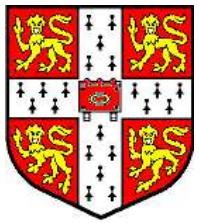
CU AGILE MT Progress and Plans

CU MT Team

August 2007



Cambridge University Engineering Department

MT System Combination (1)

- Confusion network decoding widely used in MT system combination:
  - previous methods use single reference to construct network
  - system weights not considered
  - unreliable consensus and combination performance poor
- Improving system combination using multiple confusion networks:
  - each system's 1-best hypothesis used for alignment in turn
  - system weights represented as subnetwork priors
  - flexibly incorporating other ranking information
- TER reduction of 2.0%-3.5% compared with using single reference
- Consistently outperforming manually selected best system

MT System Combination (b1)

- Confusion network decoding widely used in MT system combination:
  - selection of alignment skeleton crucial
  - previous methods use single reference to construct network
  - information of system weights largely ignored
  - unreliable consensus and combination performance poor
- Alternatively multiple confusion network decoding may be used:
  - each system's 1-best hypothesis used for alignment in turn
  - system weights represented as subnetwork priors
  - tunable and related to translation performance
  - final hypothesis extracted from combined network
  - other ranking information, e.g. LM, flexibly incorporated
- Significant reduction of TER

MT System Combination (b2)

System	TER/BLEU			
	mt02_05_test	ng_y1q4_test	mt06_nw	mt06_ng
bbn-class	55.53/35.76	74.09/13.39	63.65/15.82	68.13/14.24
bbn-hd	51.55/42.85	75.04/14.55	60.89/18.76	66.98/13.08
cu-1	57.75/32.55	76.86/12.01	65.33/14.70	72.03/13.00
isi-hiero	54.13/38.47	73.59/14.35	61.15/18.05	66.61/14.69
isi-pbmt	57.66/33.65	74.54/12.60	64.82/15.70	67.46/14.28
isi-sbmt	50.43/44.22	71.95/15.68	61.02/18.15	65.91/15.72
conf	52.35/41.99	72.64/15.02	61.30/18.75	64.79/15.85
mconf	49.56/42.82	69.21/15.46	59.05/18.68	62.70/15.72

Combination performance using AGILE dry-run systems N-best outputs on Chinese text sets.

MT System Combination (2)

- CN decoding tends to over-generate new sentences
- Introduces N-gram breaking points - hurts BLEU!
- Improved hypothesis selection using multiple CN decoding output
- More balanced TER and BLEU performance
- Gains up to 1.2% in TER, 0.5% in BLEU over previous selection method

MT System Combination (b3)

System	TER/BLEU			
	bcmd05	bnmd06	bcme06	bnme06
conf	67.58/14.94	63.98/17.89	72.88/9.73	69.43/12.65
confsel	68.56/15.58	65.13/18.39	73.86/10.21	70.04/13.22
mconf	67.10/15.42	63.26/18.84	72.68/10.28	68.84/13.13
mconfsel	68.15/15.88	64.40/18.98	72.77/10.39	69.62/13.31

Combination performance using AGILE dry-run systems N-best outputs on dry-run Chinese audio sets

Improving Minimum Bayes Risk Decoding

- Minimum Bayes risk decoding requires cost function computation
  - metrics affected by length penalty may be non-robust and noisy
  - expected risk unreliable and poor hypotheses ranking
- Improved BLEU cost function computation for MBR combining:
  - expected N-gram precision computed against the current reference
  - expected sentence length ratio also computed
- Computationally efficient and fast in decoding
- Significantly TER reduction of 0.5% to 1.1%
- Different N-best ranking - useful for system combination

Improving Minimum Bayes Risk Decoding (b1)

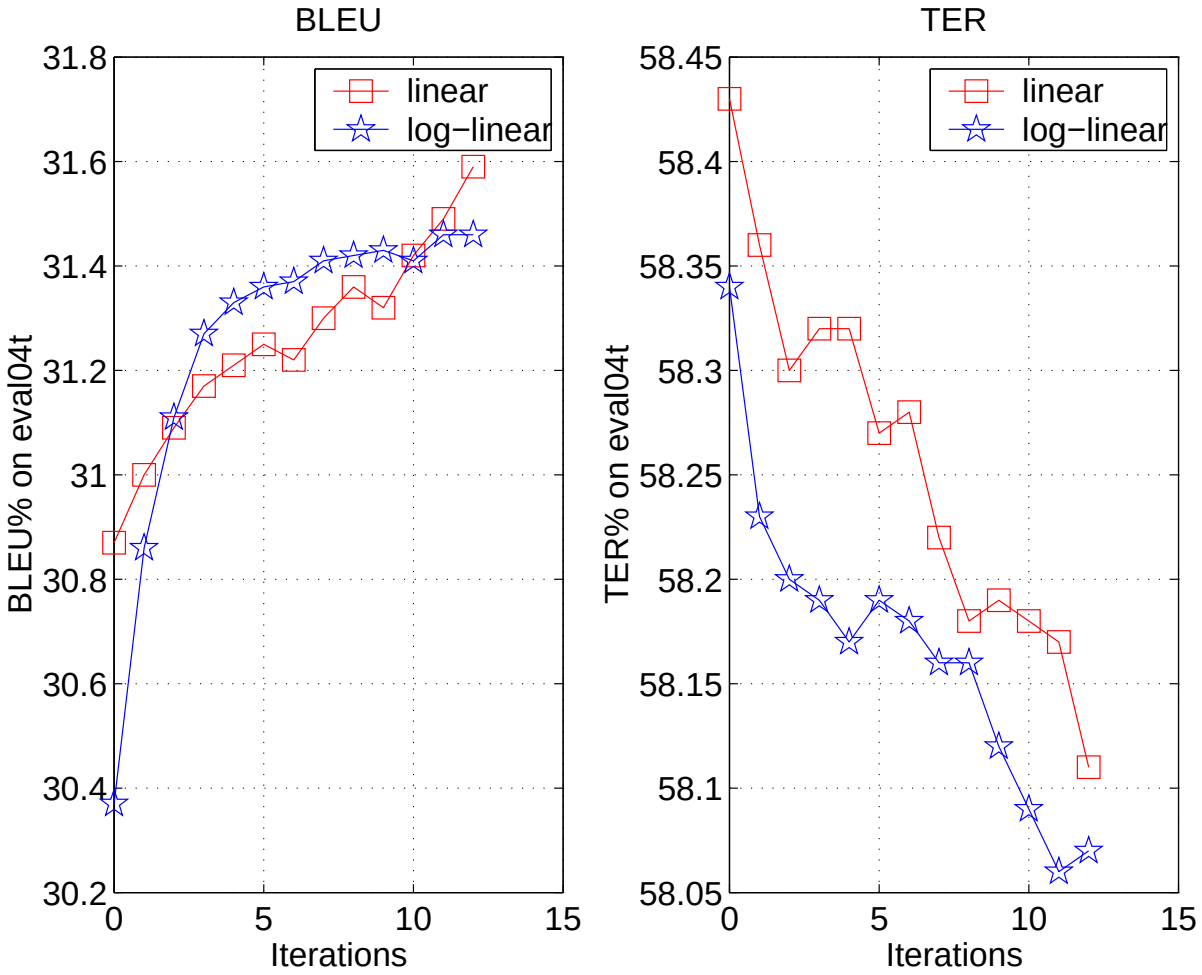
Set	System	Arabic		Chinese	
		TER	BLEU	TER	BLEU
text.nw	1-best	41.70	47.97	58.54	31.45
	mbr	41.69	47.94	58.47	31.61
	mbr.new	41.46	47.63	58.02	30.99
text.web	1-best	66.63	18.66	74.71	13.08
	mbr	66.71	18.57	74.77	13.13
	mbr.new	65.58	18.59	73.67	12.79
audio.stt	1-best	63.03	22.27	72.12	15.83
	mbr	63.05	22.27	72.27	15.80
	mbr.new	62.24	22.27	70.98	15.44

MBR performance using CU Eval07 systems N-best outputs on AGILE system combination tuning sets

Discriminative Language Model Interpolation

- Considering linear or log-linear interpolated mixture LM:
- Correlation between perplexity and error rate can be weak
- Discriminative weights estimation by minimizing Bayes risk (MBR)
- Generalize to arbitrary forms of cost functions: WER/TER/BLEU
- EBW algorithm based weights estimation stable and efficient
- 0.5%-1.0% BLEU gain over perplexity based interpolation

Discriminative Language Model Interpolation (b1)



Optimization of linear and log-linear weights on tuning set eval04t using CU 2006 C-E system's N-best outputs

Discriminative Language Model Interpolation (b2)

Int Crit	eval04d		eval03	
	TER	BLEU	TER	BLEU
eql	59.13	30.10	59.29	29.97
pp	59.27	30.37	59.44	29.88
mbr	59.03	30.64	59.22	30.41

Translation performance on held-out Chinese text sets eval04d and eval03 by rescoring CU 2006 C-E system's N-best outputs

Unsupervised Discriminative Language Model Adaptation

- MBR weights estimation used to adapt mixture LM
- Adaptation performed on text or audio document level
- Outperforming perplexity or N-best based adaptation
- Up to 0.5% TER reduction
- Applicable to integrated adaption of translation and language models

Unsupervised Discriminative Language Model Adaptation (b1)

Adapt	TER%		
	bnmd06	bcmd05	eval06
fixed	72.24	75.28	80.46
pp	72.20	75.26	80.35
nbest	72.21	75.25	80.37
mbr	71.66	74.94	79.84

Adaptation performance by rescoring dry-run CU C-E system N-best outputs on Chinese audio sets