Comparing the performance of two CSRs:
How to determine the significance level of the differences

Helmer Strik, Catia Cucchiarini, Judith M. Kessens

A^2RT, Dept. of Language and Speech, Univ. of Nijmegen, the Netherlands
{Strik, C.Cucchiarini, J.Kessens}@let.kun.nl
http://lands.let.kun.nl/

Abstract

When two CSRs are compared, it is important to test what the significance level of the difference is. For this purpose a metric and a statistical test are needed. In this paper we compare several combinations of a metric with a statistical test, in order to find a combination which is suitable for this task. Four combinations which are introduced in this paper appear to be suitable for this task.

1. Introduction

In many papers about the development of CSR systems, the performance of two CSRs is compared. This kind of evaluation is known as performance evaluation [4]. Needed are a criterion, a measure and a method [4]. The criterion usually is error rate (or recognition accuracy), the measure is word error rate (or word accuracy), and the method is a DP alignment of what has been recognized with 'the truth' (a transcription).

Already in 1989, Gillick and Cox [2] wrote: "In the development of speech recognition algorithms, it is important to know whether any apparent difference in performance of algorithms is statistically significant, yet this issue is almost always overlooked." Although it is important, testing for statistical significance is still rarely done. Usually only WERs and their differences are reported. Measures of statistical significance would be welcome, because otherwise it remains unclear how much importance should be attached to a difference in WERs. Moreover, this would make it easier to compare (the significance of) the results between experiments, which usually is quite difficult [6]. The ultimate goal of our research is to stimulate the use of statistical significance tests when comparing the performance of CSRs.

In order to determine the significance level of the difference in performance between two CSRs, a metric and a statistical test are needed. Since usually, the two CSRs are tested on the same test corpus, we will focus on this task. The goal of the research presented in this paper, is to find a suitable combination of a metric and a statistical test for this task.

This paper is organized as follows. In section 2 we define the six metrics tested in our research. In section 3 we compare various combination of metrics and statistical tests. The results are discussed in section 4, while conclusions are drawn in section 5.

2. Defining the metrics

The general procedure used to determine the performance of a CSR is the following. Each CSR is first trained and then tested with a test corpus. This test corpus has W words, S sentences, and W(s) is the number of words per sentence s.

The output of the test corpus is a string of recognized words (RECOG). RECOG is compared to what was actually spoken (SPOKEN). This is done by means of a DP alignment of SPOKEN with RECOG at the word level. The DP alignment reveals the differences between the two strings, and assigns to each of the words in RECOG one of the four following labels: correct, substitution (sub), deletion (del), or insertion (ins). Usually, only the number of substitutions, deletions and insertions for the whole test corpus are calculated, because they are needed to calculate WER. However, it is also possible to determine these numbers for every sentence s: sub(s), del(s) and ins(s). Then WER is:

$$\text{WER} = \sum_s \{\text{sub(s)+ del(s)+ ins(s)}\} \ / \ \text{W}$$

As is clear from the formula above, WER is a global measure of accuracy that expresses the proportion of words that have been recognized correctly (for the whole test corpus). However, this is not the only measure of accuracy that can be computed at word level. Above we already mentioned that, when WER is calculated by comparing SPOKEN and RECOG, the DP alignment determines for every word in RECOG whether it is recognized correctly or not. Generally, this information is only used to calculate WER. However, it can also be used to derive another word level metric which we call Word Error (WE). WE is a Boolean that can have two values: a value of 1 indicates an error (sub, del or ins), and a value of 0 is assigned if the word was recognized correctly.

WER and WE are two measures at the word level. There are similar metrics at the sentence level: SER (Sentence Error Rate) and SE (Sentence Error). Like WE, SE can have two values: 1 when there are (1 or more) errors in the sentence, and 0 when there are 0 errors (i.e. the whole sentence has been recognized completely correct). In short:

$$\text{SE(s)} = \text{signum}\{\text{sub(s)+ del(s)+ ins(s)}\}$$

SER can simply be calculated from SE:

$$\text{SER} = \sum_s \text{SE(s)} \ / \ \text{S}$$

In [7] we introduced a new metric NES (Number of Errors per Sentence):

$$\text{NES(s)} = \text{sub(s)} + \text{del(s)} + \text{ins(s)}$$

Here, we introduce another new sentence level metric called WES (Word Error rate per Sentence). For a sentence s:

$$\mathrm{WES}(s) = \mathrm{NES}(s) \,/\, \mathrm{W}(s)$$

The relations between these new metrics (NES and WES) and the 'old' metric WER is:

$$\mathrm{WER} = \textstyle\sum_s \mathrm{NES}(s) \,/\, \mathrm{W} = \sum_s \mathrm{W}(s)*\mathrm{WES}(s) \,/\, \mathrm{W}$$

WER thus is a weighted average of WES(s), with weighting coefficients W(s). Furthermore, WER is related to the average NES(s):

$$<\mathrm{NES}(s)> = \textstyle\sum_s \mathrm{NES}(s) \,/\, \mathrm{S} = \mathrm{WER} * \mathrm{W} \,/\, \mathrm{S}$$

i.e. they only differ by a constant factor. Finally, SER can also be expressed in terms of NES and WES:

$$\mathrm{SER} = \textstyle\sum_s \mathrm{signum}\{\mathrm{NES}(s)\} \,/\, \mathrm{S} = \sum_s \mathrm{signum}\{\mathrm{WES}(s)\} \,/\, \mathrm{S}$$

3. Comparing the metrics

Six metrics were defined in section 2: WER, WE, SER, SE, NES and WES. The question we try to answer is: Which one is most suitable for significance testing?

WER and SER are both single numbers that describe the performance of a CSR (for a given test corpus). In [7] we concluded that these two metrics are less suitable to test statistical significance. In [7] we also made clear that WE is not suitable for significance testing, mainly for the following two reasons: (1) the errors are not independent, and (2) insertions are problematic for many tests (see [7]).

What remains are the sentence level metrics SE, NES, and WES. These three metrics are compared below. First for artificial data in section 3.1, and then for real data in section 3.2.

3.1. Artificial data

First of all, it should be noted that SE is not a detailed measure and is therefore little informative about the differences in performance between two CSRs. While a zero value means that the sentence in question does not contain any error, a value of one can mean a lot of different things, varying from one error through all gradations up to all words recognized incorrectly. This clearly appears from the examples presented in Table 1. NES and WES, on the other hand, are in line with our intuitions that, e.g., in a sentence containing 2 errors recognition accuracy is higher than in a sentence containing 6 errors, whereas this difference would not be revealed by SE (see Table 1).

Let us now compare CSR1 with CSR2 for the examples given in Table 1. For sentence 1 we observe an improvement from 2 to 1 recognition errors. Since errors are present for both CSRs, SE is 1 for both sentences, and thus ΔSE is 0. NES does decrease from 2 for CSR1 to 1 for CSR2, and ΔNES is 1. Similarly, WES decreases from 20% to 10%, and ΔWES = 10%. In other words, the differences in recognition results for sentence 1 are reflected in ΔNES and ΔWES, but not in ΔSE. Analogously, the changes for sentences 2 and 3 are 'noticed' by ΔNES and ΔWES, but not by ΔSE. Not only do ΔNES and ΔWES reflect the differences for the sentences 1 to 3, they also reveal that the magnitude of the improvements increases when going from sentence 1 to 3. In short, NES and WES do measure some differences that are not measured by SE.

In the sentences 1, 2 and 3 NES and WES show a similar pattern, mainly because the number of words for these three sentences is the same. For sentences 4, 5 and 6 the number of words is twice as large, while the number of errors remain the same as those in sentences 1, 2 and 3, respectively. Therefore, the SE and NES values remain exactly the same, but the WES values are different. Let us take a closer look at the differences. In sentences 1 and 4 an improvement from 2 to 1 errors is observed. For both sentences ΔSE is 0 and ΔNES is 1. However, ΔWES is 10% for sentence 1 and 5% for sentence 4. Similarly, the ΔWES values for the sentence pairs 2-5 and 3-6 are 30%-15% and 50%-25%, respectively. The observed values of ΔWES are more in line with our intuition than those of ΔNES and ΔSE. After all, an improvement from 2 to 1 errors for a sentence of 10 words is better than the same improvement for a sentence of 20 words. This is exactly the reason why the ASR community uses WER to compare the performances of 2 CSRs. Also in that case, the decrease in WER depends both on the number of improvements and the size of the corpus. An example to clarify this. Given the same number of improvements (say M) for a first corpus of size N1 and a second, larger corpus of size N2 (N2 > N1); then

$$\Delta\mathrm{WER1} = \mathrm{M/N1} \;>\; \Delta\mathrm{WER2} = \mathrm{M/N2}.$$

To summarize, NES and WES seem to be more appropriate metrics than SE, because they are more detailed and more in line with our intuitions.

Table 1. Artificial recognition results.

Sentence s	**1**	**2**	**3**		**4**	**5**	**6**		**7**
W	10	10	10		20	20	20		10
CSR1									
sub	1	2	2		1	2	2		1
del	1	1	2		1	1	2		0
ins	0	1	2		0	1	2		0
NES	2	4	6		2	4	6		1
WES (%)	20	40	60		10	20	30		10
SE	1	1	1		1	1	1		1
CSR2									
sub	1	1	1		1	1	1		0
del	0	0	0		0	0	0		1
ins	0	0	0		0	0	0		0
NES	1	1	1		1	1	1		1
WES (%)	10	10	10		5	5	5		10
SE	1	1	1		1	1	1		1
CSR1-CSR2									
ΔNES	1	3	5		1	3	5		0
ΔWES (%)	10	30	50		5	15	25		0
ΔSE	0	0	0		0	0	0		0

Finally, it should be noted that there are some changes that are not even reflected in ΔNES and ΔWES, e.g. like those for sentence 7. As a matter of fact, we can see that NES and WES make no distinction between substitutions, deletions and insertions. However, this distinction is not made by SE either and, consequently, the changes in sentence 7 are not reflected in ΔSE either.

3.2. Real data

In the previous section we used artificial data to compare the metrics. Here we will use real data from past experiments. Furthermore, we will not restrict ourselves to comparing metrics alone (as in section 3.1). Instead, we will compare several combinations of metrics and statistical significance tests. After all, in order to determine statistical significance, one needs a metric and a test, and our goal is to find the most suitable combination. We will present results for the 11 combinations given in Table 2. The first 2 combinations were mainly included because they have been used in the past (see [1,3] and [2,3,5], respectively). The p value for combination 1 was calculated by means of the formula given in [1]. All other p values were calculated with the statistical software package SPSS 10.0.

Table 2. Combinations of metrics and statistical tests.

nr.	metric	statistical test	p (%)
1	WER	confidence intervals	5.8
2	SE	McNemar	11.3
3	SE	Signed Pair	11.3
4	SE	WSR	10.2
5	SE	T test	10.2
6	NES	Signed Pair	2.9
7	NES	WSR	0.9
8	NES	T test	0.1
9	WES	Signed Pair	2.9
10	WES	WSR	0.1
11	WES	T test	0.1

We first compared two CSRs for a test corpus of 5000 sentences. In Table 3 we present some descriptive statistics, i.e. total number of sentences and words in the test set, SER and WER for the two CSRs, and the difference in the WERs: ΔWER = WER1 – WER2 = 0.98%. These are the kind of numbers that are usually provided when two CSRs are compared. However, just mentioning WERs of the two CSRs and ΔWER is not sufficient, since then it is not known what the importance of this difference is. Furthermore, how should one compare such a ΔWER of 0.98% to another ΔWER of, e.g., 0.95%. This remains unclear. Therefore, the significance level of the differences of two CSRs should be calculated and mentioned in papers. The significance levels of the differences between the two CSRs was determined for the 11 metric-test combinations. The results are shown in Table 2.

Table 3. Descriptive statistics.

S	5000
W	16357
SER1	26.54%
SER2	25.92%
WER1	15.64%
WER2	14.67%
ΔWER	0.98%

It can be observed that for none of the first five combinations the difference is significant at the 5% level. The lowest p values were found for the metrics NES and WES: in combination with the Signed Pair test the p values are lower than 3%, and in combination with Wilcoxon Signed Rank (WSR) and T test the p values are even smaller ($p < 1\%$). Let us try to understand these results by taking a closer look at the data, the metrics and the statistical tests.

The significance levels for SE (combinations 3-5) are much larger than the significance levels for the other two sentence level metrics (combinations 6-11). This large difference can be explained by examining the data: there are a lot of sentences for which the number of errors is reduced but does not become zero. Consequently, for these sentences SE remains one, while NES and WES are reduced (as was the case for sentences 1 to 3 in Table 1). SE finds 195 improvements and 164 deteriorations, while NES and WES find many more: 345 improvements and 289 deteriorations. This corroborates our findings of section 3.1. Finally, the difference between McNemar and Signed Pair test on the one hand, and WSR and T test on the other, is a reflection of the fact that the last two statistical tests are more powerful than the first two.

From the results above, the picture emerges that combinations 7, 8, 10 and 11 are more suitable than the other seven combinations. Since this picture is based on one comparison of two CSRs, we decided to study more cases. We selected 11 new comparisons of two CSRs: 5 more cases for the same test corpus of 5000 sentences, and 6 cases for a test corpus of 6276 sentences. For these 11 new cases, we calculated the significance levels for all 11 combinations in Table 2. Also for these 11 other cases the p values for the first five combinations were higher than the other ones, followed by those of the combinations 6 and 9, while the lowest p values were found for the combinations 7, 8, 10, and 11. This thus seems to be a general picture that emerges from the results of all 12 cases. However, none of these four combinations (7, 8, 10, and 11) gave the best results for all 12 cases. In other words, on the basis of the data of these 12 cases it cannot be decided which of these four combinations is the most suitable one.

For three pairs of combinations (i.e. 2-3, 4-5, and 6-9) the resulting significance levels are always identical (see Table 2). This can easily be explained. However, this will not be done here, because of space limitations, and because none of these 6 combinations are amongst the four most suitable ones.

4. Discussion

In this paper we have discussed several metrics that can be used to express recognition accuracy when comparing the performance of two CSRs. In particular, we have discussed these metrics with respect to their informative properties and their possibilities of being submitted to statistical significance tests. It turned out that many of these metrics are not really satisfactory, either because they are not informative enough, or because there is no suitable significance test. In [7] we already concluded that WER, SER and WE are not suitable for statistical significance testing. What remains are the three sentence level metrics: SE, NES and WES. These three metrics were studied in combination with three appropriate statistical tests: Signed Pair, WSR and T test. These nine metric-test combinations are the combinations 3-11 in Table 2. Two other combinations were also tested (i.e. combinations 1 and 2 in Table 2), mainly because they had been used for significance testing in previous studies. Of the resulting 11 combinations, four turned out to be most appropriate: NES and WES in combination with WSR and T tests. This finding is plausible, because NES and WES seem to be superior to SE (see section 3.1), and because WSR and T test are more powerful than the Signed Pair test. These four metric-test combinations thus seem to be suitable to determine statistical significance of the differences between two CSRs.

On the basis of the examined data (the 12 cases) it does not become clear which of these four combinations is most appropriate. Is it possible to decide whether one of these four combinations is most suitable for the task at hand? Let us first have another look at the statistical tests. The Signed Pair test is a nonparametric test that only looks at the direction of the change (the sign). WSR is a nonparametric test too, but the difference with the Signed Pair test is that in WSR the ranking is done on the basis of the direction and the magnitude of the change. Finally, the T test is a parametric test that also takes the magnitude of the difference into account. The nonparametric tests (Signed Pair and WSR) make no assumptions about the data. The T test, on the other hand, does make a number of assumptions about the data, e.g. (in short) interval level, normal distribution, and homoscedasticity (equality of variances). It is questionable whether all assumptions of the T test are met for the data under study. However, the T test is robust against violations of the assumptions. In other words, even if not all assumptions of the T test are met, it may still be applied. WSR and T test are more powerful than McNemar and Signed Pair test. The difference in power between WSR and T test is small. WSR has about 95% of the power of the T test, if all the assumptions of the T test are met. The small difference in power between WSR and T test could explain the small differences in significance levels observed for these two tests.

Besides small differences in significance levels for WSR and T test, we also observed small differences between NES and WES. In section 3.1 we illustrated the differences between the metrics NES and WES. These two metrics detect the same number of improvements and deteriorations, but the magnitudes of the changes are different for the two metrics. Both WSR and T test do take the magnitude of the change into account. Still, given that neither NES nor WES always showed the lowest significance levels for all 12 cases, the effect of the differences in magnitude on the resulting significance levels appears to be small.

To summarize, of the 11 combinations tested, four combinations yielded the lowest significance levels. These four combinations seem to be suitable for significance testing. At the moment we have no objective criteria to decide which of these four combinations is the best one. More research is needed to get a better understanding of this issue.

The T test is probably more well-known and easier to interpret than WSR. Furthermore, the T test is more powerful than WSR. Regarding the data, the data presented in section 3.1 indicate that WES is more in line with our intuition than NES. Furthermore, since in the ASR community WER is the metric which is used most often to describe the performance of a CSR, it seems logical to use a similar metric at sentence level (WES is the WER for a sentence) for statistical significance testing. On the other hand, the average NES and WER differ only by a constant factor (see section 2). Since the T test essentially compares averages, and since WER is used most often, the combination of NES and T test might be the best alternative.

On these grounds, one could have a slight preference for NES or WES in combination with the T test. However, since there are no objective grounds to vote in favor of one of the four, it is probably best to use all four of them, in a similar way as NIST uses four significance tests for the DARPA evaluations [see 5].

There seems to be no practical objection against using all four of them, since they can all be calculated quite easily. The DP algorithm that is used to calculate WER for a whole corpus, can be employed to calculate the two metrics at sentence level (see section 2). Subsequently, the resulting numbers can be fed into any statistical package that includes WSR and T test, to calculate the significance levels.

5. Conclusions

The goal of the research presented in this paper was to find a metric-test combination that is suitable for determining the statistical significance of the differences between two CSRs. Of the 11 tested combinations, four turned out to be most suitable: the metrics NES and WES in combination with WSR and T test. There is no conclusive evidence to determine which is the best one of these four. Therefore, we suggest to use all four of them in combination.

6. Acknowledgements

We would like to thank Toni Rietveld for his valuable comments on a previous version of this manuscript. The research by Judith M. Kessens was carried out within the framework of the Priority Programme Language and Speech Technology, sponsored by NWO (Dutch Organization for Scientific Research).

7. References

[1] Ferreiros, J., and Pardo, J.M. “Improving continuous speech recognition in Spanish by phone-class semicontinuous HMMs with pausing and multiple pronunciations”, *Speech Communication,* Vol. 29: 65-76, 1999.

[2] Gillick, L., and Cox, S.J. “Some statistical issues in the comparison of speech recognition algorithms”, *Proc. ICASSP:* 532-535, Glasgow, May 1989.

[3] Harborg, E., *Hidden Markov Models Applied to Automatic Speech Recognition*, Ph.D. thesis, Norwegian Institute of Technology, Trondheim, 1990.

[4] Hirschman, L., and Thompson, H.S. “Overview of evaluation in speech and natural language processing”, In: R. Cole et al. (eds.) “Survey of the State of the Art in Human Language Technology”, Cambridge University Press, 1997.

[5] Pallett, D., Fiscus, J., Fisher, W., and Garofolo, J. “Benchmark tests for the DARPA spoken language program”, *Proc. of the 1993 ARPA workshop*: 7-18, 1993.

[6] Strik, H., and Cucchiarini, C. “Modeling pronunciation variation for ASR: A survey of the literature”, *Speech Communication,* Vol. 29: 225-246, 1999.

[7] Strik, H., Cucchiarini, C., Kessens, J.M. “Comparing the recognition performance of CSRs: in search of an adequate metric and statistical significance test”, *Proc. ICSLP-2000, Beijing, 2000, pp. 740-743.*