Confidence Based Lattice Segmentation and Minimum Bayes-Risk Decoding

Vaibhava Goel
I.B.M. T.J. Watson Research Center,
Yorktown Heights,
NY 10598, USA
vgoel@us.ibm.com

Shankar Kumar and William Byrne
CLSP/Johns Hopkins University,
3400 N. Charles St.,
Baltimore, MD 21218, USA
{skumar,byrne}@jhu.edu

Introduction

- Minimum Bayes Risk (MBR) decoders have shown improvements over MAP decoders using N-best List Rescoring and A^* search over word lattices.
- Segmental MBR decoders simplify implementation of MBR decoders by segmenting the N-best lists or Lattices.
- We present:
  - A method to segment Word Lattices into regions of low confidence and high confidence.
  - Three SMBR procedures that can be applied on low confidence segment sets.
  - Results on the SWITCHBOARD speech corpus showing error rate improvements with SMBR decoding.

1 Minimum Bayes-Risk Speech Recognizers

- Minimum Bayes-risk (MBR) decoder on an utterance A:

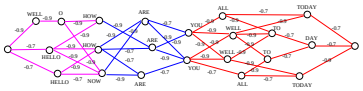
$$\delta(A) = \underset{W' \in \mathcal{W}_h}{\operatorname{argmin}} \sum_{W \in \mathcal{W}_e} l(W, W') P(W|A)$$

- A word lattice

- N-best rescoring [1] or lattice based A^* [2] implementations.

2 Segmental Minimum Bayes-Risk (SMBR) Decoders

- A segmented word lattice



- We want to obtain a segmentation such that :

$$l(W, W') = l(W_1, W_1') + l(W_2, W_2') + l(W_3, W_3')$$

- *Segmental MBR decoder is a concatenation of MBR decoders* :

$$\delta(A) = \begin{cases} \underset{W' \in \mathcal{W}_1 W \in \mathcal{W}}{\operatorname{argmin}} \sum l(W, W') P_1(W|A) \\ \underset{W' \in \mathcal{W}_2 W \in \mathcal{W}_2}{\operatorname{argmin}} \sum l(W, W') P_2(W|A) \\ \underset{W' \in \mathcal{W}_3 W \in \mathcal{W}_3}{\operatorname{argmin}} \sum l(W, W') P_3(W|A) \end{cases}.$$

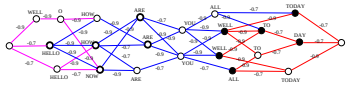
3 Issues in the SMBR Approach

- Primary benefit of SMBR is reduction in search space: A large search problem is broken down into many smaller search problems.
- SMBR decoder generates hypotheses not present in original lattice.
- SMBR involves tradeoff between search errors (no segmentation) and errors in loss function approximation.
- We have developed Lattice Cutting as a procedure that segments a word lattice into low confidence and high confidence regions.
- Segmentation restricts the possible string alignments :
  - We want the induced loss due to the segmentation to approximate the desired loss function.

$$l(W, W') \approx \sum_{i=1}^{N} l(W_i, W_i')$$

4 A Lattice Cutting Method Based on Node Sets

- Cutting a lattice based on N_s and N_e



○ : Node of N_s
● : Node of N_e

- How to select Node Sets N_s and N_e ?
  - Along each path in $\mathcal{W}$, there needs to be at least one node in N_s and one node in N_e
  - Nodes of $\mathcal{N}_s$ precede nodes of $\mathcal{N}_e$ on that path
- Need to compute marginal probabilities of paths in the lattice segment $\mathcal{W}_i$
  - Marginal probabilities should be computed over longest subpaths in $\mathcal{W}_i$

5 Marginal Probabilities of Paths in Lattice Cuts

- *Goal*: To compute $P_i(W_i|A)$, the marginal probability of word string W_i in Lattice segment $\mathcal{W}_i$
- Lattice Forward Probability of the first node n_1 of W_i, $n_1 \in N_s$:

$$F(n_1) = \sum_{W_p : E^f(W_p) = n_1} e^{L_f(W_p)}$$

Notation:
  - W_p is a prefix of a lattice path W
  - $E^f(W_p)$: Last Node of W_p
  - $L_f(W_p)$: Joint acoustic and language model log-likelihood of W_p
  - $E(W_p)$: All lattice nodes through which W_p passes
- Define a Restricted Lattice Forward Probability of node n_1

$$F(n_1; N_s) = \sum_{\substack{W_p \,:\, E^f(W_p) = n_1 \\ E(W_p) \cap N_s = \{n_1\}}} e^{L_f(W_p)}$$

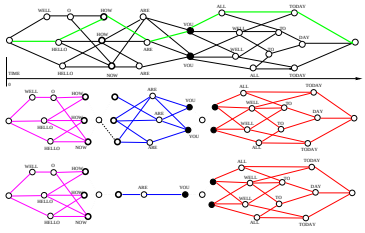
  - For a node $n_1 \in N_s$, Lattice Paths that pass through a node of N_s before n_1 contribute a segment longer than W_i to $\mathcal{W}_i$. Their probabilities are to be excluded.
- Lattice Backward Probability of the final node n_2 of W_i, $n_2 \in N_e$

$$B(n_2) = \sum_{W_s \,:\, E^b(W_s) = n_2} e^{L_b(W_s)}$$

Notation:
  - W_s is a suffix of a lattice path W
  - $E^b(W_s)$ is the starting node of W_s
  - $L_b(W_s)$ is the joint acoustic and language model log-likelihood of W_s conditioned on starting node of W_s.
- Marginal Probability of string W_i in section i:

$$P_i(W_i|A) = \frac{F(n_1; N_s) P(W_i, A(W_i)|n_1) B(n_2)}{P(A)}$$

6 A Confidence Based Lattice Cutting Procedure



7 SMBR decoding of Lattice Cuts

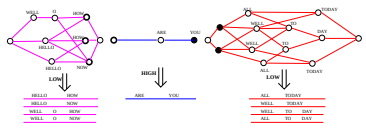
- Three MBR decoding procedures on lattice cuts
  - A^* search
  - N-best list rescoring
  - Lattice Cutting E-ROVER (LCER)
- For A^* search and N-best List rescoring, MBR decoding on a lattice cut $\mathcal{W}_i$:

$$\hat{W}_i = \underset{W_i' \in \mathcal{W}_i}{\operatorname{argmin}} \sum_{W_i \in \mathcal{W}_i} l(W_i, W_i') P_i(W_i|A).$$

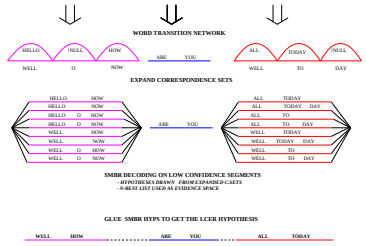
- For LCER, the search and the sum are over different spaces

8 Lattice Cutting E-ROVER

- *Evidence Space Generation*



- *Hypothesis Space Generation*



9 Results on SWITCHBOARD

- Johns Hopkins University Hub-5 LVCSR 2001 Evaluation System.
  - Test sets: Swbd2-portion of 1998 evalset (swbd2-98), Swbd-1 portion of 2000 evalset (swbd1-00) and 2001 evalsets (swbd1-01,swbd2-01,swbd2cell-01)
  - SMBR Methods : N-Best List Rescoring and E-ROVER

Decoding Strategy		WER				
		swbd2-98	swbd1-00	swbd1-01	swbd2-01	swbd2cell-01
MAP (baseline)		41.1	26.0	26.8	32.4	37.7
N-best rescoring	Entire lattice	40.4	25.6	26.6	32.1	37.2
	Lattice cutting	40.2	25.4	-	-	-
e-ROVER	Entire lattice	40.5	25.7	-	-	-
	Lattice cutting	40.0	25.2	26.5	31.5	36.8

- Results Not Included in JHU 2001 Evaluation
  - Test Set: The 1997 JHU LVCSR workshop test set
  - SMBR Procedures : N-Best List Rescoring and A^* decoding

Decoding Strategy		WER
MAP (baseline)		38.5
N-best rescoring	Entire lattice	37.9
A^* search	Entire lattice	37.5
	Lattice cutting	37.1

10 Conclusions and Future Work

- A lattice cutting procedure that segments lattices into regions of low and high confidence
- Three procedures for MBR decoding on segmented lattices
- Performance of MBR procedures improves when applied to lattice cuts
- Ongoing Research
  - Improved lattice segmentation: not relying on time information
  - A^* search procedures with separate hypothesis and evidence spaces

[1] A. Stolcke, Y. Konig, and M. Weintraub, "Explicit word error minimization in N-best list rescoring," Eurospeech-97, pp. 163–165, Rhodes, Greece, 1997.
[2] V. Goel and W. Byrne, "Minimum Bayes-risk automatic speech recognition," Computer Speech and Language, Vol. 14(2), pp. 115–135, 2000.