

# Appendix A

## Sign-Test

Today, improvements due to any one system component in speech recognition based on state-of-the-art systems are usually small. Unless we have systematic evidence of differences in a consistent direction between model A and model B, we cannot claim model A is significantly better (or worse) than model B, even though the results of A may be slightly better than those of B, because this marginal difference may come from randomness. The sign test is used by National Institute of Standards and Technology (NIST) to evaluate the significance of the difference between speech recognition systems.

The principle of the sign test is quite simple. To check whether model A is significantly better than model B, we compare, for each test utterance, the number of word errors in the hypothesis generated by model A to the number in that produced by model B. We exclude the utterances for which both models produce the same number of errors, and then assign the utterances with fewer errors A with a positive sign (+) and the others with a negative sign (-). If there is no significant difference between the two models, the number of positive signs and that of negative ones should be the same or similar. On the other hand, if these two numbers differs considerably, the probability of random generation of this event is extremely small, so this result is most probably due to the fact that model A outperforms model B.

The probability of the event  $N$  out of  $M$  signs are positive is

$$\mathcal{B}(M, N, 0.5) = \binom{M}{N} 0.5^N (1 - 0.5)^{M-N}$$

if these two models have no difference in performance (*i.e.*, the null hypothesis holds). If

$$\sum_{n=N}^M \mathcal{B}(M, n, 0.5) < p$$

where  $p$  is a small positive number close to zero, then the null hypothesis is rejected, and therefore A is significantly better than B. NIST selects the significance level of  $p = 0.05$ , whereas for this dissertation, we select a more strict one of  $p = 0.01$ .

Table A.1 illustrates the difference in word error rate between model pairs compared in the Switchboard experiments, and the corresponding p-value  $\sum_{n=N}^M \mathcal{B}(M, n, 0.5)$ . The first column in the table indicates the sections in which the corresponding models are compared. Usually, a WER difference of 0.3% absolute is regarded as significant by NIST. It is apparent that all improvements acquired from non-local dependencies based on the baseline trigram models are significant. It is interesting that even though

Section	Model A	Model B	Diff. WER	P-value
4.3.1	BO 3-gram	ME 3-gram	-0.2%	0.0091
4.4.1	BO 3-gram	Interp. Topic	-0.4%	0.0022
4.4.2	BO 3-gram	Cache-based	0.4%	0.034
4.3.2	BO 3-gram	ME Topic	-0.7%	2.1e-6
5.5	BO 3-gram	ME Syntax	-1.0%	6.2e-4
6.3	BO 3-gram	ME Composite	-1.5%	6.6e-9
4.4.1	Interp. Topic	ME Topic	-0.3%	0.0011

Table A.1: Significance of WER difference between language models (SWBD). Negative values in Diff. WER indicate that Model B is better than Model A.

the WER difference of the trigram model and the cache-based model is 0.4%, the p-value of 0.034 is still higher than the standard of 0.01 used in this dissertation. This statistic may be due to the high WER in the cache, which causes high randomness in the cache-based model.

Section	Model A	Model B	Diff. WER	P-value
4.5	BO 3-gram (14M)	ME Topic (14M)	-0.7%	0.00042
5.8	BO 3-gram (14M)	ME Syntax (14M)	-0.7%	0.00087
6.5	BO 3-gram (14M)	ME Composite (14M)	-1.2%	8.1e-6
4.5	BO 3-gram (130M)	Interp. Topic (130M)	-0.6%	0.00012
6.5	BO 3-gram (130M)	ME Composite (14M)	-0.5%	0.00065

Table A.2: Significance of WER difference between language models (BN). Negative values in Diff. WER indicate that Model B is better than Model A.

Table A.2 shows the WER difference and the corresponding p-value for model pairs compared on Broadcast News. Again, the significance test results show that gains in WER are very significant for topic and syntactic models. It is worth noting that the composite model trained using 14M words is significantly better than the baseline trigram model trained using 130M words.