Dynamic n-best Selection and Its Application in Dialog Act Detection

Junling Hu, Fabrizio Morbini, Fuliang Weng
Bosch Research and Technology center
4009 Miranda Ave.
Palo Alto, CA 94304
{junling.hu, fabrizio.morbini, fuliang.weng}@us.bosch.com

Xue Liu
School of Computer Science
McGill University
Montreal, QC H3A 2A7
Canada
xueliu@cs.mcgill.ca

Abstract

We propose dynamically selecting n for n-best outputs returned from a dialog system module. We define a selection criterion based on maximum drop among probabilities, and demonstrate its theoretical properties. Applying this method to a dialog-act detection module, we show consistent higher performance of this method relative to all other *n*-best methods with fixed n. The performance metric we use is based on ROC area.

1 Introduction

Recent years have seen increasing application of machine learning in dialog systems. From speech recognizer, to natural language understanding and dialog manager, statistical classifiers are applied based on more data available from users. Typically, the results from each of these modules were sent to the next module as n-best list, where n is a fixed number.

In this paper, we investigate how we can dynamically select the number n for *n-best* outputs returned from a classifier. We proposed a selection method based on the maximum drop between two adjacent probabilities of the outputs, where all probabilities are sorted from the highest to lowest. We call this method n*-best selection, where n* refers to a variable n.

We investigated the theoretical property of n*-best, particularly its optimality relative to the fixed n-best where n is any fixed number. The optimality metric we use is ROC (Receiver Operating Characteristic) area, which measures the tradeoff of false positive and false negative in a selection criterion. We test the empirical performance of n*-best vs. n-best of fixed n for the task of identifying the confidence of dialog act classification. In two very different datasets we use, we found consistent higher performance of n*-best than n-best for any fixed n.

This paper is the first attempt in providing theoretical foundation for dynamically selecting n-best outputs from statistical classifiers. The ROC area measure has recently been adopted by machine learning community, and starts to see its adoption by researchers on dialog systems.

Even though n*-best method is demonstrated here only for dialog act detection domain, it can be potentially applied to speech recognition, POS (part-of-speech) tagging, statistical parser and any other modules that return n-best results in a dialog system.

2 Dynamically selecting n for *n*-best outputs

The *n*-best method has been used extensively in speech recognition and NLU. It is also widely used in machine translation (Toutanova and Suzuki, 2007). Given that the system has little information on what is a good translation, all potential candidates are sent to a later stage, where a ranker makes a decision on the candidates. In most of these applications, the number of candidates *n* is a fixed number. The *n*-best method works well when the system uses multi-pass strategy to defer decision to later stage.

2.1 n*-best Selection

We call *n*-best* a variant of *n*-best where *n* is a

Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 59–62,
Antwerp, September 2007.

variable, specifically the *n**-best method selects the number of classes returned from a model, such that the number *n** satisfies the following property:

$$n^* = \arg\max_n (p_n - p_{n+1}) \qquad (1)$$

where p_n and p_{n+1} are the probabilities of class *n* and class *n*+1 respectively. In other words, *n** is the cut-off point that maximizes the drop $p_n - p_{n+1}$.

2.2 Theoretical Property of n*-best

We have the following observation: When the output probabilities are ranked from the highest to the lowest, the accumulated probability distribution curve is a concave function.

We further show that our derivation of *n** is equivalent to maximizing the second derivative of the accumulative probability curve, when the number of classes approaches infinity. In other words,

$$n^* = \arg\max_n (-P''(n+1)),$$

Due to the page limit, we omit the proof here.

3 Evaluation Metric

To compare the performance of the *n**-best method to *n*-best selection of fixed n, we need to define an evaluation metric. The evaluation is based on how the n-best results are used.

3.1 The Task: Dialog Act Detection

The task we study here is described in Figure 1. The dialog-act classifier uses features computed from the parse tree of the user utterance to make predictions on the user's dialog acts.

The n-best results from the dialog-act classifier are sent to the decision component that determines whether the system is confident about the result of the classifier. If it is confident, it will pass the result to later stages of the dialog system. If it is not confident, the system will respond "I don't understand" and save the utterance for later training.

The decision on how confident we are about inter preting a sentence translates into a decision on whether to select that sentence for re-training. In this sense, this decision problem is the same as active leaning.

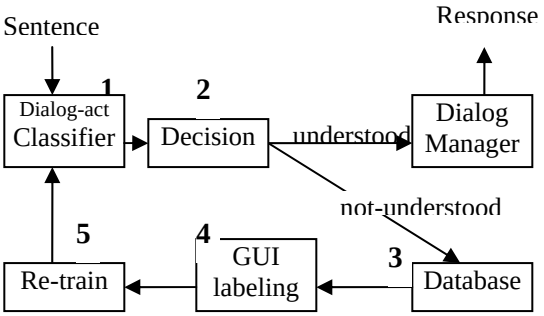

Figure 1. Detection Dialog Act with Confidence

3.2 Error Detection as Active Learning

Let *S* be the collection of data points that are marked as low confidence and will be labeled by a human. Let N_2 be the set of all new data. Let *h* be the confidence threshold and n the number we return from n-best results. We can see that (Figure 2) *S* is a function of both *n* and *h*. For a fixed *h*, the larger *n* is, the smaller *S* will be.

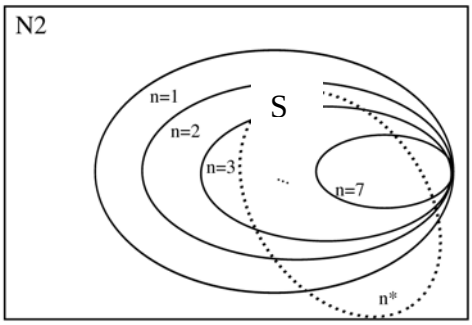

Figure 2 The Decreasing set of S as n increases

Our goal is to choose the selection criterion that produces a good *S*. The optimal *S* is one that is small and contains only true negative instances.

In active learning research, the most commonly used evaluation metric is the error rate (Tur et al, 2005; Osugi et al, 2005). The error rate can also be written as $1 - \frac{TP}{TP + FP}$, where *TP* is the number of true positives and *FP* is the number of false positives. This measure does not capture the trade off between giving the user wrong answers (false positive) and rejecting too many properly classified

user utterances (false negatives). We find a better measure that is based on ROC curve.

3.3 ROC curve and ROC Area

ROC (Receiver Operating Characteristic) curve is a graphical plot of the fraction of true positives vs. the fraction of false positive. ROC curve is an alternative to classical machine learning metrics such as misclassification rate.

An ROC space is defined by *FPR* (False Positive Rate) and *TPR* (True Positive Rate) as *x* and *y* axes respectively, where

$$FPR = 1 - \frac{TN}{FP + TN}, \quad TPR = \frac{TP}{TP + FN}$$

The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing the case in which all only true positives are returned by a particular model. The 45 degree diagonal line is called the no-discrimination line and represents the classifier that returns the same percentage of true positive and false positive.

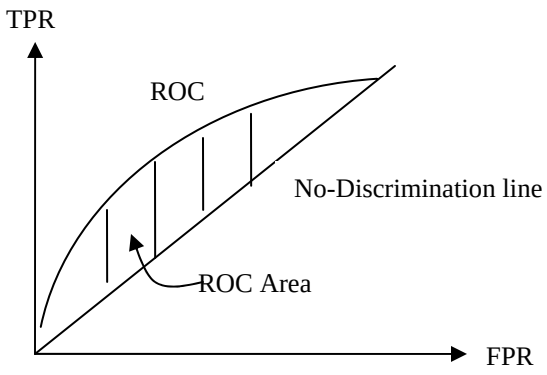

Figure 3. ROC curve and ROC area

4 Experimental Results

We tested the performance of our *n**-best method on two datasets. The first dataset contains 1178 user utterances and the second one contains 471 utterances. We use these two sets to simulate two situations: **Case 1,** a large training data and a small testing set; **Case 2,** a small training data and a large testing set.

4.1 Experimental data

All utterances in both datasets were hand labeled with dialog acts. There can be more than one dialog act associated with each utterance. An example of training instance is: "*(a cheap restaurant), (Query:restaurant, Answer, Revision)*" the first part is the user utterance, the second part (referred as L_d) is the set of human-labeled dialog acts. In total, in the domain used for these tests, there are 30 possible user dialog acts.

We compared *n**-best with fixed *n*-best methods with *n* from 1 to 6. For each of these methods, we calculate *TP*, *FP*, *TN* and *FN* for values of the threshold *h* ranging from 0.1 to 1 in steps of 0.05. Then we derived *TPR* and *FPR* and plotted the ROC curve.

Figure 4 shows the ROC curves obtained by the different methods in **Case 1**. We can see that the ROC curve for *n**-best method is better in most cases than the other methods with fixed *n*.

Figure 5 shows the ROC curves in **Case 2**, where the model is trained on a small dataset and tested on a large dataset. We can see that the ROC curves for all methods are nearer to the non-discrimination line than in the previous case. This suggests that the classifier has a lower discrimination quality given the small set used for training. However, the *n**-best method still out-performs the other *n*-best methods in the majority of scenarios.

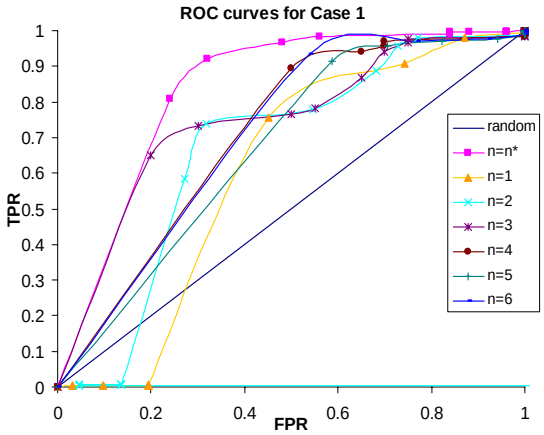

Figure 4. ROC curves from n*-best and n-best

To get a summary statistics, we calculated the size of the ROC area. Figures 6 and 7 plot the size of the ROC area of the various methods in the two test cases. We can see that *n**-best out-performs all other *n*-best methods.

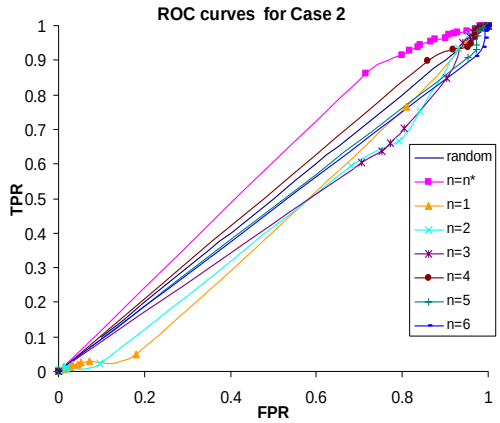

Figure 5. ROC curves obtained by n* and n-best .

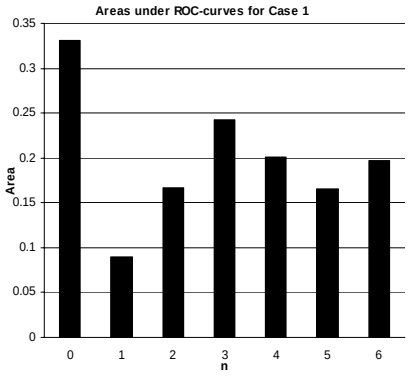

Figure 6. ROC Area for n*-best and n-best (*n** is represented as n=0)

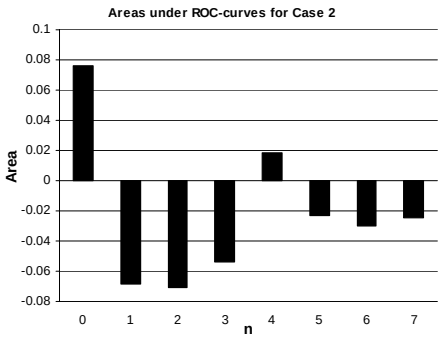

Figure 7. ROC Area for n*-best and other n-best methods (*n** is represented as n=0)

5 Conclusions

We propose dynamic selecting n for n-best outputs returned from a classifier. We define a selection criterion based on maximum drop among probabilities, and call this method *n**-best selection. We demonstrate its theoretical properties in this paper.

We measured the performance of our *n**-best method using the ROC area that has been designed to provide a more complete performance measure for classification models. We showed that our *n**-best achieved better ROC curves in most cases. It also achieves better ROC area than all other *n*-best methods in two experiments (with opposite properties).

Our method is not limited to detection of dialog acts but can be used also in other components of dialog systems.

References

C. Cortes, M. Mohri. 2004. AUC optimization vs. error rate minimization. *Advances in Neural Information Processing Systems* 16, eds., Sebastian Thrun, Lawrence Saul, and Bernhard Sch olkopf, MIT Press, Cambridge, MA.

Matt Culver, Deng Kun, and Stephen Scott. 2006. Active Learning to Maximize Area Under the ROC Curve. *Proceedings of the Sixth International Conference on Data Mining, IEEE Computer Society.* 149-158.

Sangkeun Jung, Cheongjae Lee, Gary Geunbae Lee. 2006. Dialog Studio: An Example Based Spoken Dialog System Development Workbench. 2006. *Proceedings of the Dialogs on dialog: Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*. Interspeech2006-ICSLP satellite workshop, Pittsburgh.

Thomas Osugi, Deng Kun, and Stephen Scott. 2005. Balancing Exploration and Exploitation: A New Algorithm for Active Machine Learning boundaries. *Proceedings of the Fifth IEEE International Conference on Data Mining* (ICDM'05). 330-337.

Kristina Toutanova and Hisami Suzuki. 2007. Generating Case Markers in Machine Translation. *Proceedings of NAACL-HLT 2007*, Rochester, New York. 49-56.Matt Culver, Deng Kun, and Stephen Scott. 2006. Active Learning to Maximize Area Under the ROC Curve. *Proceedings of the Sixth IEEE International Conference on Data Mining*. 149-158.

Gokhan Tur, Dilek Hakkani-Tür and Robert E.Schapire. 2005. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171-186.