COMPARISON OF 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

Automatic detection of sentences in speech is useful to enrich speech recognition output and ease subsequent language processing modules. In the recent NIST evaluations for this task, an error rate was used to evaluate system performance. A variety of metrics such as F-measure, ROC or DET curves have also been explored in other studies. This paper aims to take a closer look at the evaluation issue for sentence boundary detection. We employ different metrics, NIST error rate, classification error rate per word boundary, precision and recall, ROC curve, DET curve, precision-recall curve, and the area under the curves, to compare different system output. In addition, we use two different corpora in order to evaluate the impact of different imbalance in the data set. We show that it is helpful to use curves as well as a single performance metric, and that different curves show different advantages in visualization. Furthermore, the data skewness also has an impact on the metrics.

Index Terms— speech processing

1. INTRODUCTION

Sentence boundary detection has received much attention recently in order to enrich speech recognition output for better readability and help subsequent language processing modules. Automatic sentence boundary detection was evaluated in the recent NIST rich transcription evaluations. In addition, studies have been conducted to evaluate the impact of sentence segmentation on downstream tasks such as speech translation, parsing, and speech summarization [1, 2, 3].

It is not clear what is the best performance metric for the sentence boundary detection task. In the NIST evaluation, system performance was evaluated using an error rate, that is, the total number of inserted and deleted boundaries divided by the number of reference boundaries. ROC curve, DET curve, and F-measure have also been used in different other studies [2, 4]. Of course, since the ultimate goal is to help 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 just for sentence boundary detection (using aforementioned NIST metric).

It helps system development to use a stand alone metric for the sentence boundary task itself. In this paper, our goal is to examine various evaluation metrics and their relationship. In addition, we evaluate the effect of different priors of the event of interest (i.e., sentence boundaries) by using different corpora. 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, such as disfluency detection, story segmentation.

The rest of this paper is organized as follows. Section 2 describes the different metrics we use and their relationship. In Section 3, we use the RT04 NIST evaluation data to analyze different measures. Summary appears in Section 4. (DELETE IF NOT ENOUGH SPACE)

2. METRICS

The task is to determine where the sentence boundaries are when given a word sequence (typically from a speech recognizer) along with the speech signal. We use the reference transcription for the study in this paper, and thus focusing on the evaluation issues and avoiding the compound effect due to speech recognition errors. We can represent this as a classification or detection task, i.e., for each word boundary, is there a sentence boundary or not?

Table 1 shows a confusion matrix and the notation we use in order to easily describe various metrics for sentence boundary detection evaluation. For a given task, the total number of samples is $tp + fp + fp + tn$, 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, i.e., sentence boundaries in this task.

2.1. Metrics

Many metrics have been used for evaluating sentence boundary detection or similar tasks, in addition to the ones examined in this study (details discussed in the following). For example, it can be evaluated for a particular downstream processing, parsing [2], machine translation [1], summarization [3]. In [5, 6], metrics are developed that treat the sentences as units and measure whether the reference and hypothesized sentences match exactly. Slot error rate [7] was introduced first for information extraction task, 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, i.e., treating system output as a 'human' annotation. There are other metrics in the general classification tasks that have not been widely used for sentence boundary detection. For example, cost curves [8] were introduced to easily show the expected cost versus the operating points. The following describes the metrics we will examine in this paper.

- 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:

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

Note that the NIST evaluation tool *mdeval*[1] allows boundaries within a small window to match up, in order to take into account the different alignments from speech recognizers. We ignore those in this study and simply treat the task as a straightforward classification task.

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

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

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

$$precision = \frac{tp}{tp + fp}$$
$$recall = \frac{tp}{tp + fn} \quad (1)$$

A single metric is often used to account for the trade off between these two:

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

- ROC curve. Receiver operating characteristics (ROC) curves are used for decision making in many detection tasks. It shows the relationship between the true positive ($=\frac{tp}{tp+fn}$) and the false positive ($=\frac{fp}{fp+fn}$) as the decision threshold varies.
- Precision-recall (PR) curve. This curve shows what happens to precision and recall as we vary the decision threshold.
- DET curve. Detection error tradeoff (DET) curve plots the miss rate (= 1 - true positive) versus the false alarms (i.e., false positive), using the normal deviate scale [9]. It is widely used in the speaker recognition task, but not so often in 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. This is used for both ROC and PR curves, but not much for the DET curves.[2]

2.2. Relationship

For a task being evaluated, 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 thus the NIST error rate and classification error rate. Each of the two error rates can uniquely determine the other one, as they are proportional. 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 [10]. 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 probability (i.e., 1 - true positive) versus false positive on the scale of the normal deviation in DET curves. Since normal deviation is a monotonic function, changing the axis to normal deviation scale still preserves the property of being dominant.

3. ANALYSIS ON RT04 DATA SET

3.1. Sentence boundary detection task setup

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 percentage of sentences[3] is different across the corpora, about 14% on CTS and 8% on BN. Comparing the two corpora allows us to investigate the effect of imbalanced data on the metrics.

System output is based on the ICSI+SRI+UW sentence boundary detection system [4]. Five different models are used in this study, prosody alone, language model (LM) alone, HMM, maximum entropy (Maxent) model, and the combination of HMM and Maxent.[4] For all these approaches, there is a posterior probability generated for each interword boundary, which we use to plot the curves or set the decision threshold for a single metric.

3.2. Analysis

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. Note that the results shown here are slightly different from those in [4], due to the difference in the practice of scoring. In addition to not using the NIST scoring tool *mdeval*, we used the recognizer forced alignment output (slightly different from the original transcripts) as the word sequence and performed sentence boundary detection upon it. The 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 models on 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 the combination of them are close to each other, and thus we did not use separate arrows for them.

In Table 2, for almost all the cases (except the recall on CTS), the combination of HMM and Maxent achieves the best performance. 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 output and how to evaluate them. 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, on both CTS and BN.

- Domain and metric

  BN and CTS have different speaking style and class distributions (priors of sentence boundaries), and thus comparisons across the two domains using some single metrics may not be informative. For example, the CER is similar across the two domains, but to some extent that is because of the higher skewness on BN than CTS. Using other metrics such as NIST

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

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

[3] In the EARS program, the sentence-like units were called "SU"s. See [11] for the definition of them in spoken language.

[4] Details of the modeling approaches can be found in [4].

	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	0.984	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.

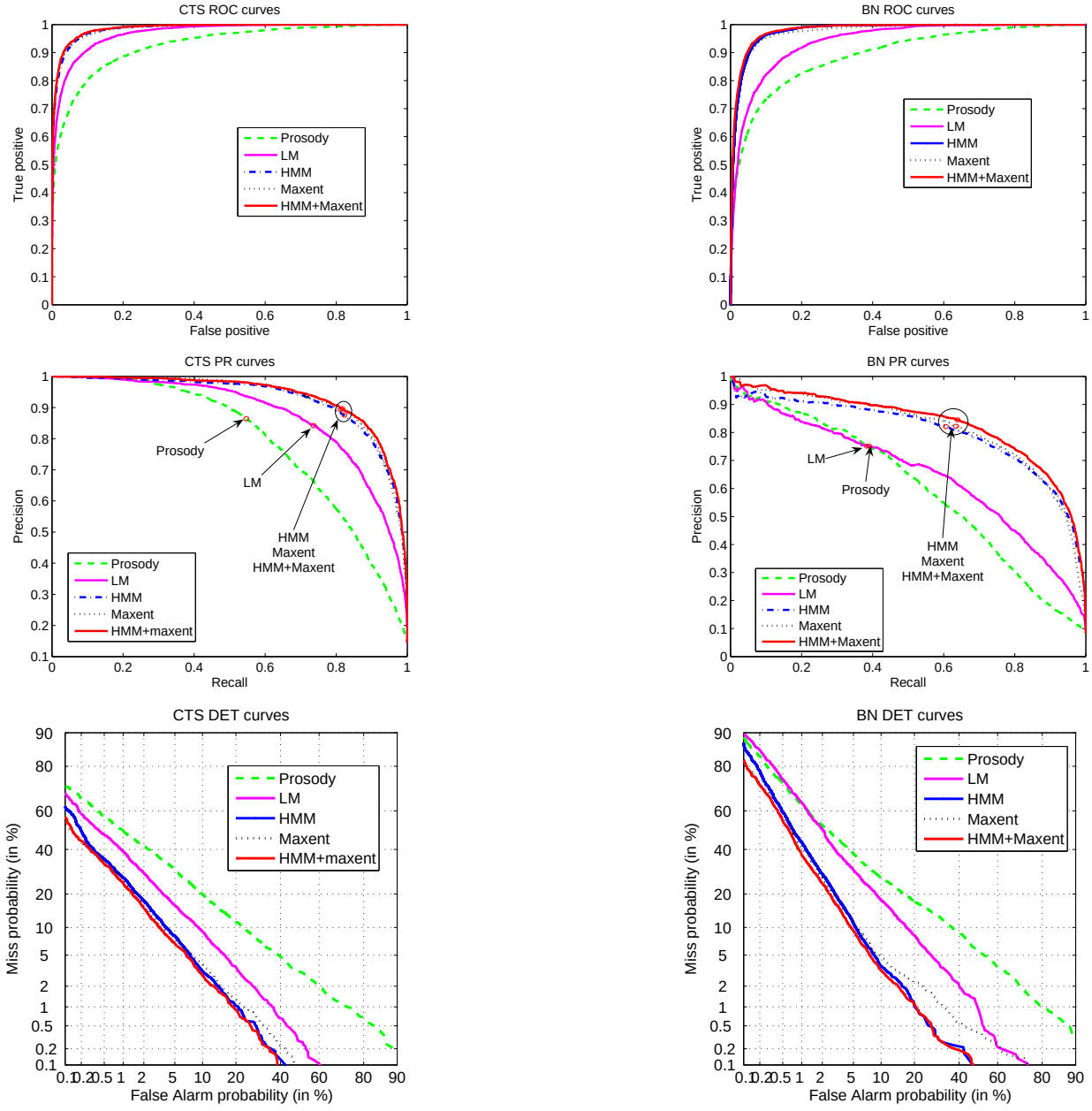

Fig. 1. ROC, PR, and DET curves for CTS for five different systems: Prosody, LM, HMM, Maxent, and the combination of HMM and Maxent.

error rate, precision/recall can better account for such data imbalance. As expected, using ROC curves for imbalanced data may hide some difference among classifiers and also between different tasks. 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. DET curves also better illustrate the difference between the curves across the two corpora (e.g., the slopes of the curves).

- Domain, models, and metrics

  There is some difference between models across the two domains. On 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 is worse than LM, for both ROC and PR curves. As shown from the PR curve, 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 LM, but not in other regions. Overall, the AUC from the prosody PR curve is worse than LM. Therefore, using the curves helps to determine what model or system output is better for the region of interest. In BN, the PR curves for the prosody model and the LM cross in the middle, but not so on CTS, where the LM alone achieves better performance than prosody using most of the measurement (except precision). The difference between models and across CTS and BN domain is also easier to observe from the DET curves than the ROC curves.

- 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. 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 in one space also means dominance in other spaces. Additionally, if a curve for one algorithm is dominant than another one, then the AUC is greater. However, that AUC is better does not mean that curves are 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 the imbalanced data set. DET curves are more easily to visualize than ROC curves and show better the difference between algorithms.

4. CONCLUSIONS

Studies on evaluation for general classification or detection tasks have been performed in machine learning. In this paper, we use a real spoken language processing task — sentence boundary detection, to compare different performance metrics. We have examined single metric including NIST error rate, classification error rate, precision, recall, and AUC, as well as decision curves (ROC, PR, and DET). The three different curves are one-to-one mapping; however, they have different advantages in visual representation. Some differences among algorithms are more visible in one curve than the others. Generally for the imbalanced data set, the PR curves provide better visualization than ROC curves. A single metric only provides limited information. It shows the performance corresponding to one decision point; whereas decision curves illustrate what model is better for a specific region and may be more preferable for downstream language processing. Note that this study is based on a particular sentence boundary detection system and its posterior probability estimation, therefore the conclusion about the models is system dependent; however the focus in this paper is rather on general analysis on system evaluation. Furthermore, even though the analysis in this paper is based on sentence boundary detection, the property of this task is similar to many other language processing applications (e.g., story segmentation), hence, the understanding of the evaluation metrics is generalizable to other similar tasks. For future work, it would be interesting to examine the different cost for different errors (MAYBE DELETE THIS SENT?).

5. ACKNOWLEDGMENT

Thanks to Mary Harper, Andreas Stolcke, Mari Ostendorf, Dustin Hillard, and Barbara Peskin for the joint work on developing the sentence boundary detection system used in this paper, and also the discussion on performance evaluation. This work is supported by DARPA under Contract No. HR0011-06-C-0023.

6. REFERENCES

[1] 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.

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

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

[4] 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*, 2006.

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

[6] 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.

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

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

[9] 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.

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

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