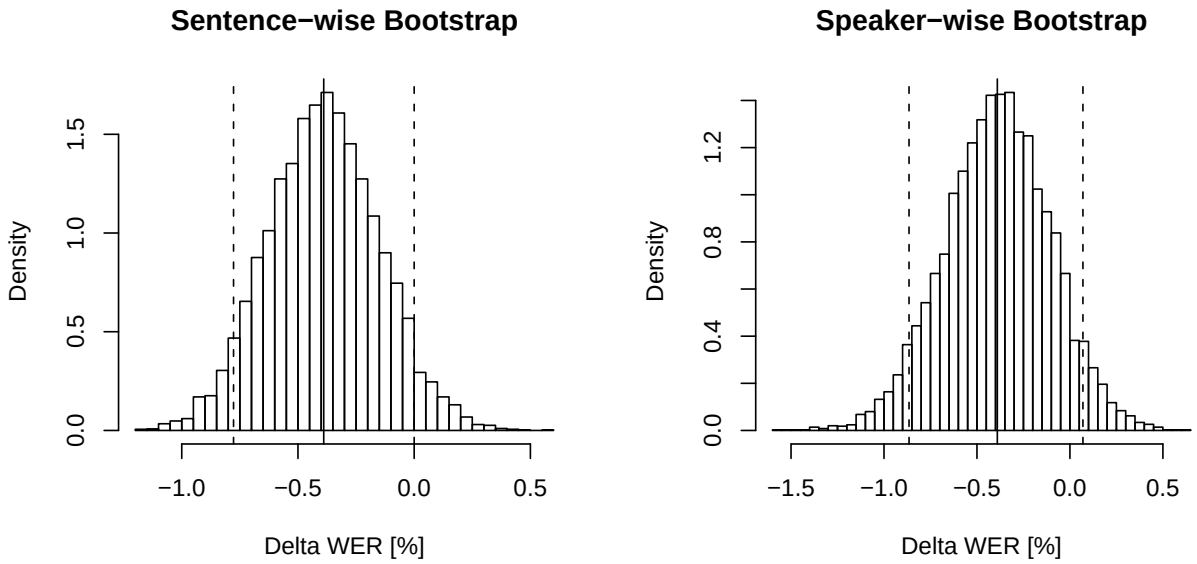

Fig. 2. Histogram of bootstrap replications of the word error rate difference ΔW^* on the NAB dev'94 test set. The system corresponding to figure 1 and table 1 plays the role of system B here, while system A uses a different acoustic model. Apparently system B is better by 0.4% absolute in terms of word error rate (solid line). The dashed lines mark the 90% bootstrap-t confidence interval.

tribution of errors, results are directly related to word error rate, and the "probability of improvement" provides an intuitive figure of significance. Furthermore it can easily take into account the variation of performance across speakers. We have demonstrated that failing to do so may lead to underestimating the error of a performance metric. Future experiments will show if the proposed confidence measures prove useful on the long run.

8. REFERENCES

[1] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Glasgow, Scotland, May 1989, pp. 532 – 535.

[2] D. Pallett, J. Fiscus, W. Fisher, and J. Garofolo, "Benchmark tests for the DARPA spoken language program," in *Proc. of the 1993 ARPA Human Language Technology Workshop*, Plainsboro (NJ), Mar. 1993, pp. 7 – 18.

[3] H. Strik, C. Cucchiarini, and J. M. Kessens, "Comparing the recognition performance of CSRs: In search of an adequate metric and statistical significance test," in *Proc. Int. Conf. Spoken Language Processing*, Beijing, China, Oct. 2000, vol. IV, pp. 740 – 743.

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

[5] A. Sixtus and H. Ney, "From within-word model search to across-word model search in large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 16, no. 2, pp. 245 – 271, May 2002.

Table 2. Differential word error rates of two "Wall Street Journal" dictation systems (A and B) on three different test sets of the NAB task with confidence estimates: 90% confidence intervals based on standard error and on bootstrap-t with $B = 10^4$ and probability of improvement.

Test Set		dev'94	eval'94	dev'95
words		7397	8347	7361
ΔW	[%]	-0.39	-0.30	-0.31
		sentence-wise bootstrap		
sentences		310	316	309
ΔW_{boot}	[%]	-0.39	-0.30	-0.30
$1.64\ \mathrm{se}_{\mathrm{boot}}(\Delta W)$	[%]	0.40	0.35	0.48
$\Delta W^{-0.05}_{\mathrm{boot}}$	[%]	-0.79	-0.65	-0.80
$\Delta W^{+0.05}_{\mathrm{boot}}$	[%]	0.00	0.04	0.17
$\mathrm{poi}_{\mathrm{boot}}$	[%]	94.5	91.4	84.5
		speaker-wise bootstrap		
speakers		20	20	20
ΔW_{boot}	[%]	-0.39	-0.30	-0.31
$1.64\ \mathrm{se}_{\mathrm{boot}}(\Delta W)$	[%]	0.47	0.35	0.54
$\Delta W^{-0.05}_{\mathrm{boot}}$	[%]	-0.87	-0.66	-0.88
$\Delta W^{+0.05}_{\mathrm{boot}}$	[%]	0.07	0.05	0.21
$\mathrm{poi}_{\mathrm{boot}}$	[%]	91.3	92.0	81.8