A PAIRED TEST FOR RECOGNIZER SELECTION WITH UNTRANSCRIBED DATA

Bhiksha Raj, Rita Singh and James Baker

Carnegie Mellon University, USA

ABSTRACT

Traditionally, the use of untranscribed speech has been restricted to the unsupervised or semisupervised training of acoustic models. On the other hand, comparison of recognizers has required labeled data. In this paper we show how a recognizers (and classifiers in general) may be rank-ordered in terms of their performance using only a large quantity of untranscribed (unlabeled) data, given a third "reference" recognizer (or classifier). We develop statistical tests for the comparison of recognizers in this scenario. Interestingly, the actual recognition accuracy of the reference recognizer is immaterial to the test, provided it is better than random. We also show, through detailed experiments, that the predictions based on untranscribed data are indeed valid and correct.

***Index Terms*—** Hypothesis testing, Untranscribed data, Unsupervised learning, Speech recognition.

1. INTRODUCTION

1.1. The case for untranscribed data

Speech recognition systems have traditionally been trained with transcribed data. Much of the cost of training a system properly is the cost of acquiring this data – collecting appropriate data and having them transcribed correctly is expensive. This has often been an impediment to the training of systems for new languages and domains.

Untranscribed data, however, are relatively cheap, and have in fact become readily available. Recent improvements in communication, recording, HCI and storage technologies, have resulted in an explosive growth in the amount of recorded and stored speech data. This provides a rich new source of data to train speech recognition systems. Much of this data is untranscribed, however.

This has led to explorations into techniques that can train speech recognition systems from large amounts of untranscribed data, such as unsupervised or semi-supervised training techniques *e.g.* [1, 2] and data selection methods for untranscribed training data [3].

1.2. Comparing recognition systems

However, *comparison* of trained systems still requires transcribed data. In other words, if were given a pair of systems and had to determine which of these may be expected to recognize speech better, we still require *transcribed* data to make this evaluation.

Typically, to compare two recognizers, their recognition performance on an evaluation set of transcribed speech recordings is compared. To complete the comparison however, statistical significance tests must be performed. The null hypothesis that the recognition accuracy of both systems is actually identical is evaluated using one of several possible methods. If we assume that the number of words hypothesized by the two systems is identical to, and has one to one correspondence with, the words in the actual transcription for the data, the Wilcoxon test [4] or McNemar's test [5] may be used. These tests evaluate the probability of the observed correspondences in the errors made by the two systems. In speech recognition systems, where the recognizer may also insert or delete words, a more appropriate test is the "Matched Pairs Sentence-Segment Word Error (MAPSSWE) Test" [6] which compares the recognition errors made by the two systems over entire segments of audio.

In either case, the test evaluates the hypothesis that the two recognition outputs are separate draws from the same random process. If this hypothesis can be rejected with sufficiently high probability, then it is assumed that the system that produced the better accuracy of the two is indeed better.

1.3. Comparison with *untranscribed* data

However, as mentioned earlier, these tests require transcribed data which is expensive to obtain. Untranscribed data are easier and cheaper to get. Would it then be possible to use *untranscribed* data to determine which of two recognizers performs better. That is, given a pair of recognizers and a corpus of untranscribed data, can we decide which of the two is better?

More generally, given two classifiers of unknown provenance, can we decide which of them is better, given only a collection of unlabeled data instances to evaluate them on?

The somewhat surprising answer is "yes", provided we have access to a third classifier! The only constraint on the third classifier is that it must perform better than random. It can even be less accurate than both classifiers being compared. By comparing the output of the classifiers being evaluated to that of the third classifier, it is possible to compute a probability for the truth of the null hypothesis that states the two classifiers are identical in performance. More generally, it is possible to determine which of two recognizers is likely to perform better on a given test data, given only their output, and that of a third recognizer, on a common set of untranscribed test data.

1.4. Contributions of this paper

In this paper we describe statistical tests that can be performed to compare two recognizers, given only unlabeled data and a third "reference" recognizer (or, more generally, a reference classifier).

We begin by showing analytically, for simple binary classifiers, how the comparison of their output to that of the reference classifier relates to the true underlying relation between the classifiers. We describe a simple statistical test to compare the classifiers when the true labels of the test data are not known, but are obtained from the reference classifier. We continue to show how McNemar's test can be generalized to the problem of statistical comparison from unlabeled data. Interestingly, the tests do not depend upon the actual accuracy of the reference classifier employed to obtain labels, provided it is better than random. Finally we also show how the tests generalize to multi-class classifiers.

To conclude, we show through a detailed set of experiments that the predictions made by the tests are indeed validated by performance observed on other test data.

2. COMPARING TWO CLASSIFIERS WITH A REFERENCE CLASSIFIER

2.1. Agreement between two binary classifiers

Consider two binary classifiers P and a reference classifier R. Let p be the probability of correct classification for P and r the probability of correct classification for R. Let A_{PR} represent the event that P and R agree on any classification. The probability t_{pr} of A_{PR} is given by $t_{pr} = pr + (1-p)(1-r)$. Note that $\partial a_{pr}/\partial p = 2r - 1$, *i.e.* for $r > 0.5$, a_{pr} increases linearly with p.

2.2. A simple unpaired test for binary classifiers

Now consider any instance that has been classified by binary classifiers P, Q and R, with probability of success p, q and r respectively. The probability that P and R will agree is given by $t_{pr} = pr + (1-p)(1-r)$. The probability that Q and R will agree is given by $t_{qr} = qr + (1-q)(1-r)$. Since $rx + (1-r)(1-x)$ is a monotonically increasing function of x for $r > 0.5$, $p > q \Leftrightarrow t_{pr} > t_{qr}$.

We can now follow the conventional formalism for proving statistical significance [5]. Let N_{PR} represent the number of instances on which P and R agree, and N_{QR} the number of instances on which Q and R agree. Let N represent the total number of evaluation instances. Assuming all test instances to be independently classified, the maximum likelihood estimates of t_{pr} and t_{qr} are given by Equation 1.

$$\hat{t}_{pr} = \frac{N_{PR}}{N}, \qquad \hat{t}_{qr} = \frac{N_{QR}}{N} \tag{1}$$

Let $N_{PR} > N_{QR}$. The null hypothesis that must be rejected in order to accept that $t_{pr} > t_{qr}$, and therefore that $p > q$ is that $t_{pr} = t_{qr}$, or equivalently that $d_{pq} = t_{pr} - tqr = 0$. The maximum likelihood estimate of d_{pq} is $\hat{d}_{pq} = \hat{t}_{pr} - \hat{t}_{qr}$. The variance of this estimate is given by $var\hat{d}_{pq} = var\hat{t}_{pr} + var\hat{t}_{qr}$. The mean of the distribution of a maximum likelihood estimate of the probability parameter x of a Bernoulli distribution is x. The variance of the estimator is $x(1-x)/N$. Thus, under the null hypothesis, the mean and variance of the estimator for d_{pq} is 0 and a good estimate for its variance is given by

$$\sigma^2_{pq} = \frac{2t_{pq}(1-t_{pq})}{N} \tag{2}$$

where $t_{pq} = 0.5(\hat{t}_{pr} + \hat{t}qr)$. Thus, for large N, under the null hypothesis $\tilde{d} = \hat{d}_{pq}/\sigma_{pq}$ (which can now be computed entirely from N_{PR}, N_{QR} and N) is a random variable drawn from $\mathcal{N}(x; 0, 1)$, a normal distribution with 0 mean and unit variance. In order to reject the hypothesis, we compute the two sided probability $P_{test} = 2\int_{|\tilde{d}|}^{\infty} \mathcal{N}(x; 0, 1)dx$ that the absolute value of a randomly drawn value can be equal to or greater than $\tilde{d}$. If P_{test} is smaller than a rejection threshold (typically 0.05, or 0.01), then the null hypothesis is rejected and it is assumed that $p > q$, *i.e.* that P is a more accurate classifier than Q.

Note that the above procedure does not require knowledge of the accuracy r of the reference classifier. In fact, r can be lesser than p or q. All that is required is knowledge of the number of instances where P and Q agreed with R respectively.

		PQ			
		TT	TF	FT	FF
R	T	pqr	$p(1-q)r$	$(1-p)qr$	$(1-p)(1-q)r$
	F	$pq(1-r)$	$p(1-q)(1-r)$	$(1-p)q(1-r)$	$(1-p)(1-q)(1-r)$

Table 1. Possible outcomes and their probabilities. "T" represents correct classification; "F" represents misclassification. The columns represent outcomes of P and Q. *e.g.* the column "TT" represents events in which both P and Q classify an instance correctly. Rows represent outcomes of R. The table entries represent probabilities, *e.g.* the probability that P, Q and R will all classify correctly is pqr.

2.3. A Generalized McNemar's Test

As in the case of evaluation on labeled data, the test of Section 2.2 can be modified to account for the fact that both P and Q are classifying the same data set.

Let N_{pqr} represent the number of instances on which the the classification outcomes are $P \rightarrow p$, $Q \rightarrow q$ and $R \rightarrow r$. *E.g.* N_{TFT} is the number of events that P and R classified correctly, but Q misclassified.

Since our data are not labeled, we are only able to observe *agreements* between the classifiers, but we cannot determine the correctness of the output of any classifier. Thus we only observe the following counts:

$$\begin{aligned} N_{P\bar{Q}R} &= N_{TFT} + N_{FTF} \\ N_{\bar{P}QR} &= N_{FTT} + N_{FTF} \\ N_{PQ\bar{R}} &= N_{TTF} + N_{FFT} \\ N_{PQR} &= N_{TTT} + N_{FFF} \end{aligned}$$

where $N_{P\bar{Q}R}$ is the number of times P and R agree, but Q does not agree with them, $N_{\bar{P}QR}$ is the number of times Q and R agree with one another, but not P, $N_{PQ\bar{R}}$ is the number of times P and Q agree with one another, but not R and N_{PQR} is the number of times all of them agree.

Let $T_{P\bar{Q}R}$ be the probability that P and R agree, but not Q, and $T_{\bar{P}QR}$ the probability that Q and R agree, but not P. Under the null hypothesis, $T_{P\bar{Q}R} = T_{\bar{P}QR}$, or equivalently, $T_{PQ} = T_{P\bar{Q}R}/(T_{P\bar{Q}R} = T_{\bar{P}QR}) = 0.5$. In other words, according to the null hypothesis, the conditional probability of P agreeing with R, given that only one of P or Q agree with R is 0.5.

The standard derivation of McNemar's test now follows. The conditional probability distribution of $N_{P\bar{Q}R}$ is a binomial distribution $\mathcal{B}(N_{PQ}, 0.5)$, where $N_{PQ} = N_{P\bar{Q}R} = N_{\bar{P}QR}$. The probability of getting at least $N_{P\bar{Q}R}$ out of N_{PQ} agreements under the null hypothesis is given by

$$P_{test} = \begin{cases} 2P(N_{P\bar{Q}R} \leq n \leq N_{PQ}) & \text{if } N_{P\bar{Q}R} > 0.5N_{PQ} \\ 2P(0 \leq n \leq N_{P\bar{Q}R}) & \text{if } N_{P\bar{Q}R} < 0.5N_{PQ} \\ 1.0 & \text{if } N_{P\bar{Q}R} = 0.5N_{PQ} \end{cases} \tag{3}$$

where

$$P(N_1 \leq n \leq N_2) = (0.5)^{N_{PQ}} \sum_{n=N_1}^{N_2} \binom{N_{PQ}}{n} \tag{4}$$

If P_{test} is smaller than a chosen significance threshold (*e.g.* 0.05 or 0.01) the null hypothesis is rejected. If so, then the sign of $p - q$ is assumed to be the same as the sign of $N_{P\bar{Q}R} - N_{\bar{P}QR}$.

Note once again that the test does not depend on the actual value of r. The effect of r is indirect – reducing r reduces N_{PQ}.

Model	Devset1	Devset2	Joint
R1	54.3	50.1	52.6
R2	64.3	60.1	62.6
R3	73.6	68.8	71.7
R4	20.2	16.0	18.5
R5	54.6	50.1	52.8

Table 2. Recognition accuracy of the various reference models on the three validation sets.

3. TESTING MULTI-CLASS CLASSIFIERS

The simple test based on binary classifiers described above assumes that when two classifiers misclassify data, they agree with one another. This is generally not the case for multi-class classification problems.

The effect of this on the simple test of Section 2.2 is minimal. Given two classifiers P and R, the probability that they will agree is now given by $t_{pr} = pr + \alpha(1-p)(1-r)$, where α is the fraction of all instances that are misclassified by both P and R,where they agree (*i.e.* they both select the same wrong class). The required relation for t_{pr} to be monotonic in p is $r > \alpha/(1+\alpha)$. Since α can be as low as $1/(n_{class}-1)$, where n_{class} is the number of classes, this means that the requirement on r is less strict than in the case of the binary classifier where r was required to be greater than 0.5. The rest of the test does not change.

In the case of the generalized McNemar's test, the picture is similar. Table 1 shows the various possible outcomes and their probabilities. $T_{P\bar{Q}R}$ now takes the form $T_{P\bar{Q}R} = p(1-q)r + \alpha(1-p)q(1-r) + \alpha(1-\beta)(1-p)(1-q)(1-r)$. Similarly $T_{\bar{P}QR} = (1-p)qr + \beta p(1-q)(1-r) + (1-\alpha)\beta(1-p)(1-q)(1-r)$. $\beta = \alpha$ under the null hypothesis. Consequently, the rest of the test still applies, with the additional factor that r can be less than 0.5 (and as low as $1/n_{class}$.

4. EXPERIMENTAL EVALUATION

We ran a number of experiments to evaluate the predictions made by the proposed tests on a speech recognition task. The CMU Sphinx-III speech recognition system was used for all experiments. Since the form of the proposed generalized McNemar's test assumes a one-to-one correspondence between words in the word sequences being compared, the recognition performances (agreement or accuracy) reported below do not include insertions (deletions were equated to substitution by a $\emptyset$ symbol). We trained five acoustic models using various corpora. The goal was to determine the relative performance of the models using untranscribed data. We refer to these models as the "evaluation" models. The models were trained on two corpora: the far-field training set of the wall street journal (WSJ) database, and a *Librivox* (www.librivox.org) recording of the book "Emma" by Jane Austen read by Elizabeth Klett comprising 15 hours of data (which we will refer to as the "L1" data below). The five evaluation models were:

A: Models trained from Wall Street Journal (WSJ) far-field data.

B: "A" adapted to the L1 data using 3-class MLLR.

C: "A" adapted to the L1 data by 12-class MLLR.

D: "A" adapted to the L1 data by 40-class MLLR.

E: Models trained entirely on the L1 data.

On Devset1					
Ref. $\rightarrow$	R1	R2	R3	R4	R5
models $\downarrow$	Acc(%)	Acc(%)	Acc(%)	Acc(%)	Acc(%)
A	37.5	36.4	35.6	18.9	36.8
B	45.6	49.7	50.6	19.9	45.2
C	47.9	53.2	56.4	20.0	47.5
D	47.8	53.8	58.1	19.8	47.8
E	53.9	63.4	72.9	19.7	54.1
On Devset2					
A	36.1	35.2	34.8	17.1	36.1
B	43.4	47.8	49.5	17.5	43.2
C	45.5	51.1	55.0	17.5	45.5
D	45.2	51.0	56.2	17.2	45.4
E	50.8	59.5	68.6	16.9	50.9
On Jointset					
A	37.0	35.9	35.3	18.2	36.5
B	44.8	48.9	50.1	19.0	44.4
C	46.9	52.4	55.8	19.0	46.7
D	46.8	52.7	57.3	18.8	46.8
E	52.7	61.8	71.2	18.6	52.8

Table 3. Agreement percentages between the outputs of recognizers A-E and the outputs from the reference models R1-R5, on each of three validation sets.

All models represented triphones as 3-state HMMs with 6000 tied states, each modeled by mixtures of 8 Gaussians.

As unlabeled validation data we used two more recordings of Jane Austen's books from Librivox, both read by Elizabeth Klett: "Pride and Prejudice" (total running time: $\tilde{1}1$ hours), which we will refer to as "Devset1", and "Persuasion" (total running time: $\tilde{8}$ hours), which we will refer to as "Devset2". We also used the combination of both recordings as a larger validation set, which we refer to as the "Joint" set. Devset1 had a total of 123923 words, Devset2 had 84503 words and the Joint set had 208426 words.

The accuracy of the tests depends only indirectly on the accuracy of the reference recognizer; nevertheless we tried a several reference recognizers with different accuracies to evaluate the effect of the reference recognizer:

R1: Models trained on Hub4 97 data.

R2: R1 adapted by to Deveset1 using 4-class MLLR.

R3: R1 adapted to Devset1 using 45-class MLLR.

R4: Models trained on Resource Management data.

R5: Models trained on Hub4 98 data.

Reference models R1 and R5 are similar, being trained on similar data. R4 was purposely chosen to be of poor accuracy. R2 and R3 were obtained by adapting R1 to one of the two Devset recordings, in order to adapt them to the speaker (who is the same for both sets). The intention here was to create two reference models that had higher accuracy than R1, with R3 being more accurate than R2. The reference recognizers all used 5000 tied states with 8 Gaussians per state. In all recognition experiments, a trigram language model trained from two Jane Austen novels, "Mansfield Park" and "Northanger Abbey" was used.

Table 2 shows the actual accuracies of the various reference models on the validation sets: Devset1, Devset2 and Joint. The performance is as predicted above, with only the difference between R1 and R5 being statistically insignificant at the 0.01 level. We did not, of course, use these known accuracies in our statistical significance

	Devset1	Devset2	Joint
R1	ABDCE	ABDCE	ABDCE
R2	ABCDE	ABDCE	ABCDE
R3	ABCDE	ABCDE	ABCDE
R4	AEDBC	CBDAE	AEDBC
R5	ABCDE	ABDCE	ABCDE

Table 4. Summary of decisions made by the reference classifiers about which of A,B,C,D and E are likely to perform better on a Test-set, based on recognition of various validation sets. A decision X>Y is written as XY, and indicates that X performs better than Y. *E.g*, the string ABCDE indicates that the ranking is $A > B > C > D > E$.

tests. To compare the models A-E, the three validation sets were recognized using the reference models and the evaluation models. The recognition output of the evaluation models were compared to the that of the reference model. Table 3 shows the individual agreements beteween the outputs of each of the evaluation models and the various reference models on each validation set. We note that Devset1 is larger than Devset2, and the Joint set is larger than both.

Based only on the comparison of the agreements between the various reference models and the evaluation models, the rank ordering of the various models, as predicted by each of the reference models is given by Table 4 The various pairwise comparisons in the table must be accepted if corresponding null hypothesis (that both models in the comparison are equivalent) is rejected.

In principle, the results of Table 3 should be sufficient to perform the simple test of Section 2.2. However, we employed the generalied McNemar test to compare the models, for which agreement-disagreement counts such as $N_{P\bar{Q}R}$ are required. The outcome of this computation is shown in Figure 1. Each grid in the figure shows the comparisons between the evaluation models A-E, based on recognition of a specific validation set, using one of the five reference recognizers. Each column represents the comparison of a particular recognizer to the other recognizers. A grey or black box show that the recognizer for the column shows better agreement with the reference recognizer than the recognizer for the row. However, only black boxes represent instances where the null hypothesis that the two recognizers are actually the same has been rejected with a confidence level of 0.01. In these cases the tests predict that the recognizer represented by the column will be superior to that for the row with a probability of at least 99%.

We note that in the boxes shaded black all tests agree, although the grey boxes are often inconsistent. We further note that the number of comparisons about which we can be confident depends on the accuracy of the reference recognizer. The worst reference recognizer, R4, does not make confident predictions about anything. We also note that the number of black boxes (confident predictions) increases as the validation set increases in size.

To test if these predictions actually compare to reality, we compared all recognizers on a labeled test set, which was yet another recording from Librivox: "Sense and Sensibility" by Jane Austen, also read by Elizabeth Klett. The recording time for this data was almost 11 hours. The data had 112174 spoken words. Table 5 shows the recognition results for the various models on the test set. All differences are significant to the 0.01 level. Note that all of the confident predictions made in Figure 1 are confirmed by the test.

5. DISCUSSION

The new generalized McNemar's test is observed to be accurate at making predictions about performance on a test set using only unlabeled validation data. The predictions are also observed to follow expected trends. The number of confident predictions it is able to make increases with the size of the validation data. It also increases as the accuracy of the reference recognizer improves. Confident predictions are less likely when the recognizers being compared are close in accuracy. The ability to predict generally follows the expected trend in the number of words on which one recognizer agrees and another disagrees with the reference recognizer.

The proposed tests can be improved. We do not consider insertions. The equivalent of MAPSSWE [6], that does consider these factors, can easily be developed. We also do not have a model for agreements on misclassification for multi-class data. Generalized tests that consider these will be presented at a later venue.

	A	B	C	D	E
% Acc.	34.2	47.0	52.3	54.3	81.1
Wds	38306	52717	58701	60927	90990

Table 5. % Accuracy, and number of words correctly recognized in the test set by various models.

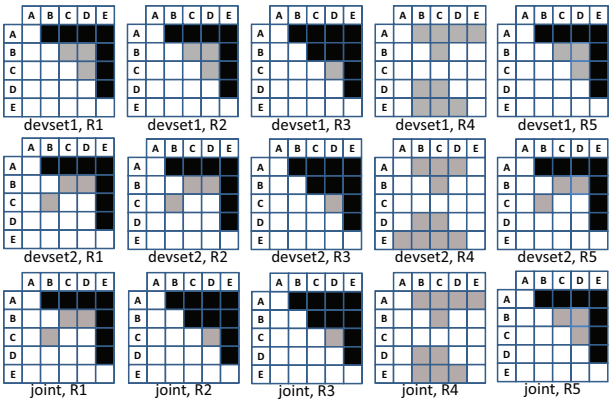

Fig. 1. Predictions and confidence levels obtained from various reference models on various validation sets.

6. REFERENCES

[1] L. Lamel, J.-L. Gauvain, and G. Adda, "Unsupervised acoustic model training," in *Proc. ICASSP*, 2002, pp. 877–880.

[2] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *Proc. ICASSP*, 2009, pp. 4297–4300.

[3] R. Singh, B. Lambert, and B. Raj, "The use of sense in unsupervised training of acoustic models for hmm-based asr systems," in *Proc. INTERSPEECH*, 2010.

[4] G. W. Corder and D. I. Foreman, *Non parameteric statistics for non-statisticians: a step-by-step approach*, Wiley, New Jersey, 2009.

[5] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, 1989, pp. 532–535.

[6] D. S. Pallet, W. M. Fisher, and J. G. Fiscus, "Tools for analysis of benchmark speech recognition tests," in *Proc. ICASSP*, 1990, pp. 97–100.