Minimum Hypothesis Phone Error as a Decoding Method for Speech Recognition

Haihua Xu,Daniel Povey

August 28, 2009

Maximum a Posteriori (MAP)

The standard decoding formula normally used in speech recognition is Maximum A Posteriori (MAP) as follows:

$$\begin{aligned} W^* &= \mathbf{argmax}_W P(W|\mathcal{O}) \\ &= \mathbf{argmax}_W P(W) p(\mathcal{O}|W) \end{aligned}$$

Known limitation

Such a criterion guarantees a sentence that minimizes sentence error can be decoded; however, it is usually the Word Error Rate (WER) not the sentence error used as the evaluation criterion for the recognition system performance. To make up for this mismatch, Minimum Bayes Risk (MBR) criterion is a natural alternative.

Minimum Bayes Risk (MBR)

$$W^* = \textbf{argmin}_{W_i} \sum_{j=1}^{N} P(W_j|\mathcal{O})E(W_i|W_j)$$

where $E(W_i|W_j)$ is the number of errors (Levenshtein distance) given W_i as a reference.

Problems

Direct calculating the criterion in a subspace of W(generally represented as word graph/lattice) is prohibitive, thus many approximated strategies are attempted.

Known approaches to WER minimization

- N-best sentence list based decoding scheme proposed by A.Stockle et al.
- Consensus network proposed by L.Mangu et al.
- Time-frame word error proposed by F.Wessel et al. al.

MPE/MWE as a criterion for lattice rescoring

We approximate MBR with Minimum Phone Error (MPE) discriminative training criterion as decoding criterion, the approach is

$$W^* = \mathbf{argmax}_W \sum_{W'} P^\kappa(W'|\mathcal{O}) Acc(W'|W)$$

where $\mathbf{argmax}_W$ is taken over an N-best list that we derive from the decoding lattice, and so is $\sum_{W'}$. In other words we find the hypothesis W^* to maximize the objective criterion.

Advantages of the proposed method

- Explicitly optimizing the objective criterion, the correctness of which has been proofed by MPE/MWE discriminative training.
- Conceptually simple and clear, as simplified forward-backward algorithm is performed on the decoding lattice.
- Much flexible on the accuracy criterion $Acc(W'|W)$ calculating, such as on phone error, time-frame phone error, time-frame word error, and word error criteria can be implemented under the same framework.

How large N for a desired W^*? (1)

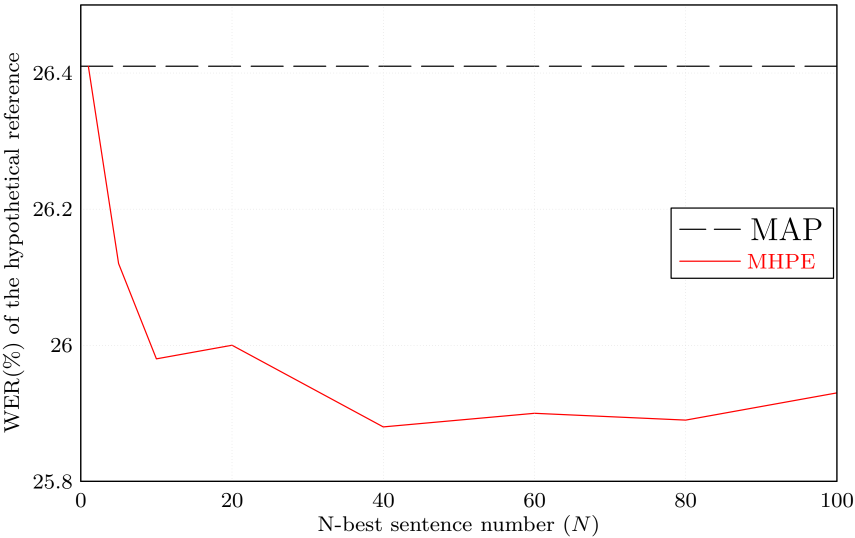

Figure: WER versus N on MSRA data (trained with MPE)

How large N for a desired W^*? (2)

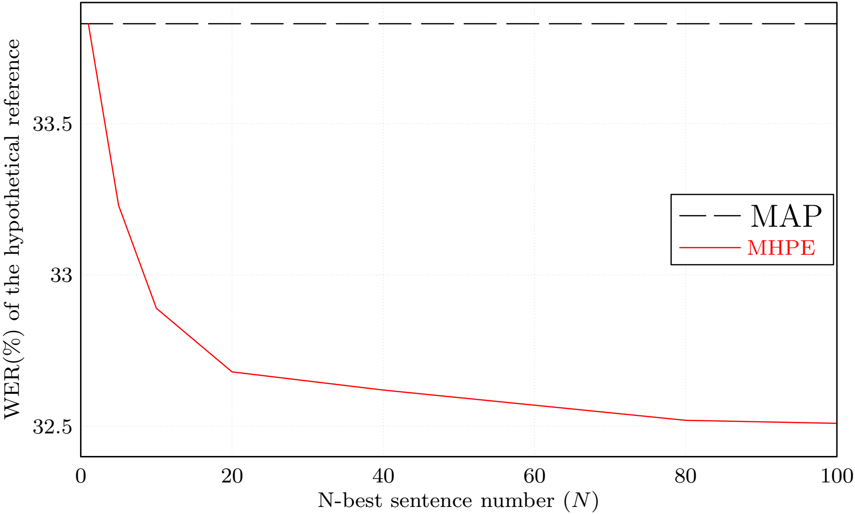

Figure: WER versus N on broadcast data(trained with MAP+MPE)

How large N for a desired W^*? (3)

As illustrated from the figures, the desired WER can be gained when N is ranged from 20 to 40. Therefore with a very limited N, we can approximate WER minimization. Similar experiments has also been performed on English test data, and we reach the same conclusions.

Experimental results

Recognition system (A) trained with the MLE criterion

Table: Baseline vs. MHPE on MSRA test data.

Methods	MSRA (MLE, N=40)				
	#ins	#sub	#del	SER	WER
Base	24	2333	171	95.40%	26.41%
MHPE	51	2319	107	95.40%	25.88%
Δ	+27	-12	-64	-0.0%	-0.53%

Experimental results

Recognition system (B) trained with the MAP criterion

Table: Baseline vs. MHPE on BDC test data.

Methods	BDC (MAP,N=40)				
	#ins	#sub	#del	SER	WER
Base	114	7065	963	98.02%	33.83%
MHPE	187	7014	651	98.18%	32.62%
Δ	+73	-51	-312	+0.16%	-1.21%

Experimental results

Recognition system (C) trained with the MPE criterion

Table: Baseline versus MHPE on MSRA test data.

Methods	MSRA (MPE, N=40)				
	#ins	#sub	#del	SER	WER
Base	26	2074	183	94.60%	23.85%
MHPE	47	2035	108	93.40%	22.90%
Δ	+21	-39	-75	-1.20%	-0.95%

Experimental results

Recognition system (D) trained with MAP+MPE criterion

Table: Baseline versus MHPE on BDC test data.

Methods	BDC (MPE,N=40)				
	#ins	#sub	#del	SER	WER
Base	109	6296	959	96.45%	30.60%
MHPE	178	6254	568	96.20%	29.08%
Δ	+69	-42	-391	-0.25%	-1.52%

Experimental results

MHPE versus Consensus Network on lattice decoding

Table: MHPE versus CN

Test sets	Base	CN	MHPE
MSRA(MLE)	26.41%	25.92%	25.88%
BDC(MAP)	33.83%	32.80%	32.62%
MSRA(MPE)	23.85%	23.42%	22.90%
BDC(MPE)	30.60%	29.41%	29.08%

Conclusions and future work

We have introduced a new decoding method for lattice rescoring that aims to get closer to the Minimum Bayes Risk decision rule with respect to the Word Error Rate. Future work will be focused on

- To have a full comparison on other criteria to implement $Acc(W_i, W)$.
- More sophisticated approach will be investigated to take the place of N-best sentence list.
- Based on the proposed criterion, new system combination scheme will be studied, not just Confusion Network Combination.

Thanks !