Weight Estimation for N-Best Rescoring*

Ashvin Kannan†, *Mari Ostendorf*†, *J. Robin Rohlicek*‡

† Boston University
44 Cummington St.
Boston, MA 02215

‡ BBN Inc.
10 Moulton St.
Cambridge, MA 02138

ABSTRACT

This paper describes recent improvements in the weight estimation technique for sentence hypothesis rescoring using the N-Best formalism. Mismatches between training and test data are also explored.

1. INTRODUCTION

The N-Best rescoring paradigm involves the generation of a list of the N best sentence hypotheses by a recognition system and the subsequent rescoring of these hypotheses by other knowledge sources. The sentence hypotheses are then reranked according to a weighted linear combination of the different scores. This paradigm has the potential of achieving better performance than that of any individual knowledge source, if these scores are combined in an "optimal" manner. This paper discusses the key issues related to estimation of robust weights for a linear combination of scores.

2. WEIGHT ESTIMATION

In the initial work [1], the weights used in the linear score combination were chosen to minimize the generalized mean of the rank of the correct hypothesis using an iterative search algorithm based on Powell's method [2]. Further experience using this technique suggested that the result was very sensitive to the large number of local minima in the optimization criterion.

Several steps have been taken to address this issue. The optimization criterion now minimizes the average word error in the top ranking hypothesis. The use of this criterion results in a "smoother" weight space, i.e., having fewer local minima. Also addressing the problem of local minima, we examine a large number of points in the weight space on a lattice spanning the range of probable weights. Powell's method may be used with points on the grid as the initial estimate of weights to find the best performance, or the points on a fine grid may be evaluated directly.

The error function is piece-wise constant over the weight space. A particular ranking of the hypotheses corresponds to a region (cell) defined by a set of inequalities that describe a polytope. In the hope of obtaining a more robust estimate, we measure the amount of slack for the different coefficients along the coordinate axes such that the weight remains within the cell as well as determine the "center" of the cell. The product of the slacks in the different coordinate directions at the "center" is an approximate indicator of the "volume" of the cell. If more than one cell gives the same performance, we choose the one with the largest "volume". Weights which correspond to the "center" of this cell are used for combining scores in the test set.

3. EXPERIMENTS

Experiments were conducted to gain a better understanding of the weight space. In our implementation of the N-Best rescoring paradigm [1], the N-Best list (N = 20) is generated by the BBN BYBLOS system [3]. This list is rescored by the BU system, which is based on the stochastic segment model (SSM) [4, 5], a statistical model for the sequence of observations that comprise a phoneme segment. The SSM models are based on independent-frame assumptions, are gender-dependent and are context-dependent with context tying based on automatic clustering. Results are reported on the speaker-independent Resource Management corpus using the Word-Pair grammar. The weights were trained on the Feb 89 test set and then used to combine scores for the Oct 89 test set. The training of weights may be either gender-dependent or gender-independent.

Figure 1 and Figure 2 show contour plots for the word error distribution as a function of normalized HMM and SSM scores on the two test sets, keeping the phoneme and word insertion penalties fixed at typical values. The contours have been drawn for the ten lowest word errors, with intensity being inversely proportional to error. The HMM and SSM scores were normalized by the average of the respective scores for the correct sentences to better illustrate their relative weight in the combined score.

Figure 1 represents the case for gender-dependent op-

*This research was jointly funded by NSF and DARPA under NSF grant number IRI-8902124.

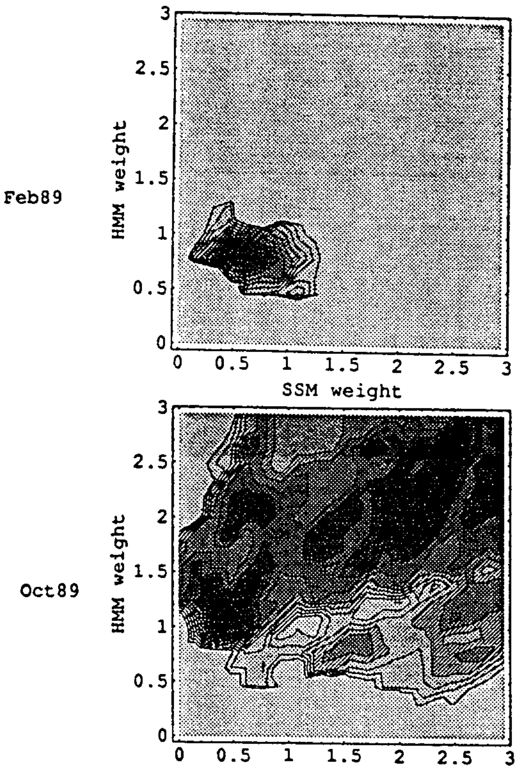

Figure 1: Error function for optimization over male speakers. Range: 2.9-3.6% (Feb89), 2.8-3.3% (Oct89).

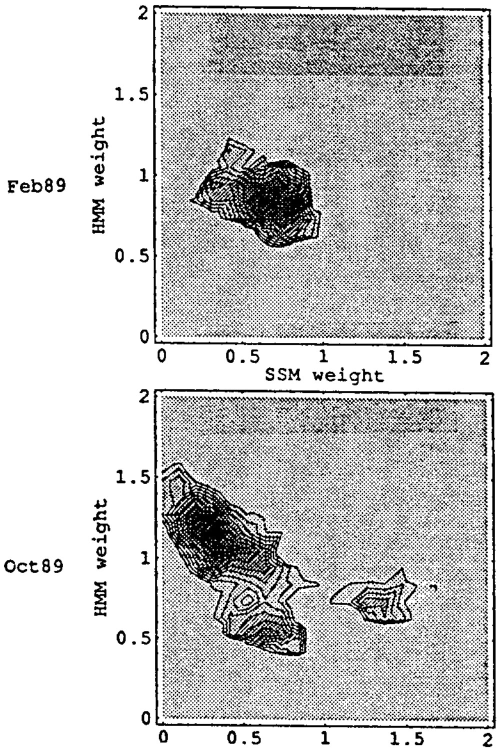

Figure 2: Error function for gender-independent optimization. Range: 2.8-3.2% (Feb89), 3.2-3.6% (Oct89).

timization over male speakers. The weight space for the two test sets appears vastly different. The effects of gender-independent optimization is shown in Figure 2. Though the Oct 89 figure has fewer local optima, it must be noted that the best region for one test set still does not match that of the other. Normalizing the acoustic scores shows that the HMM is weighted higher than the SSM, but the weights are of the same order of magnitude. The word vs. phoneme count contours (not shown) suggest that typical values of the word penalty are about 3-5 times that of the phoneme penalty.

Our current word recognition results on the Feb 89 test set are 4.2% for SSM and 2.8% for the combined system (HMM-SSM) using weights estimated on this test set. Using the same weights and testing on the Oct 89 test set, the results are 4.8% for the SSM and 3.3% for the combined system. Combining the SSM with the BBN HMM yields a 13% reduction in error rate over the HMM performance alone which was 3.8%.

4. DISCUSSION

In summary, we have described techniques that alleviate the problem of sensitivity to local optima in weight estimation for N-Best rescoring. However we find that there still exists a significant problem of mismatch between training and test sets. By comparing the contour plots we see that gender-independent optimization seems to be less sensitive to mismatch. This leads us to believe that we must estimate weights over a larger set of speakers.

References

1. Ostendorf, M., Kannan, A., Austin, S., Kimball, O, Schwartz, R., Rohlicek, J. R., "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses," *Proceedings of the DARPA Workshop on Speech and Natural Language*, pp. 83–87, February 1991.
2. Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T., *Numerical Recipes* , Cambridge University Press, Cambridge 1986.
3. Schwartz, R., and Austin, S., "Efficient, High Performance Algorithms for N-Best Search", *Proceedings of the Third DARPA Workshop on Speech and Natural Language*, pp. 6–11, June 1990.
4. Ostendorf, M., and Roukos, S., "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. on Acoust., Speech and Signal Processing*, pp. 1857–1869, December 1989.
5. Roukos, S., Ostendorf, M., Gish, H., and Derr, A., "Stochastic Segment Modeling Using the Estimate-Maximize Algorithm," *Proceedings IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 127–130, New York, New York, April 1988.