Context Dependent Language Model Adaptation

X. Liu, M. J. F. Gales & P. C. Woodland

September 14, 2008

Cambridge University Engineering Department

ISCA Interspeech'08 September 2008

Language Model Adaptation

- Language models are often constructed as mixture of n-gram models.

$$P(w_i|h_i^{n-1}) = \sum_m \lambda_m P_m(w_i|h_i^{n-1})$$

- Multiple component models trained on diverse sources are combined.
- Directly adapting n-gram probabilities to target domain impractical.
- Standard approaches re-adjust global interpolation weights using:
  - maximum likelihood, or minimizing perplexity (PP),

$$\ln P(\mathcal{W}) = \sum_m \sum_{i=1}^{L_m} \ln P(w_i|h_i^{n-1})$$

  - discriminative criteria, e.g., minimum Bayes risk (MBR)

Context Dependent Language Model Adaptation

- "*Usefulness*" of sources vary between contexts in:
  - modeling resolution: n-gram coverage, cut-off settings used in training.
  - form of parameter estimation: choice of smoothing schemes.
  - topics and styles.

- Global interpolation unable to capture context specific variability.

- Using history dependent weights to incorporate more context information.

$$P(w_i|h_i^{n-1}) = \sum_m \phi_m(h_i^{n-1}) P_m(w_i|h_i^{n-1})$$

- Baum-Welch (BW) or extended BW (EBW) algorithms for weights estimation.

- Robust estimation schemes required given limited adaptation data.

MAP Weight Adaptation

- Applicable to both perplexity based and discriminative adaptation.

$$\hat{\phi}_m(h_i^{n-1}) = \frac{\mathcal{C}_m(h_i^{n-1}) + \tau \phi_m^{\mathsf{Pr}}(h_i^{n-1})}{\sum_m \mathcal{C}_m(h_i^{n-1}) + \tau}$$

- Choices of smoothing priors $\phi_m^{\mathsf{Pr}}(h_i^{n-1})$ may be used:
  - ***dynamic*** prior estimated on the supervision data: closer to target domain.
  - ***static*** prior estimated on the training data: more informative.

- Using ***normalized perplexity*** (nPP) to remove bias to corpus size:

$$\ln P_{\mathsf{norm}}(\mathcal{W}) = \sum_m \frac{L}{L_m} \sum_{i=1}^{L_m} \ln P(w_i | h_i^{n-1})$$

- Suitable for building task independent LMs of context dependent interpolation.

Class Context Dependent Weight Adaptation

- Class based approach effective to address data sparsity problem.
- Dimensionality of history space too high for clustering.
- Only consider word level clustering.
- Sharing weights between word histories mapped to identical class contexts.

$$P(w_i|h_i^{n-1}) = \sum_m \phi_m(\mathcal{G}_i^{n-1}) P_m(w_i|h_i^{n-1})$$

- Deriving a suitable weight class mapping $\mathcal{G}_i^{n-1}$ important:
  - Standard clustering algorithms for class LMs inappropriate for weights.
  - Bottom-up iterative merging scheme derived for context dependent weights.
  - Efficient log-likelihood computation using a lower bound approximation.

Class Context Dependent Weight Adaptation (Cont)

Hist Type	Clustering Algorithm	Num Class	PP Trn	nPP Trn	PP Test
class	exchange algorithm	50	166.6	76.1	217.9
		100	164.4	74.8	216.6
		200	160.9	73.5	214.8
		400	159.7	73.0	213.9
	weight merge	50	165.0	73.6	213.5
		100	163.5	73.4	213.1
		200	162.4	73.1	213.0
		400	161.2	72.8	212.7

nPP and PP performance of weight clustering algorithms on a 10 source, 1 giga word Mandarin Chinese broadcast transcription setup for 1-gram (single word) history dependent weights.

- Bottom-up weight merging consistently outperforms exchange algorithm.
- More appropriate for context dependent weights.

Weights Back-off for Unobserved Contexts

Contexts that are not seen in the training or adaptation data:

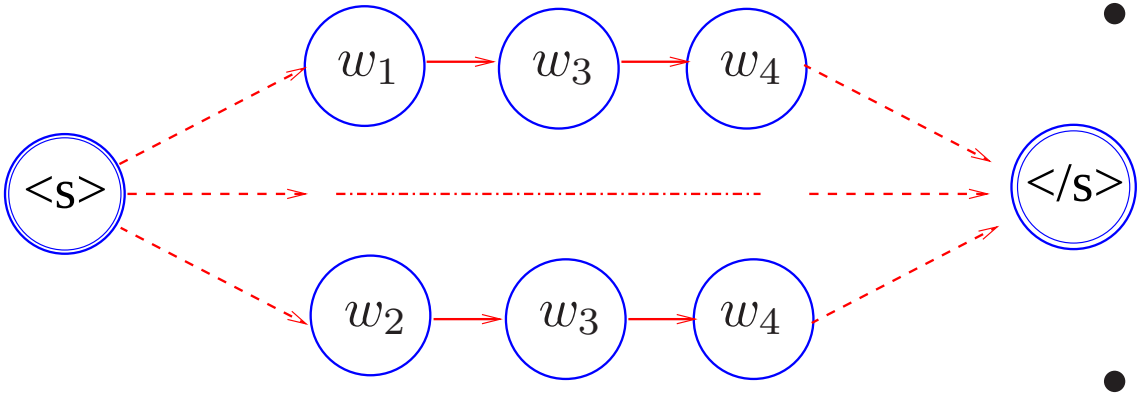
- handled by a simple back-off scheme:

$$\phi_{\text{bo}}(h_i^{n-1}) = \begin{cases} \phi(h_i^{n-1}) & \text{if } \exists\ \phi(h_i^{n-1}) \\ \phi(h_i^{n-2}) & \text{else if } \exists\ \phi(h_i^{n-2}) \\ \cdots & \cdots \\ \phi(\text{null}) & \text{otherwise} \end{cases}$$

- apply to both word and class based weights.
- no normalization required.
- stored in tree structure parallel to component n-gram models.
- efficient caching and search algorithms required for fast access.

Using Context Dependent LM Weights In Decoding

- Longer contexts of weight models give richer LM information.
- More paths with unique LM scores need to be kept distinct.
- Expansion jointly decided by component n-gram and weight back-off models.



$$P_1(w_4|w_3) \quad P_2(w_4|w_3) \quad \phi(w_1, w_3)$$

differs from

$$P_1(w_4|w_3) \quad P_2(w_4|w_3) \quad \phi(w_2, w_3)$$

- Significant lattice size increase of 20% to 120% over baseline language models using standard global interpolation.
- Flexible use of context dependent weights: off-line (static interpolation) or on the fly (dynamic interpolation) .

Mandarin Broadcast Speech Transcription Training Setup

- Performance evaluation used the CU-HTK Mandarin LVCSR system:
  - adapted HLDA MPE models trained on 942 hours of broadcast speech.
  - 58k word list, interpolated 4-gram word back-off language model.
  - 1.0G words from 10 text sources were used in LM training.
- Three evaluation sets were used:
  - 3.4 hour bndev06
  - 2.5 hour bcdev05
  - 1.8 hour eval06
- Baseline 4-gram 1-best hypothesis as supervision.
- Top 1000 hypotheses were extracted for MBR adaptation.

Mandarin Broadcast Speech Transcription Training Setup

Comp LM	Text (M)	Train Config	Global Weight Tuning		
			PPTest	PPTrn	nPPTrn
bcm	4.83	111,kn	0.2325	0.0049	0.1426
bnm	3.78	111,kn	0.1501	0.0066	0.1729
giga-xin	277.6	123,gt	0.1036	0.2428	0.1079
giga-cna	496.7	123,gt	0.0791	0.4577	0.0815
phoenix	76.89	112,kn	0.1542	0.1125	0.1225
voarfabbc	30.28	112,kn	0.0966	0.0270	0.0734
cctvcnr	26.81	112,kn	0.0592	0.0391	0.0844
tdt4	1.76	112,kn	0.0318	0.0060	0.0717
papersjing	83.73	122,kn	0.0444	0.0919	0.0928
ntdtv	12.49	122,kn	0.0485	0.0114	0.0503

Table 1: Text source, 2/3/4-gram cut-off settings, smoothing scheme used in training and global ML weights tuned using test set PP, training data PP and nPP scores for component LMs.

Performance of Unadapted LMs

Train Crit	Wgt Cntxt Type/Len	nPP Crit	Reference Perplexity			CER%		
			bn06	bc05	eval06	bn06	bc05	eval06
eql	-/-	84	201	252	247	8.3	19.3	19.1
PP.supv	-/-	84	199	224	229	8.2	19.0	19.0
PP	-/-	131	225	404	361	8.6	20.7	19.9
nPP	-/-	82	198	240	235	8.1	19.1	19.1
nPP	word/1g	70	193	224	222	8.1	19.1	19.1
	word/3g	52	***179***	***213***	***215***	***8.1***	***18.9***	***18.8***

Table 2: PP and lattice rescoring 1-best CER% of ML interpolated LMs.

- nPP ranking consistent with test data PP, bias to corpus size removed.
- Building task independent LMs, no test test information required!
- Context dependent interpolation outperformed global interpolation.

Performance of ML Adapted LMs

MAP Prior	Adapt (PP)	Reference Perplexity			CER%		
		bn06	bc05	eval06	bn06	bc05	eval06
-	-	199	224	229	8.2	19.0	19.0
-	glob	150	201	201	8.1	18.9	18.8
dynamic	word/1g	139	188	188	8.1	18.9	18.7
	word/3g	***133***	***176***	***184***	***8.0***	***18.8***	***18.7***
static	word/1g	141	187	187	8.1	18.8	19.0
	word/3g	132	182	186	8.0	18.8	18.7
dynamic	cls100/1g	146	196	194	8.1	18.9	18.7
	cls100/3g	128	180	172	8.1	19.0	18.9

Table 3: PP and CER% performance of ML adapted LMs.

- Context dependent adaptation outperformed global adaptation.
- Weak correlation between PP and CER, discriminative adaptation preferred.

Performance of MBR Adapted LMs

Crit	LM Adapt	CER% bn06 bc05 eval06
-	-	8.2 / 19.0 / 19.0
MBR	word/1g	8.1 / 18.7 / 18.6
	word/3g	***8.0*** / ***18.6*** / ***18.6***
MBR	cls100/1g	8.0 / 18.9 / 18.7
	cls100/3g	8.1 / 18.8 / 18.7

Table 4: CER% performance of MBR adapted LMs.

- Criterion closer approximation to error rate.
- Discriminative adaptation outperformed ML adaptation.
- Statistically significant CER reduction up to 0.4%.

Conclusion and Future Work

- Context dependent LM adaptation outperforms global adaptation.
- nPP criterion may be used to train task independent models.
- Future research focuses on:
  - hierarchical discriminative interpolation in LM training and adaptation;
  - improving weight parameter smoothing;
  - integrated discriminative weight clustering and estimation;
  - non-deterministic, soft-tying of word to class contexts mapping.

Thank you!