COMPARING EVALUATION METRICS FOR SENTENCE BOUNDARY DETECTION

Yang Liu[1] *Elizabeth Shriberg*[2,3]

[1]University of Texas at Dallas, Dept. of Computer Science, Richardson, TX, U.S.A.
[2]SRI International, Menlo Park, CA, U.S.A.
[3]International Computer Science Institute, Berkeley, CA, U.S.A.

ABSTRACT

In recent NIST evaluations on sentence boundary detection, a single error metric was used to describe performance. Additional metrics, however, are available for such tasks, in which a word stream is partitioned into subunits. This paper compares alternative evaluation metrics—including the NIST error rate, classification error rate per word boundary, precision and recall, ROC curves, DET curves, precision-recall curves, and area under the curves—and discusses advantages and disadvantages of each. Unlike many studies in machine learning, we use real data for a real task. We find benefit from using curves in addition to a single metric. Furthermore, we find that data skew has an impact on metrics, and that differences among different system outputs are more visible in precision-recall curves. Results are expected to help us better understand evaluation metrics that should be generalizable to similar language processing tasks.

Index Terms— sentence boundary detection, precision, recall, ROC curve

1. INTRODUCTION

Sentence boundary detection has received increasing attention in recent years, as a way to enrich speech recognition output for better readability and improved downstream natural language processing [1, 2, 3]. Automatic sentence boundary detection itself was part of the recent NIST rich transcription evaluations.[1] In addition, studies have been conducted to evaluate the impact of sentence segmentation on subsequent tasks, including speech translation, parsing, and speech summarization [2, 4, 5].

In the NIST evaluations, system performance for sentence boundary detection has been evaluated using an error rate (total number of inserted and deleted boundaries, divided by the number of reference boundaries). Research studies ([2, 3]) have also begun to look at the ROC curve, DET curve, and F-measure. Of course, since the ultimate goal is to aid downstream language processing tasks, a proper way to evaluate sentence boundary detection would be to look at the impact on the downstream tasks. In fact, in [2] it was shown that the optimal segmentation for parsing is indeed different from that obtained when optimizing for sentence boundary detection accuracy using the aforementioned NIST metric. Other application-based studies include the impact of sentence boundary detection for machine translation [4] and for summarization [5].

This paper examines various evaluation metrics and discusses their advantages and disadvantages. We focus on the boundary detection task itself, rather than its impact on downstream applications. In addition, we evaluate the effect of different priors of the event of interest (i.e., sentence boundaries) by using different corpora. Highly skewed priors are inherent to this and related tasks, since boundary events are typically rare compared to nonboundaries. Unlike most studies in machine learning, this work focuses on a real language processing task. The study is expected to help us better understand evaluation metrics that will be generalizable to many similar language processing tasks involving segmentation of the speech stream into subunits — for example, topic, story, or dialog act.

[1]See http://www.nist.gov/speech/tests/rt/rt2004/fall/ for more information on NIST evaluations.

2. METRICS

The task of sentence boundary detection in speech is to determine the location of boundaries given a word sequence (typically from a speech recognizer) and associated audio signal. In this study, we use reference transcriptions, to avoid confounding discussion with the effects of recognition errors themselves; the metrics for recognition output apply similarly after aligning recognition output with reference transcriptions. We can represent the task as two-way classification or detection, that is, label each interword boundary as either "sentence" or "no-sentence". Table 1 shows the notation we use for the confusion matrix result. For a given task, the total number of samples is $tp + fn + fp + tn$, and the total number of positive samples is $tp + fn$.

	system true	system false
reference true	tp	fn
reference false	fp	tn

Table 1. A confusion matrix for the system output. "True" means positive examples, that is, sentence boundaries in this task.

2.1. Metrics Description

Various metrics have been used for evaluating sentence boundary detection or similar tasks, in individual studies. For example, in [6, 7], metrics are developed that treat the sentences as units and measure whether the reference and hypothesized sentences match exactly. Slot error rate [8] was introduced first for information extraction tasks, and later used for sentence boundary detection. Kappa statistics have often been used to evaluate human annotation consistency, and can also be used to evaluate system performance, that is, treating system output as a 'human' annotation. Other metrics in the general classification literature, such as cost curves [9], have not been widely used for evaluating sentence boundary detection.

In this paper we focus on the following metrics:

- NIST metric. The NIST error rate is the sum of the insertion and deletion errors per the number of reference sentence boundaries. Using the notation in Table 1, this becomes

$$\text{NIST error rate} = \frac{fn + fp}{tp + fn}$$

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2007		**00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Comparing Evaluation Metrics for Sentence Boundary Detection	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **International Computer Science Institute,1947 Center Street Suite 600,Berkeley,CA,94704**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**4**	

Note that the NIST evaluation tool *mdeval*[2] allows boundaries within a small window to match up, in order to take into account different alignments from speech recognizers. We ignore that detail in this study and simply treat the task as straightforward classification.

- Classification error rate. If this task is represented as a classification task for each interword boundary point, then the classification error rate (CER) is

$$CER = \frac{fn + fp}{tp + fn + fp + tn}$$

- Precision and recall. These metrics are widely used in information retrieval, and are defined as follows:

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

A single metric is often used to reflect both precision and recall, and their tradeoff:

$$\text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

- ROC curve. Receiver operating characteristic (ROC) curves are used for decision making in many detection tasks. They show the relationship between true positives ($=\frac{tp}{tp+fn}$) and false positives ($=\frac{fp}{fp+tn}$) as the decision threshold varies.
- PR curve. The precision-recall (PR) curve shows what happens to precision and recall as the decision threshold is varied.
- DET curve. A detection error tradeoff (DET) curve plots the miss rate ($=\frac{fn}{tp+fn}$) versus the false alarms (i.e., false positive), using the normal deviate scale [10]. It is widely used in the task of speaker recognition, but less used for other classification problems.
- AUC. The curves above provide a good view for the system's performance at different decision points. However, a single number is often preferred when comparing two curves or two models. Area under the curves (AUC) is used for this purpose. It is often used for both ROC and PR curves, but less so for DET curves.[3]

2.2. Relationship

For a given task, the number of positive samples (i.e., $np = tp+fn$) and the total number of samples (i.e., $tp + fn + fp + tn$) are fixed. Therefore, precision and recall uniquely determine the confusion matrix, and hence the NIST error rate and classification error rate. Each of the two error rates can uniquely determine the other since the denominators in them are proportionally. However, from the two error rates (without detailed information about insertion or deletion errors), we cannot infer the precision and recall rate.

The ROC and PR curves are one-to-one mapping curves. Each point in one curve uniquely determines the confusion matrix, and thus the point in the other curve. For the ROC and PR curves, it has been shown that if a curve is dominant in one space, then it is also dominant in the other [11]. Such a relationship also holds for the ROC and DET curves. This is straightforward from the definition of these curves—true positive versus false positive in ROC curves, and miss rate (i.e., 1 − true positive) versus false positive on the scale of the normal deviate in DET curves. Since normal deviation is a monotonic function, changing the axis to a normal deviate scale preserves the property of being dominant.

3. ANALYSIS ON RT04 DATA SET

3.1. Sentence Boundary Detection Task Setup

To study the behavior of different metrics on real data, we evaluated sentence boundary detection for a state-of-the-art system on two different corpora. We used the RT04 NIST evaluation data, conversational telephone speech (CTS) and broadcast news speech (BN). The total number of words in the test set is about 4.5K in BN and 3.5K in CTS. The prior probability of a sentence boundary is different across the two corpora, about 14% for CTS and 8% for BN.[4] Comparing the two corpora allows us to investigate the effect of data skew differences on the metrics.

System output is based on the ICSI+SRI+UW sentence boundary detection system [3]. Five different system outputs are used in this study: decision tree classifiers using prosodic features only, 4-gram language model (LM), hidden Markov model (HMM) that combines prosody and language model, maximum entropy (Maxent) model using prosodic and textual information, and the combination of HMM and Maxent.[5] For all these approaches, there is a posterior probability generated for each interword boundary, which we use to plot the curves or to set the decision threshold for a single metric.

3.2. Results and Discussion

Table 2 shows different single performance measures for sentence boundary detection for CTS and BN. A threshold of 0.5 is used to generate the hard decision for each boundary point. We used the recognizer forced alignment output (slightly different from the original transcripts) as the input to sentence boundary detection. Reference boundaries were obtained by matching the original sentence boundaries to the alignment output.

Figure 1 shows the ROC, PR, and DET curves for the five system outputs, for both CTS and BN. The points shown in the PR curves correspond to using 0.5 as the decision threshold (i.e., the results shown in Table 2). The points for HMM, Maxent, and their combination are close to each other, and thus are not individually labeled with arrows.

In Table 2, for almost all the cases (except recall on CTS), the combination of HMM and Maxent achieves the best performance. The curves also show that generally HMM, Maxent, and their combination are close to each other, and much better than the other two curves for the prosody and LM, for both CTS and BN. However, in this study, our goal is not to determine the best model to optimize a single performance metric. We are more interested in looking at different system outputs and how they behave with respect to evaluation metrics.

3.2.1. Domain and metrics

BN and CTS have different speaking styles and class distributions (priors of sentence boundaries), and thus comparisons across the two domains using a single metric may not be informative. For example, the CER is similar across the two domains (for HMM, Maxent, and

[2]The scoring tool is available from http://www.nist.gov/speech/tests/rt/rt2004/fall/tools/.

[3]For DET curves, single metrics such as the EER (equal error rate) and DCF (detection cost function) are often used in speaker recognition.

[4]In the EARS program, each sentence-like unit was called an "SU" [12].

[5]Details of the modeling approaches can be found in [3].

	BN					CTS				
	Prosody	LM	HMM	Maxent	HMM+Maxent	Prosody	LM	HMM	Maxent	HMM+Maxent
NIST error rate (%)	73.86	74.31	52.58	50.21	47.87	53.94	40.22	29.42	28.38	27.78
CER (%)	6.10	6.14	4.34	4.15	3.96	7.76	5.79	4.23	4.08	4.00
Precision	0.751	0.751	0.821	0.822	0.845	0.864	0.842	0.876	0.894	0.896
Recall	0.391	0.384	0.606	0.635	0.639	0.547	0.736	0.823	0.812	0.817
F-measure	0.514	0.508	0.698	0.717	0.727	0.670	0.785	0.848	0.851	0.855
ROC AUC	0.893	0.941	0.978	0.975	0.981	0.928	0.969	0.985	0.985	0.987
PR AUC	0.601	0.652	0.804	0.815	0.832	0.791	0.878	0.929	0.934	0.938

Table 2. Different performance measures for sentence boundary detection in CTS and BN. The decision threshold is 0.5.

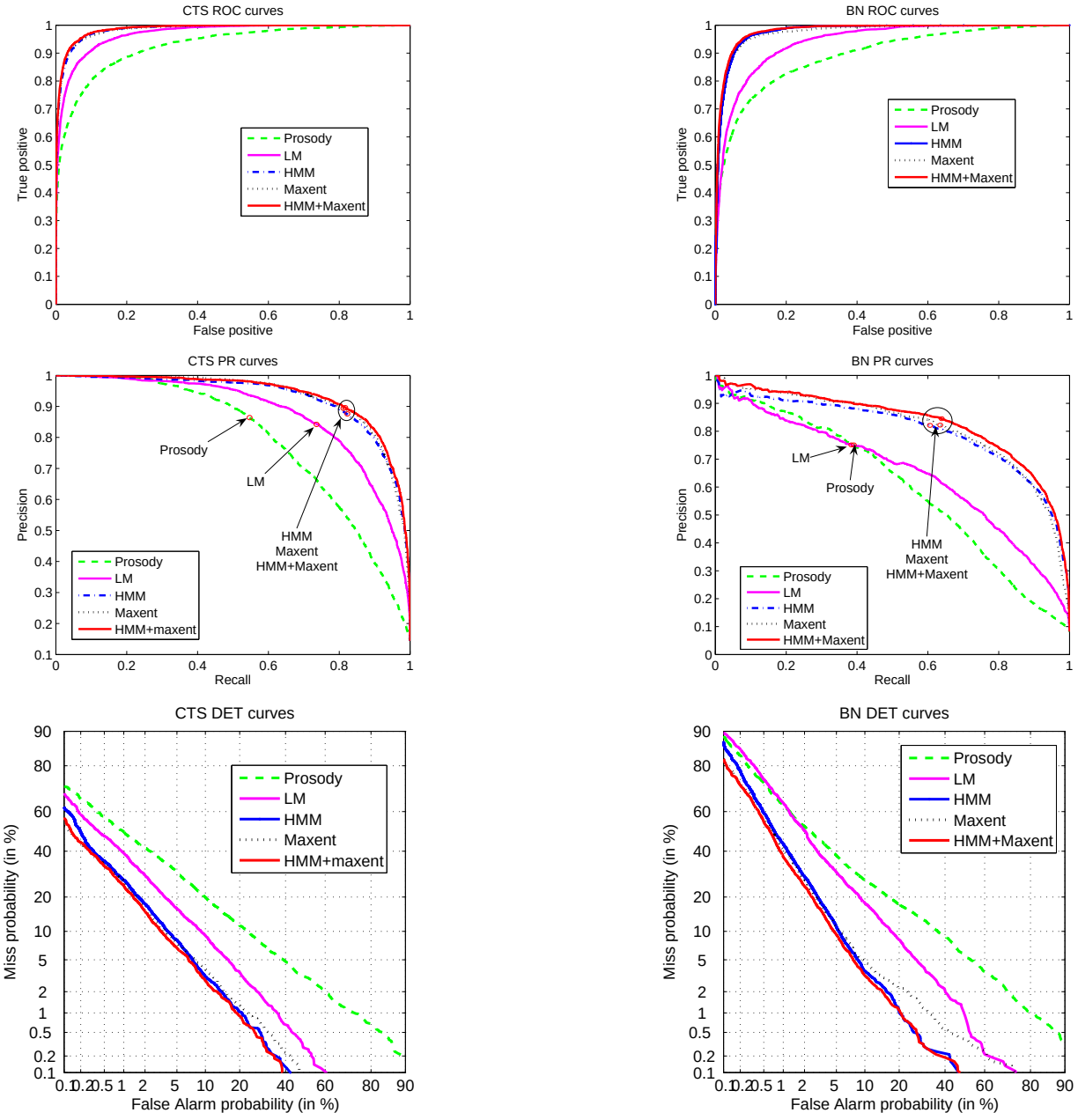

Fig. 1. ROC, PR, and DET curves for CTS and BN from five different systems: Prosody, LM, HMM, Maxent, and the combination of HMM and Maxent.

their combination), but to some extent that is because of the higher degree of skew for BN than for CTS. Using other metrics such as the NIST error rate and precision/recall can better reflect inherent performance differences. As expected, for imbalanced data sets where the majority is negative examples, ROC curves show weakness in distinguishing different classifiers or comparing across tasks, since the large number of negative examples (i.e., $tn + fp$) often results in small differences between the false positive rates. The AUC for the ROC curves is quite high for both BN and CTS, whereas in the PR space the difference between BN and CTS is more noticeable. The PR curves and the associated AUC values are much worse in BN than CTS. For the imbalanced data, PR curves often have advantages in exposing the difference between algorithms, as shown in Figure 1. DET curves also better illustrate the difference between the curves across the two corpora (e.g., the slopes of the curves).

3.2.2. Interaction among domains, models, and metrics

There is some difference between models across the two domains. For BN, using only the prosody model performs similarly to or slightly better than the LM alone, in terms of error rate, precision, and recall. However, the AUC values for the prosody model are worse than those for the LM, for both ROC and PR curves. As shown from the PR curves, in the region around the decision threshold (and also the region to the left, i.e., with lower recall), the prosody curve is better than the LM, but not in other regions. Therefore, using the curves helps to determine what model or system output is better for the region of interest. For BN, the PR curves for the prosody model and the LM cross in the middle, but this is not so for CTS, where the LM alone achieves better performance than prosody alone, using most of the metrics (except precision, as shown in Table 2).

3.2.3. Single metrics versus curves

Table 2 shows that the different measurements for this sentence boundary task are highly correlated for one corpus — an algorithm is often better than another using many single metrics. However, one single metric does not provide all the information, since it is the measure for one particular chosen decision point. As described earlier, the NIST error rate and CER cannot determine confusion matrix, or precision and recall, as they combine insertion and deletion errors (although that information can be available). For downstream processing, if a different decision region is more preferable, using the curves will easily expose such information as which model performs better. For example, [2] shows that the optimal point for parsing is different from that chosen to optimize the single NIST error rate (intuitively, shorter utterances are more appropriate for parsing).

For the PR, ROC, and DET curves, from the discussion in Section 2, we know that the dominance of a curve in one space implies dominance in other spaces. Additionally, if a curve for one algorithm is dominant over another one, then the AUC is greater. However, that the AUC of a curve is better than another does not mean that the curve is dominant. Similarly, the AUC comparison for the PR and ROC curves can be different. For example, comparing HMM and Maxent on both corpora, Maxent has better AUC in the PR space (not very significant), but not in ROC, as shown in Table 2.

In many cases, curves for different algorithms cross each other; therefore, it is not easy to conclude that one classifier outperforms the other. The decision is often based on downstream applications (e.g., improve readability, input to machine translation, or information extraction). For this situation, using both the curves, along with single value measurement is a better idea. For visualization, PR curves expose information better than ROC, especially for imbalanced data sets. DET curves are also easier to visualize than ROC curves, and more effectively show differences between algorithms.

4. CONCLUSIONS

We used a real-world spoken language processing task to compare different performance metrics for sentence boundary detection. While this study is based on a particular sentence boundary detection system and posterior probability distribution, the focus was on general comparison of alternative metrics, rather than on performance specifics. We examined single metrics, including the NIST error rate, classification error rate, precision, recall, and AUC, as well as decision curves (ROC, PR, and DET). Single metrics provide limited information; decision curves illustrate which model is best for a specific region and should be preferable for downstream language processing. Furthermore, data skew has an impact on metrics. For an imbalanced data set, PR curves generally provide better visualization than do ROC curves, for viewing differences among different algorithms. Finally, while the analysis in this paper is based on sentence boundary detection, the nature of this task is similar to many other language processing applications (e.g., story segmentation). Hence, findings should be generalizable to other similar tasks.

5. ACKNOWLEDGMENTS

We thank Mary Harper, Andreas Stolcke, Mari Ostendorf, Dustin Hillard, and Barbara Peskin for the joint work on the sentence boundary detection system used, and discussion of performance evaluation. This work is supported by DARPA under Contract No. HR0011-06-C-0023. Approved for public release; distribution unlimited.

6. REFERENCES

[1] D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman, "Measuring the readability of automatic speech-to-text transcripts," in *Proc. of Eurospeech*, 2003, pp. 1585–1588.

[2] M. Harper, B. Dorr, B. Roark, J. Hale, Z. Shafran, Y. Liu, M. Lease, M. Snover, L. Young, R. Stewart, and A. Krasnyanskaya, "Final report: parsing speech and structural event detection," http://www.clsp.jhu.edu/ws2005/groups/eventdetect/documents/final report.pdf, 2005.

[3] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14(5), pp. 1526–1540, 2006.

[4] C. Zong and F. Ren, "Chinese utterance segmentation in spoken language translation," in *Proc. of the 4th International Conference on Computational Linguistics and Intelligent Text Processing*, 2003.

[5] J. Mrozinski, E. Whittaker, P. Chatain, and S. Furui, "Automatic sentence segmentation of speech for automatic summarization," in *Proc. of ICASSP*, 2006.

[6] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. of ICASSP*, 2005.

[7] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "Toward joint segmentation and classification of dialog acts in multiparty meetings," in *Proc. of MLMI Workshop*, 2005.

[8] J. Makhoul, F. Kubala, and R. Schwartz, "Performance measures for information extraction," in *Proc. of the DARPA Broadcast News Workshop*, 1999.

[9] C. Drummond and R. Holte, "Explicitly representing expected cost: An alternative to ROC representation," in *Proc. of SIGKDD*, 2000.

[10] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. of Eurospeech*, 1997.

[11] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. of ICML*, 2006.

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