SCORING STRUCTURAL MDE: TOWARDS MORE MEANINGFUL ERROR RATES

M. Ostendorf and D. Hillard

Department of Electrical Engineering, University of Washington, Seattle, WA
Contacts: {mo,hillard}@ee.washington.edu

ABSTRACT

This paper outlines issues and explores options for augmenting the current structural MDE scoring with the goal of obtaining more meaningful measures of performance. We propose event-based statistical significance measures (similar to STT), evaluation of confidence predictions (important to downstream processing applications), and accounting for annotation differences (due to inherent variability in annotations). Focusing on SUs, we give results for SRI-ICSI-UW system outputs for reference and STT conditions with preliminary implementations of these proposals.

1. INTRODUCTION

Scoring methods for speech-to-text (STT) system output are relatively well established, and software scoring tools have been widely available for several years. Several significance tests for word error rate (WER) difference and a confidence scoring technique have long been used. As error rates have dropped, the issue of differences in human annotation have become important to address, but researchers have made good progress with over a year of work looking into methods for taking into account annotator variability.

Current metadata extraction (MDE) scoring techniques, in contrast, are not as informative as they could be for several reasons. First, until now there have been no standard measures of statistic significance of performance differences, so it is difficult to draw conclusions about the value of the many techniques that give small gains. Second, the standard confidence annotation paradigm of marking confidence only on detected events omits valuable information for downstream processing, and

the standard STT confidence scoring method is complicated by the significant impact of STT word deletions on the score. Lastly, the current measures report error in terms of the number of reference events, rather than the number of words, though decisions are in fact made at the word level. There are good reasons for this choice, since the large number of words with no marked events implies that one can achieve low error rates with a word-based score even for the case when no events are ever detected. However, using the event count as the denominator also introduces some complications. In particular, the impact of annotator differences is much greater than in a word-based score, because a larger percentage of the events may be in disagreement, so the current high error rates of MDE systems are somewhat misleading.

The goal of this paper is to move MDE scoring more in the direction of STT, by adding more powerful significance tests to the current inventory, introducing a confidence scoring method, and exploring options for accounting for confusability. To restrict the scope of the implementation and experiments, we have chosen to focus on scoring sentence-like boundaries, referred to as SUs, primarily because these are the events of greatest importance for downstream processing and they are also the most frequently occurring. Hence, in the next section, we briefly review the definition of an SU and common aspects of SU detection systems. The remainder of the paper proceeds with a discussion of the scoring methods and questions, together with experiments comparing the outputs of a few systems on the RT04 evaluation test set for both conversational telephone speech and broadcast news.

2. SU DETECTION

An SU roughly corresponds to a sentence, except that SUs are for the most part defined as units that include only one independent main clause, and they may sometimes be incomplete, as when a speaker is interrupted and does not complete their sentence. There are four types of SUs considered here: statement, question, backchannel, and incomplete. A more specific annotation guideline for SUs is available [1], which we refer to as the "V6" standard. These guidelines have been used to annotate both conversational telephone speech (CTS) and broadcast news (BN) data.[1]

Errors are measured by a slot error rate (SER) similar to the WER metric utilized by the speech recognition community, i.e. dividing the combined number of SU insertions,

[1] For the RT04 evaluation, the guidelines are modified somewhat to map the V6 discourse response class to different SUs categories by rule depending on word cues.

Table 1. WER and SER for different STT systems compared here, using the best-case MDE system based on the 1-best STT output.

STT system	WER	SER ±SU	SER w/ type
Broadcast News			
reference	0.0	47.65	50.16
Superears	11.9	57.70	59.97
SRI-only	15.2	60.49	62.73
CTS			
reference	0.0	26.70	37.30
SRI-IBM	14.9	36.46	47.30
SRI-only	18.7	40.51	51.09

Table 2. SER for different MDE systems compared here on the CTS task, where the STT output is based on a pruned N-best list from the SRI-only system (18.6% WER) and the MDE system uses only the HMM and Maxent components.

STT system	SER ±SU	SER w/ type
1-best (pruned)	41.2	52.2
N-best	40.5	51.5

deletions and substitutions by the total number of reference SUs. When recognition output is used, the words will generally not align perfectly with the reference transcription and hence the SU boundary predictions will require an alignment procedure to match to the reference location. In the standard scoring software, md-eval (v19a), the alignment is based on the minimum word error alignment of the reference and hypothesized word strings and on the minimum SU error alignment when the WER is equal for multiple alignments. In many of our experiments, we make use of the word alignments, which are available via a debugging output of the tool.[2]

Tables 1 and 2 also give the WER and SER figures for the various systems compared here, all of which are based in the SRI-ICSI-UW basic framework. One series of comparisons (Table 1) covers scoring issues for cases where the STT system output reflects a range of WERs and the MDE system is fixed. In these experiments, the MDE system operates on the 1-best STT system output, incorporates all knowledge sources and combines HMM, maximum entropy and conditional random field decisions. Another series of comparisons (Table 2) looks at small system differences, where the STT system is fixed (the SRI-only systems) but the MDE system is based on either the 1-best vs. the N-best STT output. The N-best MDE system used in these experiments uses only the HMM and maximum entropy components, and to reduce the cost of prosody feature extraction, the N-best list is pruned to include only those hypotheses that represent 98% of the total probability mass (hence the 1-best WERs in Tables 1 and 2 differ slightly).

[2]We use the option with extent matching equal to 1 and the maximum gap (in words) between matching ref/sys metadata events equal to 0.0, in order to get the same error for both the exact NIST metric and the secondary extent matching error associated with the debugging output. As a consequence, the results reported here are slightly better than the offi cially reported NIST scores, because the alignment is actually optimized for the extent matching score and not the NIST score. In the default case, these are different, but with the above parameters they are equal.

3. COMPARING DIFFERENCES BETWEEN SYSTEMS

There are currently four significance tests used by NIST in assessing differences between speech recognition systems [2]. Two measures compare speaker-level performance differences, using either the sign test or the Wilcoxin signed rank test. Two other measures compare segment-level performance differences, using either the matched pair or McNemar tests introduced in [3]. Segments are defined as a unit bounded by time points where both system hypotheses correctly match the reference for a sequence of two words, so as to obtain units that can reasonably be assumed to be independent for use in the significance tests.

The speaker-level significance tests translate easily to the structural MDE scoring task, but they are relatively weak tests because the number of speakers (the "n" in the significance test) is not large. Of the two NIST speaker-level tests, the Wilcoxon signed-rank test tends to be somewhat more powerful, so we chose this one for comparing to segment-level results to be developed here.

The key problem in using the segment-level tests is the definition of a suitable "chunk" of speech on which we can assume errors are roughly independent from one to the next. Here, we consider three options: A) segments of speech bounded by speaker change points (relevant for BN only); B) segments of speech from a single speaker, bounded by a pause of at least X seconds; and C) segments of speech bounded by points where either both systems correctly recognize SUs or there is a Type B boundary. Table 3 shows that count for the different types of units as a function of pause duration threshold, using two SU-detection systems with fairly similar performance (in both cases the SRI-only system is compared to a multi-site system that has lower WER and correspondingly lower MDE error rates). The 3s pause threshold seems to give a reasonable number of segments, but was less likely than 2s to break up a segment, so in the remainder of the paper tests use X=3s. (For future work, it may be preferable to have different thresholds for

Table 3. Effect of segment definition on the "n" for significance testing in the 2004 evaluation test sets.

Test Set	spkrs	Type A	Type B			Type C
			X=4s	X=3s	X=2s	X=3s
CTS	72	–	943	1109	1430	2928
BN	235	701	731	747	777	1328

Table 4. Assessing the difference between SRI 1-best vs. N-best using Wilcoxin (WI), matched pair (MP) segment and McNemar (MN) segment significance tests, for segmentation types B and C.

Task	WI	MP-B	MN-B	MP-C	MN-C
STT WER contrasts					
CTS	10e-5	10e-5	10e-5	10e-5	10e-5
BN	.004	.0014	.855	.0014	.994
STT N-best vs. 1-best					
CTS	.076	.104	.68	.102	.88

BN and CTS.)

We ran significance tests comparing two systems which had fairly different SU detection performance (see Table 1) and two systems with similar performance (see Table 2). Table 4 summarizes the results, which at first seemed counter intuitive but on closer inspection did make sense. The pause-based type B segmentation gave very similar results compared to the hypothesis-matching type C segmentation, despite a large difference in number of segments between the two. The smaller number of segments for type B is as effective because each segment is longer and has greater variability. Since the pause-based type B segmentation is easier to obtain and does not depend on the systems being compared, we will use this approach in further experiments. The McNemar test was not very useful because it only tests on segments where one system perfectly recognizes the SUs in a segment and the other does not, which is very rare for this task. For the vast majority of cases, either both systems did well, i.e. for easy segments, or they both made errors (particularly for BN where segments tend to be quite long).

4. CONFIDENCE SCORING

The measure used for evaluating confidence estimates in STT is normalized cross entropy (NCE) [4]:

$$NCE = (H_{max} - H_{conf})/H_{max}$$

where

$$\begin{aligned} H_{max} &= -p_c \log_2 p_c - (1-p_c)\log_2(1-p_c) \\ H_{conf} &= -1/n[\sum_{w_i\ corr} \log_2 p_i + \sum_{w_i\ err} \log_2(1-p_i)] \end{aligned}$$

where $p_c = n_c/n$ is the average probability that a hypothesized word is correct, p_i is the predicted confidence that w_i is correct, and the sum is over all n hypothesized words in the test set. (The NCE score has also been referred to as normalized mutual information.) Another method for evaluating confidence scores, proposed in [5], is a DET curve or the equal error rate (EER) point associated with the DET curve. Siu and Gish show that both the EER and NCE measures interact with recognition accuracy, with EER having a linear dependence and NCE having a more complex dependence on WER. Hence, it is difficult to compare NCE results for systems with very different accuracies. Still, it has proven to be a useful method for comparing confidence predictions of systems with similar accuracy. The extension of the STT confidence scoring method to MDE requires some issues to be resolved.

First, *should confidence be marked for the binary SU/non-SU category or for the multi-valued SU categories?* The answer to this question may be application dependent. Many of the techniques for determining SU type involve language cues and might be better handled by downstream processing. If so, it would be better to provide either binary SU/non-SU confidences or the full posterior distribution for the different types (feasible given the small inventory). Since it is not clear that SU type confidence will be useful, and to simplify our analysis, we consider only binary SU confidence here.

Next, *should confidences be marked at every word boundary or just at hypothesized events?* We argue that the answer to this question should be to use confidences marked at every word boundary, for two main reasons. First, in informal polling of people working on downstream language processing about their preferences (at the 2004 TIDES meeting), the consistent answer was word-based confidences. Second, if we used event-based confidence and directly implemented the STT NCE formula, then the sum over recognized SUs would not reflect deletion errors, which can represent over half of SU errors scored without sub-type.

This highlights a potential problem with the NCE measure: since the sum is only over the hypothesized words, *deleted words do not have confidence estimates and therefore do not impact the NCE.* Although word deletions are small in number relative to hypothesized words, the percentage of deleted SUs falling on deleted words is not small, e.g. 25% for CTS data. Roughly half of the CTS cases occur for very short utterances, which perhaps are less important because they would be easy to identify if the words are present, but this still leaves a large percentage of the reference SU events unaccounted for if the above sum is used for computing NCE. One view is that the deletions are accounted for in WER, so there is no need to account for these in the NCE score. Another possibility is to add terms to the entropy sums for the deletions with a fixed confidence value

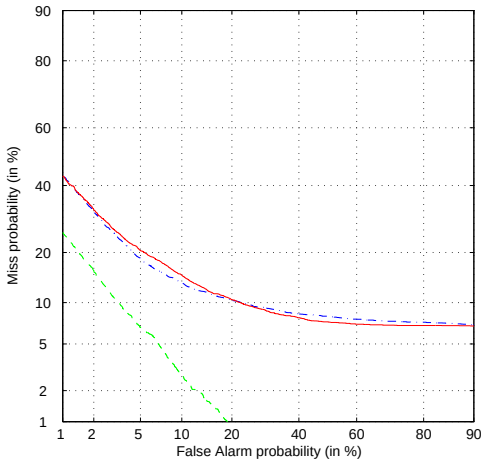

Fig. 1. DET curve for SU detection based on confidence predictions for CTS reference, SRI 1-best, and SRI N-best systems.

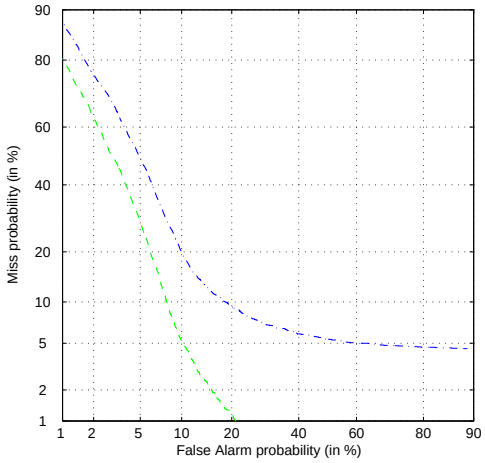

Fig. 2. DET curve for SU detection based on confidence predictions of for BN reference and Superears systems.

for SUs on deletions. Our preferred alternative is to use a DET curve for evaluation instead of NCE, where the curve is obtained by sweeping over a range of confidence thresholds for deciding on presence of an SU and measuring the false vs. missed detection rates.

To better understand the behavior of the different confidence measures, we look at DET curves and NCE scoring of confidence estimates for SU prediction based on the reference transcriptions and the SRI-only 1-best and N-best STT transcripts. The DET curves are shown in Figures 1 and 2, and the NCE scores (ignoring word deletions) are given in Table 5. The effect of deletions shows up in the DET curve as a flooring on the missed detection rate. The DET curves show dramatic differences between the performance on reference vs. STT transcripts for both CTS and BN, while the NCE scores give mixed results (e.g. for CTS, the MDE confidences score better when the system is given the STT output vs. the reference transcript, which we would not expect). Further, the DET curve shows little difference between the 1-best and N-best system confidences for the CTS task, while the NCE score shows a large difference with much worse results for the N-best case. These results demonstrate that NCE is not very useful for SU confidence evaluation. While there are problems with the DET curve in that it is threshold free (the detection threshold need not be at confidence 0.5, and confidence scores need not even be in the [0,1] range), it appears to appropriately reflect our intuitions about system differences. Further, it allows downstream processing flexibility in determining the operating point.

Table 5. NCE scores for MDE based on reference, 1-best and N-best hypotheses for SU vs. no-SU confidences.

	ref	1-best	N-best
BN	.86	.84	–
CTS	.76	.81	.61

Table 6. Average differences between independent SU annotations of the BN and CTS eval04 test sets, as measured by md-eval.

Task	Comparison	SER
BN	Anno1 vs. Anno2	12.7
BN	AnnoX vs. Adjud	6.8
CTS	Anno1 vs. Anno2	17.5
CTS	AnnoX vs. Adjud	9.2

5. ACCOUNTING FOR ANNOTATION VARIATIONS

Perhaps the biggest problem facing MDE scoring is that of accounting for annotation variability. There are several sources of variability, including the unavoidable problem of human error, ambiguities in annotation guidelines, and actual ambiguities in the data. Using the md-eval tool to compare the annotations to each other and to the adjudicated version shows a fairly large proportion of differences, as shown in Table 6 where results represent the average over the two possible assignments of the annotations as reference vs. hypothesis. Not surprisingly, there are a larger percentage of annotation differences for CTS than for BN. As we will show here experimentally, these annotation disagreements do impact the score of automatic systems, with the impact on CTS being much greater because of the larger number of annotation differences. (Note that all the BN experiments reported in this section are only on half of the BN eval data, since independent annotations were only available for that subset.)

While STT has essentially the same problem, though less severe, the proposed scoring solutions involving several annotations of the test data will not work for MDE because of the high cost of metadata annotation. In addition, the interaction between sequential events (e.g. SUs) is a bigger issue for MDE because of the longer span of events.[3] For simplicity, we chose to control for sequential effects by using segment-level scores when allowing alternative annotations. We report results for four scoring alternatives, all based on md-eval:

- compare the SU hypotheses to the adjudicated annotations, as in official score (adj.);
- for each pause-based segment (as in significance testing), score the SU hypothesis against the independent annotations and the adjudicated annotation and use the lowest error rate for that segment (3-ref);
- same as above but use only the two independent annotations (2-ref); and
- score only those pause-based segments where independent annotators agreed, ignoring regions where there is disagreement since we cannot be certain that it is a legitimate ambiguity vs. an annotator error (agree).

[3] If one annotator marks a boundary at location X but not Y, and the other marks it at location Y but not X, under what conditions is it reasonable to assume that it is acceptable for a system to recognize both (vs. just one or the other)? There are anecdotal examples we have seen where either view would be appropriate.

Table 7. SU SER as measured against the adjudicated reference only (adj), the best segment score among the 3 references (3-ref), the best segment score among 2 independent references (2-ref), and only considering segments where the independent annotations are in agreement (agree).

		SER			
Task	System	adj.	3-ref	2-ref	agree
BN	super-ears	61.2	58.1	59.1	55.2
CTS	SRI+IBM	47.3	43.0	43.5	33.0

The last case covers a little over half of the segments, but less than a third of the SUs, raising the possibility of difficulties in establishing significance in the difference between systems. However, in revisiting the SRI-only vs. multi-team system comparisons using the matched pair test on pause-based segments, we find that the CTS system differences are still highly significant and the BN differences only become slightly less significant, with the significance level changing from .001 to .005.

Assuming that the "agree" case represents the noise-free measure, then we could argue that 10% (relative) of the current error rates can be attributed to noise for BN and 30% for CTS for both reference and STT-based conditions, as summarized in Table 8. Of course, the situation is probably not quite so optimistic, given that it is likely that the cases that are difficult for humans are also difficult for our automatic systems, so the "agree" measure may also have an optimistic bias. Note that the measurement noise for MDE is higher than that for STT, which is in the range of 15-25% for CTS depending on the system [6]. The percentage of WER noise for BN is much lower, but less well studied.

We can also compare STT and MDE system error relative to a gold standard to human interannotator differences. For MDE, comparing the SER based on the adjudicated reference to the interannotator differences in Table 6, the system error rate is a factor of 2-3 times the human SU disagreements for CTS, depending on whether the reference transcripts vs. STT output is used. This can be compared to a factor of 4-5 times for STT WER vs. human word transcription disagreements on CTS data [6]. Hence, while it

Table 8. Estimates of the relative amount of noise the SER (percent difference between the adj. and agree scores) based on different annotations.

Task	System vs. adj	System vs. agree	Noise in SER
BN-stt	61.5	55.2	10%
BN-ref	51.8	46.3	11%
CTS-stt	47.3	33.0	30%
CTS-ref	37.3	26.1	30%

is clear that the automatic MDE systems are not yet close to human performance, the relative difference between the (noisy) system error rates and the rate of human agreement is actually lower for MDE than for STT. For MDE of BN, the relative difference between the systems and humans is greater (4-5 times), and possibly closer to the gap for STT.

6. SUMMARY

In summary, this paper explored MDE scoring issues related to significance testing, confidence scoring and scoring to account for annotation variation. Based on exploring different options, we recommend:

- Significance testing with the matched pair test based on speech chunks defined in terms of long pauses and/or speaker changes;
- Confidence scoring based on a DET curve; and
- Accounting for annotation variability by scoring only on speech chunks where dual annotations agree.

The focus of the examples given here has been on SUs alone, for reasons given earlier. SUs fall into the class of boundary events, i.e. events that occur between words, and the scoring methods proposed here would also apply to other boundary events, including IPs. Other types of structural metadata that are characterized as word labels, as in the current scoring of edits and fillers, would fit relatively easily into the proposed frameworks. However, the extension to scoring events with multi-word spans would be more difficult.

Acknowledgments

This work is supported in part by DARPA contract no. MDA972-02-C-0038. Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these agencies. Thanks to Barbara Peskin and Liz Shriberg for suggestions of approaches to consider, and to Yang Liu for providing confidences and SU outputs for several different conditions.

7. REFERENCES

[1] S. Strassel, *Simple Metadata Annotation Specification V6.2*, Linguistic Data Consortium, 2004.

[2] NIST, "Significance Tests for ASR," http://www.nist.gov/speech/tests/sigtests/sigtests.htm, 2000.

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

[4] NIST, "The 2001 NIST Evaluation Plan for Recognition of Conversational Speech over the Telephone," http://www.nist.gov/speech/tests/ctr/h5_2001/h5-01v1.1.pdf, 2000.

[5] M. Siu and H. Gish, "Evaluation of word confidence for speech recognition systems," *Computer Speech and Language, 13(4)*, pp. 299–318, 1999.

[6] J. Fiscus and R. Schwartz, "Analysis of scoring and reference transcription ambiguity," in *Proc. of the 2004 Rich Transcription Workshop*, 2004.