RISK-BASED LATTICE CUTTING FOR SEGMENTAL MINIMUM BAYES-RISK DECODING

Shankar Kumar and Bill Byrne

Center for Language and Speech Processing, Johns Hopkins University,
3400 N. Charles St.,
Baltimore, MD 21218, USA
{*skumar,byrne*}*@jhu.edu*

I. Introduction

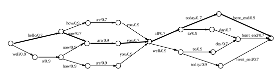
- Minimum Bayes Risk (MBR) decoders improve upon MAP decoders by directly optimizing loss function of interest: Word Error Rate
- MBR decoding is expensive when the search spaces are large
- Segmental MBR (SMBR) decoding breaks the single utterance-level MBR decoder into a sequence of simpler search problems.
  - To do this, the N-best lists or lattices need to be segmented
- We present: *A* new lattice segmentation strategy based on a risk criterion

II. Segmental Minimum Bayes-Risk Decoding

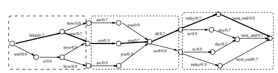
- Minimum Bayes-risk (MBR) decoder on an utterance A:

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

- MBR Implementations: N-best rescoring, lattice based A^* search and lattice based extended ROVER (e-ROVER)
- A word lattice



- Suppose we can segment this word lattice :



- This segmentation induces a loss function between any two word strings:

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

- A good segmentation is such that the true loss function is the same as the induced loss function :

$$l(W, W') = l_I(W, W')$$

- Under such a segmentation, MBR decoder reduces to a concatenation of MBR decoders:

$$\delta(A) = \begin{cases} \operatorname*{argmin}_{W' \in \mathcal{W}_1} \sum_{W \in \mathcal{W}_1} l(W, W') P_1(W|A) \;\cdot \\ \operatorname*{argmin}_{W' \in \mathcal{W}_2} \sum_{W \in \mathcal{W}_2} l(W, W') P_2(W|A) \;\cdot \\ \operatorname*{argmin}_{W' \in \mathcal{W}_3} \sum_{W \in \mathcal{W}_3} l(W, W') P_3(W|A) \end{cases}$$

III. Ideal Lattice Segmentation

- *Goal:* Reduce search space of MBR recognizer
  - Pruning the lattice could result in search errors
  - Segmentation breaks up a single search problem into many simpler search problems
- Any segmentation restricts string alignments. $l(W, W') \leq \sum_{i=1}^{N} l(W_i, W_i')$
- Therefore, segmentation involves tradeoff between search errors and errors in approximating the loss function

IV. Risk Criterion for Lattice Segmentation

- Total Bayes-Risk of all lattice word strings

$$R_T = \sum_{W' \in \mathcal{W}} \sum_{W \in \mathcal{W}} l(W, W') P(W|A)$$

- $\tilde{W} = \tilde{w}_1^K$ is the MAP string in the lattice
- ML approximation to Total Bayes-Risk

$$R_T \approx \tilde{R}_T = \sum_{W' \in \mathcal{W}} l(\tilde{W}, W') P(\tilde{W}|A)$$

- After Lattice Segmentation

$$\tilde{R}_T \leq P(\tilde{W}|A) \sum_{W' \in \mathcal{W}} \sum_{i=1}^{N} l(\tilde{W}_i, W_i')$$

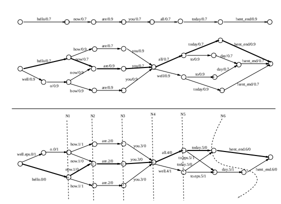
- Modified Segmentation Criterion: Minimize the upper bound on $\tilde{R}_T$

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

- This is lattice segmentation wrt the MAP hypothesis

V. Levenshtein Alignment via Weighted Finite State Transducers

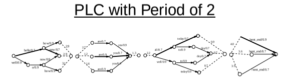
- *Goal*: Optimal alignment under Levenshtein distance between $W' \in \mathcal{W}$ and $\tilde{W}$
- We have developed an efficient Weighted Finite State procedure that computes the Levenshtein alignment between $\tilde{W}$ and all lattice word strings.
  - Algorithmic details in the paper
  - The procedure tags every lattice link with an index i along the best path: $i \in \{1, 2, ..., K\}$



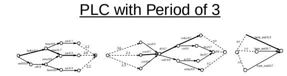
VI. Risk-Based Lattice Cutting

- *Risk-Based Lattice Cutting (RLC)*
  Segment the lattice into K segments based on the Levenshtein alignment
- *Periodic Risk-Based Lattice-Cutting (PLC)*
  Segment Lattice into $< K$ segments by choosing node sets at equal periods
  - Higher Period : Better approximation to the Levenshtein distance
  - Higher Period : More Search errors
  - RLC : PLC with a period of 1.

PLC with Period of 2



PLC with Period of 3



VII. Results on SWITCHBOARD

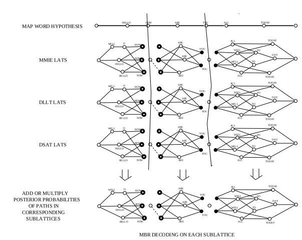
- Johns Hopkins University LVCSR Hub-5 2001 Evaluation System.
- Test sets: Swbd2-portion of 1998 evalset (SWB2), Swbd-1 portion of 2000 evalset (SWB1)

Decoding Strategy		WER(%)	
		SWB2	SWB1
MAP (baseline)		41.1	26.0
N-best rescoring	Entire lattice	40.4	25.6
	RLC	41.0	25.9
	PLC	40.1	25.4
A^* search	Entire lattice	40.4	25.5
	RLC	41.0	25.9
	PLC	40.0	25.4
e-ROVER	Entire lattice	40.5	25.7
	RLC	40.5	25.7
	PLC	39.9	25.3

- SMBR decoding on segmented lattices performs better than MAP or MBR decoding on unsegmented lattices
- SMBR decoding performs better under PLC than under RLC

VIII. An Application to Multiple-System Lattice Combination

Investigate ASR System Combination Strategies using Lattices



Johns Hopkins University RT-02 LVCSR Evaluation System

Decoding Strategy	WER(%)		
	SWB1	SWB2	SWB2C
MAP			
MMIE - 1best	24.5	39.2	39.6
DLLT - 1best	24.0	38.7	38.8
DSAT - 1best	24.5	39.3	39.5
Intersect - 1best	24.0	38.4	38.7
Sublattice-Based			
Intersect - SMBR	23.5	37.8	38.0
Union- SMBR	23.3	37.8	37.8

- SMBR decoding is better than simply intersecting lattices and rescoring
- Adding posteriors of hypotheses over sub-lattices is better than multiplying them (Using a conditional independence assumption)

IX. Conclusions

- A lattice cutting procedure based on a risk criterion
  - Segment the lattice wrt the MAP hypothesis
  - No time information or likelihoods required from the lattice
- Performance of MBR procedures improves when applied to lattice cuts
- PLC procedure performs better than RLC procedure
  - Proper tradeoff between Levenshtein distance approximation and search errors is crucial
- Applications of Lattice Segmentation
  - Multiple-system lattice combination via sub-lattice combination
  - Confidence estimation for recognized word strings