Discriminative Models for Speech Recognition

Mark Gales

1 February 2007

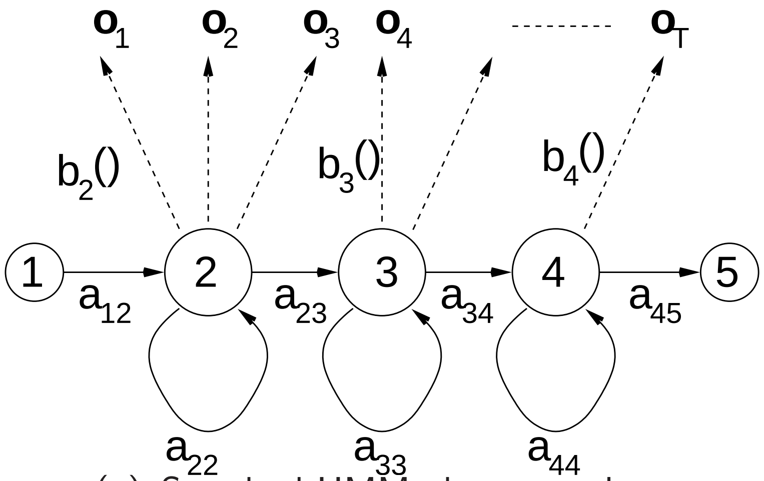
Cambridge University Engineering Department

ITA Workshop

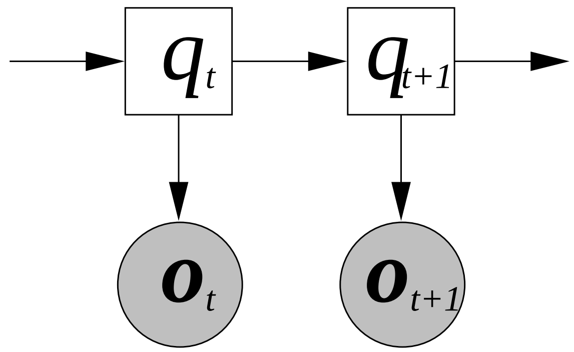
Overview

- Generative model for Speech Recognition - Hidden Markov Models
  - discriminative criteria - MMI, MCE, MPE
- Discriminative classifiers
  - maximum entropy Markov models
  - hidden conditional random fields
- Dynamic kernels - Fisher kernels, generative kernels
- Conditional augmented models

Hidden Markov Model



(a) Standard HMM phone topology



(b) HMM Dynamic Bayesian Network

- HMM generative model
  - class posteriors, $P(\mathbf{w}|\mathbf{O}_{1:T};\boldsymbol{\lambda})$, obtained using Bayes' rule
  - requires class priors, $P(\mathbf{w})$ - language models in ASR
- Maximum likelihood training criterion used in many applications
  - ASR - Gaussian Mixture Models (GMMs) as state output distributions
  - efficiently implemented using Expectation-Maximisation (EM)
- Poor model of the speech process - piecewise constant state-space.

Discriminative Training Criteria

- Discriminative training criteria commonly used to train HMMs for ASR
  - Maximum Mutual Information (MMI) [1, 2]: maximise

$$\mathcal{F}_{\texttt{mmi}}(\boldsymbol{\lambda}) = \frac{1}{R}\sum_{r=1}^{R}\log(P(\mathbf{w}_{\texttt{ref}}^{(r)}|\mathbf{O}^{(r)};\boldsymbol{\lambda}))$$

  - Minimum Classification Error (MCE) [3]: minimise

$$\mathcal{F}_{\texttt{mce}}(\boldsymbol{\lambda}) = \frac{1}{R}\sum_{r=1}^{R}\left(1+\left[\frac{p(\mathbf{O}^{(r)}|\mathbf{w}_{\texttt{ref}}^{(r)};\boldsymbol{\lambda})P(\mathbf{w}_{\texttt{ref}}^{(r)})}{\sum_{\mathbf{w}\neq\mathbf{w}_{\texttt{ref}}^{(r)}}p(\mathbf{O}^{(r)}|\mathbf{w};\boldsymbol{\lambda})P(\mathbf{w})}\right]^{\varrho}\right)^{-1}$$

  - Minimum Bayes' Risk (MBR) [4, 5]: minimise

$$\mathcal{F}_{\texttt{mbr}}(\boldsymbol{\lambda}) = \frac{1}{R}\sum_{r=1}^{R}\sum_{\mathbf{w}}P(\mathbf{w}|\mathbf{O}^{(r)};\boldsymbol{\lambda})\mathcal{L}(\mathbf{w},\mathbf{w}_{\texttt{ref}}^{(r)})$$

MBR Loss Functions for ASR

- Sentence (1/0 loss):

$$\mathcal{L}(\mathbf{w}, \mathbf{w}_{\mathtt{ref}}^{(r)}) = \begin{cases} 1; & \mathbf{w} \neq \mathbf{w}_{\mathtt{ref}}^{(r)} \\ 0; & \mathbf{w} = \mathbf{w}_{\mathtt{ref}}^{(r)} \end{cases}$$

When $\varrho = 1$, $\mathcal{F}_{\mathtt{mce}}(\boldsymbol{\lambda}) = \mathcal{F}_{\mathtt{mbr}}(\boldsymbol{\lambda})$

- Word: directly related to minimising the expected Word Error Rate (WER)
  - normally computed by minimising the Levenshtein edit distance.
- Phone: consider phone rather word loss
  - improved generalisation as more “error’s” observed
  - this is known as Minimum Phone Error (MPE) training [6, 7].

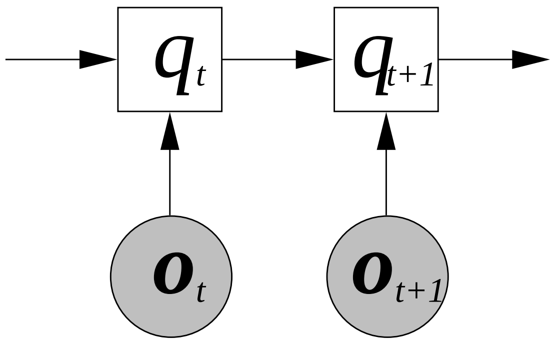
Discriminative Training for LVCSR Systems

- Modifications to direct implementation using, e.g. extended Baum Welch
  - Efficient denominator representation: lattices often used
  - Acoustic Deweighting: scale state/segment probabilities
  - Language Model "Weakening": use heavily pruned bigram/unnigram rather than tri-gram/4-gram
  - I-Smoothing: use ML estimates as priors for discriminative estimation
- Last three are important to achieve good generalisation
- Example Broadcast News LVCSR gains ($\approx 500 - 1000$ hours training data)
  - typically 200K-300K Gaussian components for each system

Language	Training	
	ML	MPE
English (WER%)	16.0	13.1
Arabic (WER%)	22.9	20.0
Mandarin (CER%)	14.4	12.7

Maximum Entropy Markov Models

- Attempt to model the class posteriors directly - MEMMs one example
  - The DBN and associated word sequence posterior [8]



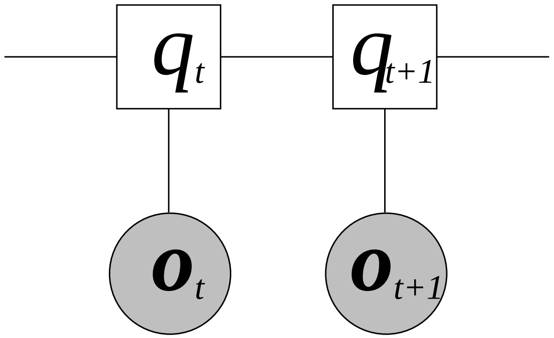
$$P(\mathbf{w}|\mathbf{O}_{1:T};\boldsymbol{\alpha}) = \sum_{\mathbf{q}} P(\mathbf{w}|\mathbf{q}) \prod_{t=1}^{T} P(q_t|\mathbf{o}_t, q_{t-1};\boldsymbol{\alpha})$$

$$P(q_t|\mathbf{o}_t, q_{t-1};\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha},\mathbf{o}_t)} \exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{T}(\mathbf{o}_t, q_t, q_{t-1})\right)$$

- Features extracted - transitions $\mathbf{T}(q_t, q_{t-1})$, observations $\mathbf{T}(\mathbf{o}_t, q_t)$
  - same features as standard HMMs
- Problems incorporating language model prior
  - gains over standard (ML-trained) HMM with no LM
  - does yield gains in combination with standard HMM

Hidden Conditional Random Fields

- Conditional random fields hard to directly apply to speech data
  - observation sequence length T doesn't word match label sequence L
  - introduce latent discrete sequence (similar to HMM)
- The feature dependencies in the HCRF and word sequence posterior [9]



$$
\begin{aligned}
&P(\mathbf{w}|\mathbf{O}_{1:T};\boldsymbol{\alpha}) \\
&\quad = \frac{1}{Z(\boldsymbol{\alpha},\mathbf{O}_{1:T})}\sum_{\mathbf{q}}\exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{T}(\mathbf{O}_{1:T},\mathbf{w},\mathbf{q})\right)
\end{aligned}
$$

$$
\mathbf{T}(\mathbf{O}_{1:T},\mathbf{w},\mathbf{q}) = \left[\begin{array}{c}\mathbf{T}_1(\mathbf{w}) \\ \mathbf{T}_{\mathsf{a}}(\mathbf{O}_{1:T},\mathbf{w},\mathbf{q})\end{array}\right]
$$

  - $\mathbf{T}_1(\mathbf{w})$ may be replaced by $\log(P(\mathbf{w}))$
  - allows LM text training data to be used

HCRF Features

- The features used with HCRFs

$$\mathbf{T}_{\mathsf{a}}(\mathbf{O}_{1:T}, \mathbf{w}, \mathbf{q}) = \left[\begin{array}{c} \vdots \\ \sum_{t=1}^{T} \delta(q_{t-1} - s_i)\delta(q_t - s_i) \\ \sum_{t=1}^{T} \delta(q_t - s_i) \\ \sum_{t=1}^{T} \delta(q_t - s_i)\mathbf{o}_t \\ \sum_{t=1}^{T} \delta(q_t - s_i)\mathsf{vec}(\mathbf{o}_t\mathbf{o}_t^{\mathsf{T}}) \\ \vdots \end{array} \right]$$

  - features the same as those associated with a generative HMM
  - state "distributions" not required to be valid individual PDFs

- Non-convex optimisation problem

Interest in modifying features extracted from sequence

Dynamic Kernels

- Dynamic kernels (or features) map sequence data into a fix dimensionality
  - standard classifiers (e.g. SVMs) can then be applied
  - examples include marginalised count kernels [10], Fisher kernels [11]
- Generative kernels [12] modified version of Fisher kernels

$$\boldsymbol{\phi}(\mathbf{O}_{1:T};\boldsymbol{\lambda}) = \left[\begin{array}{c} \log(p(\mathbf{O}_{1:T};\boldsymbol{\lambda})) \\ \boldsymbol{\nabla}_{\lambda}\log(p(\mathbf{O}_{1:T};\boldsymbol{\lambda})) \\ \vdots \\ \boldsymbol{\nabla}^{\rho}_{\lambda}\log(p(\mathbf{O}_{1:T};\boldsymbol{\lambda})) \end{array}\right]$$

  - ρ is the order of the kernel
  - $\boldsymbol{\lambda}$ specifies the parameters of the generative model.
- Can be used in generative models - augmented statistical models [13]

HMM Generative Features

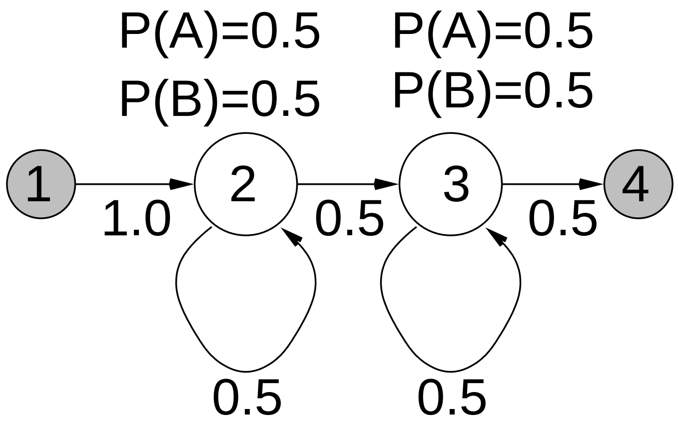
- HMM: $p(\mathbf{O}_{1:T};\boldsymbol{\lambda}) = \sum_{\mathbf{q}\in\boldsymbol{\Theta}} \left\{\prod_{t=1}^{T} a_{q_{t-1}q_t} \left(\sum_{m\in q_t} c_m \mathcal{N}(\mathbf{o}_t;\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)\right)\right\}$
- Derivative depends on posterior, $\gamma_{jm}(t) = P(q_t = \{s_j, m\}|\mathbf{O}_{1:T};\boldsymbol{\lambda})$,

$$\boldsymbol{\nabla}_{\mu_{jm}} \log\left(p(\mathbf{O}_{1:T};\boldsymbol{\lambda})\right) = \sum_{t=1}^{T} \gamma_{jm}(t)\boldsymbol{\Sigma}_{jm}^{-1}\left(\mathbf{o}_t - \boldsymbol{\mu}_{jm}\right)$$

  - posterior depends on complete observation sequence, $\mathbf{O}$
  - introduces dependencies beyond conditional state independence
  - compact representation of effects of all observations
- Higher-order derivatives incorporate higher-order dependencies
  - increasing order of derivatives - increasingly powerful trajectory model
  - systematic approach to incorporating additional dependencies

Example Generative Kernel Features

- Consider a simple 2-class, 2-symbol $\{\mathtt{A}, \mathtt{B}\}$ problem:
  - Class ω_1: `AAAA`, `BBBB`
  - Class ω_2: `AABB`, `BBAA`



Feature	Class ω_1		Class ω_2	
	`AAAA`	`BBBB`	`AABB`	`BBAA`
Log-Lik	-1.11	-1.11	-1.11	-1.11
∇_{2A}	0.50	-0.50	0.33	-0.33
$\nabla_{2A}\nabla'_{2A}$	-3.83	0.17	-3.28	-0.61
$\nabla_{2A}\nabla'_{3A}$	-0.17	-0.17	-0.06	-0.06

- ML-trained HMMs are the same for both classes
- First derivative classes separable, but not linearly separable
  - also true of second derivative within a state
- Second derivative across state linearly separable

Conditional Augmented Models

- Features from dynamic kernels can be included in a discriminative fashion
  - maximise

$$P(\mathbf{w}|\mathbf{O};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\lambda},\boldsymbol{\alpha})}\exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\begin{bmatrix}\mathbf{T}_1(\mathbf{w})\\ \mathbf{T}_{\mathsf{a}}(\mathbf{O}_{1:T},\mathbf{w})\end{bmatrix}\right)$$

$$\mathbf{T}_{\mathsf{a}}(\mathbf{O}_{1:T},\mathbf{w}) = \begin{bmatrix}\vdots\\ \delta(\mathbf{w}-\tilde{\mathbf{w}})\log(p(\mathbf{O}_{1:T};\boldsymbol{\lambda}^{(\tilde{\mathbf{w}})}))\\ \vdots\\ \delta(\mathbf{w}-\tilde{\mathbf{w}})\boldsymbol{\nabla}_{\lambda}\log(p(\mathbf{O}_{1:T};\boldsymbol{\lambda}^{(\tilde{\mathbf{w}})}))\\ \vdots\end{bmatrix}$$

- Standard gradient descent approaches may be used to train parameters
  - optimising $\boldsymbol{\alpha}$ is a convex optimisation problem - unique, global solution
  - optimising $\boldsymbol{\lambda}$ is non-convex ...

TIMIT Classification Experiments

- TIMIT phone-classification experiments
  - 48 base-phones modelled (mapped to 39 for scoring)
  - context-independent phone base models. 3-emitting state HMMs

Classifier	Training		Components	
	λ	α	10	20
HMM	ML	–	29.4	27.3
C-Aug	ML	CML	24.2	–
HMM	MMI	–	25.3	24.8
C-Aug	MMI	CML	23.4	–

Classification error on the TIMIT core test set

- C-Aug outperforms HMMs for comparable numbers of parameters
  - currently not as good as the best HCRF numbers

Summary

- Discriminative training criteria used in state-of-the-art ASR system
  - underlying acoustic model still a generative HMM
- Recent interest in discriminative acoustic models for ASR, e.g.
  - maximum entropy Markov models,
  - hidden conditional random fields
  - dynamic kernels/condition augmented models
- Consistent gains over discriminatively trained HMMs
  - majority of evaluation on small tasks (TIMIT phone classification/recognition)
- Hard to predict whether gains will map to LVCSR tasks
  - various techniques necessary for good discriminative training generalisation

References

[1] P.S. Gopalakrishnan, D. Kanevsky, A. Nádas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. Information Theory*, 1991.

[2] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech & Language*, vol. 16, pp. 25–47, 2002.

[3] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, 1992.

[4] J Kaiser, B Horvat, and Z Kacic, "A novel loss function for the overall risk criterion based discriminative training of HMM models," in *Proc. ICSLP*, 2000.

[5] W Byrne, "Minimum Bayes risk estimation and decoding in large vocabulary continuous speech recognition," *IEICE Special Issue on Statistical Modelling for Speech Recognition*, 2006.

[6] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, May 2002.

[7] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 2004.

[8] H-K Kuo and Y Gao, "Maximum entropy direct models for speech recognition," *IEEE Transactions Audio Speech and Language Processing*, 2006.

[9] A. Gunawardana, M. Mahajan, A. Acero, and J.C. Platt, "Hidden conditional random fields for phone classification," in *Interspeech*, 2005.

[10] K. Tsuda, T. Kin, and K. Asai, "Marginalized kernels for biological sequences," *Bioinformatics*, vol. 18, pp. S268–S275, 2002.

[11] T. Jaakkola and D. Hausser, "Exploiting generative models in disciminative classifiers," in *Advances in Neural Information Processing Systems 11*, S.A. Solla and D.A. Cohn, Eds. 1999, pp. 487–493, MIT Press.

[12] M. Layton and M.J.F. Gales, "Maximum margin training of generative kernels," Tech. Rep. CUED/F-INFENG/TR.484, Department of Engineering, University of Cambridge, June 2004.

[13] M. J. F. Gales and M. I. Layton, "Training augmented models using svms," *IEICE Special Issue on Statistical Modelling for Speech Recognition*, 2006.