Adaptation Techniques for Language Models

2003/12/16

Louis Tsai
Speech Lab, CSIE, NTNU
louis@csie.ntnu.edu.tw

Reference Paper

- Unsupervised Language Model Adaptation
  - AT&T Labs, ICASSP 2003
- Discriminative Training of Language Models for Speech Recognition
  - Bell Labs, C.H. Lee, ICASSP 2002
- Efficient Language Model Adaptation Through MDI Estimation
  - Marcello Federico, EUROSPEECH 99

Introduction

to find the optimal word sequence $\tilde{W}$ for a given speech signal X :

$$\tilde{W} = \arg\max_{W} P(W \mid X) = \arg\max_{W} \frac{P(X \mid W)P(W)}{P(X)}$$

$$\cong \arg\max_{W} P(X \mid W)P(W)$$

where $P(X \mid W)$ is the acoustic model

$P(W)$ is the language model

MAP

- Language Model (LM) and acoustic model (AM) adaptation attempt to obtain models for a new domain with little training data
- AM adaptation bas been studied extensively
- LM adaptation has received much less attention
- The most widespread approaches to supervised LM adaptation in a large vocabulary setting are model interpolation and count mixing

MAP

- Both count mixing and model interpolation can both be viewed as a *maximum a posteriori* (MAP) adaptation strategy with a different parameterization of the prior distribution
- The model parameters θ are assumed to be a random vector in the space Θ, and x is a given observation sample
- The MAP estimate is the posterior distribution of θ

$$\theta_{\text{MAP}} = \arg\max_{\theta} g(\theta \mid \text{x}) = \arg\max_{\theta} f(\text{x} \mid \theta) g(\theta)$$

MAP

- The prior distribution of the weights $\omega_1, \omega_2, \ldots, \omega_K$ is Dirichlet density

$$g(\omega_1, \omega_2, ..., \omega_K \mid v_1, v_2, ..., v_K) \propto \prod_{i=1}^{K} \omega_i^{v_i - 1}$$

where $v_i > 0$ are the parameters of the Dirichlet distribution

MAP

- c_i : expected counts for the i-th component

$$f(\mathrm{x}\,|\,\theta) = f(x_1,\ldots,x_T\,|\,\omega_1,\ldots,\omega_K) \propto \prod_{i=1}^{K} \omega_i^{c_i}$$

$$\therefore f(\mathrm{x}\,|\,\theta)g(\theta)$$

$$= f(x_1,..,x_T\,|\,\omega_1,\ldots,\omega_K)\cdot g(\omega_1,\ldots,\omega_K\,|\,v_1,\ldots,v_K)$$

$$= \prod_{i=1}^{K} \omega_i^{v_i-1+c_i}$$

Apply Largrange Multiplier

$$\sum_{i=1}^{K} \log \omega_i^{v_i-1+c_i} = \sum_{i=1}^{K} (v_i - 1 + c_i)\log \omega_i + l(\sum_{i=1}^{K} \omega_i - 1)$$

MAP

Differentiate w.r.t ω_i

$$(v_i - 1 + c_i)\frac{1}{\omega_i} + l = 0 \quad \Rightarrow \quad \omega_i = -\frac{v_i - 1 + c_i}{l} \quad \ldots\ldots (1)$$

$$\sum_{i=1}^{K} \omega_i = -\sum_{i=1}^{K} \frac{v_i - 1 + c_i}{l} = 1 \quad \therefore \; l = -\sum_{i=1}^{K} (v_i - 1 + c_i) \quad \text{代入(1)}$$

得 $$\omega_i = \frac{v_i - 1 + c_i}{\sum_{k=1}^{K} (v_k - 1 + c_k)}$$

MAP
count mixing

- Mixing parameters α and β

$$v_i = \tilde{c}(h)\frac{\alpha}{\beta}\tilde{P}(w_i \mid h) + 1$$

$$\hat{P}(w_i \mid h) = \frac{\tilde{c}(h)\frac{\alpha}{\beta}\tilde{P}(w_i \mid h) + \bar{c}_d(hw_i)}{\sum_{k=1}^{K}\left[\tilde{c}(h)\frac{\alpha}{\beta}\tilde{P}(w_k \mid h)\right] + \bar{c}(h)}$$

$$= \frac{\alpha\tilde{c}_d(hw_i) + \beta\bar{c}_d(hw_i)}{\alpha\tilde{c}(h) + \beta\bar{c}(h)}$$

MAP
model interpolation

$$v_i = \bar{c}(h)\frac{\lambda}{1-\lambda}\tilde{P}(w_i \mid h) + 1$$

$$\hat{P}(w_i \mid h) = \frac{\bar{c}(h)\frac{\lambda}{1-\lambda}\tilde{P}(w_i \mid h) + \bar{c}_d(hw_i)}{\sum_{k=1}^{K}\left[\bar{c}(h)\frac{\lambda}{1-\lambda}\tilde{P}(w_k \mid h)\right] + \bar{c}(h)}$$

$$= \frac{\frac{\lambda}{1-\lambda}\tilde{P}(w_i \mid h) + \bar{P}(w_i \mid h)}{\frac{\lambda}{1-\lambda} + 1}$$

$$= \lambda\tilde{P}(w_i \mid h) + (1-\lambda)\bar{P}(w_i \mid h)$$

MCE

Given an observation sequence X_i representing the speech signal and a word sequence $W = w_1, w_2, \ldots, w_n$,

$$\tilde{W} = \arg\max_W P(W \mid X_i) = \arg\max_W \frac{P(X_i \mid W)P(W)}{P(X_i)}$$

$$\cong \arg\max_W P(X_i \mid W)P(W)$$

define a discriminate function that is a weighted combination of acoustic and language model scores :

$$g(X_i, W; \Lambda, \Gamma) = \alpha \log P(X_i \mid W, \Lambda) + \log P(W \mid \Gamma)$$

Λ is the acoustic model, Γ is the language model,

α is the inverse of the language model weight

MCE

$$W_1 = \arg\max_{W} g(X_i, W; \Lambda, \Gamma)$$

$$W_\gamma = \arg\max_{W \neq W_1,\ldots,W_{\gamma-1}} g(X_i, W; \Lambda, \Gamma)$$

W_1 has the large value for $g()$

W_r is the rth best hypothesized word sequence

W_0 is the known correct word sequence

- Compare the discriminate function for W_0 and that for N competing word sequences $\{W_1, W_2, \ldots, W_N\}$ hypothesized by the recognizer
  - misclassification function

MCE

misclassification function :

$$d(X_i;\Lambda,\Gamma) = -g(X_i,W_0;\Lambda,\Gamma) + G(X_i,W_1,...,W_N;\Lambda,\Gamma)$$

anti-discriminant function :

$$G(X_i,W_1,...,W_N;\Lambda,\Gamma) = \log\left(\frac{1}{N}\sum_{\gamma=1}^{N}\exp[g(X_i,W_\gamma;\Lambda,\Gamma)\eta]\right)^{\frac{1}{\eta}}$$

if $\eta \to \infty$, the anti - discriminant function is dominated by the biggest competing discriminant function :

$$G(X_i,W_1,...W_N;\Lambda,\Gamma) \to g(X_iW_1;\Lambda,\Gamma)$$

MCE

class loss function :

$$l(X_i) = l(d(X_i)) = \frac{1}{1+\exp(-\gamma d(X_i)+\theta)}$$

γ and θ are constants which control the slope and the shift of the sigmoid function, respectively

Using the GPD algorithm, the parameters of the language model can be adjusted iteratively (with step size ε) using the following update equation to minimize the recognition error:

$$\Gamma_{t+1} = \Gamma_t - \varepsilon \nabla l(X_i; \Lambda_t, \Gamma_t)$$

MCE

- We keeping the acoustic model constant the gradient of the loss function becomes

$$\nabla l = \frac{\partial l_i}{\partial d_i} \frac{\partial d(X_i;\Lambda,\Gamma)}{\partial \Gamma}$$

$$\frac{\partial l_i}{\partial d_i} = \gamma l(d_i)(1 - l(d_i))$$

- Using bigram :

$$\frac{\partial d(X_i;\Lambda,\Gamma)}{\partial p_{w_x w_y}} = \left[-I(W_0, w_x w_y) + \sum_{\gamma=1}^{N} C_\gamma I(W_\gamma, w_x w_y) \right]$$

$I(W, w_x w_y)$ denotes the number of times the bigram $w_x w_y$ appears in word sequence W

$$C_\gamma = \frac{\exp[g(X_i, W_\gamma;\Lambda,\Gamma)\eta]}{\sum_{j=1}^{N} \exp[g(X_i, W_j;\Lambda,\Gamma)\eta]}$$

$$\frac{\partial l_i}{\partial d_i} = \frac{\partial \dfrac{1}{1+\exp(-\gamma d(X_i)+\theta)}}{\partial d_i}$$

$$= \left(\frac{1}{1+\exp(-\gamma d(X_i)+\theta)}\right)^2 \cdot \left(-\exp(-\gamma d(X_i)+\theta)\right)\cdot(-\gamma)$$

$$= \gamma \cdot l(d_i) \cdot \frac{\exp(-rd(X_i)+\theta)+1-1}{1+\exp(-rd(X_i)+\theta)}$$

$$= \gamma \cdot l(d_i) \cdot (1-l(d_i))$$

$$\frac{\partial d(X_i;\Lambda,\Gamma)}{\partial P_{w_x w_y}} = \frac{\partial\left(-g(X_i,W_0;\Lambda,\Gamma)+G(X_i,W_1...,W_N;\Lambda,\Gamma)\right)}{\partial P_{w_x w_y}}$$

$$= \frac{\partial\left(-\log P(W_0\mid\Gamma)+\log\left(\frac{1}{N}\sum_{r=1}^{N}\exp[g(X_i,W_r;\Lambda,\Gamma)\eta]\right)^{\frac{1}{\eta}}\right)}{\partial P_{w_x w_y}}$$

$$= \frac{\partial\left(-(\sum P_{w_i w_j})+\frac{1}{\eta}\log(\frac{1}{N}\sum_{r=1}^{N}\exp[g(X_i,W_r;\Lambda,\Gamma)\eta])\right)}{\partial P_{w_x w_y}}$$

$$= -I(W_0,w_x w_y)+\frac{1}{\eta}\cdot\frac{1}{\frac{1}{N}\sum_{j=1}^{N}\exp[g(X_i,W_j;\Lambda,\Gamma)\eta]}\cdot\frac{1}{N}\left[\sum_{r=1}^{N}e^{g(X_i,W_r;\Lambda,\Gamma)\eta}\times\eta\times(1\mid 0)\right]$$

$$= -I(W_0,w_x w_y)+\sum_{r=1}^{N}\frac{e^{g(X_i,W_r;\Lambda,\Gamma)\eta}}{\sum_{j=1}^{N}\exp[g(X_i,W_j;\Lambda,\Gamma)\eta]}\cdot I(W_r,w_x w_y)$$

MDI

- Minimum discrimination information (MDI)
- A new LM is estimated so that it is “as close as possible” to a general background LM

Background LM

An n-gram LM approximates the probability $\Pr(W_1^T)$ of a text of words $W_1^T = w_1 \ldots, w_t, \ldots, w_T$, from a finite vocabulary V, with the product :

$$\Pr(W_1^T) = \prod_{t=1}^{T} \Pr(w_t \mid h_t)$$

where $h_t = w_{t-n+1} \ldots w_{t-1}$

smooth

- Data sparseness of real texts suggest to *smooth* *n*-gram probabilities

$$P_B(w \mid h) = f_B^*(w \mid h) + \lambda_B(h) P_B(w \mid h')$$

where $f_B^*(w \mid h)$ is the *discounted* frequency, $\lambda_B(h)$ is the zero - frequency probability :

$$\lambda_B(h) = 1.0 - \sum_{w \in V} f_B^*(w \mid h)$$

$$e.g., \; f_B^*(w \mid h) = \max\left\{ \frac{c_B(hw) - \beta}{c_B(h)}, 0 \right\} \;, \; \beta = \frac{n_1}{n_1 + 2n_2}$$

n_i represents the number of *n*-grams that occurred exactly *I* times in *B*

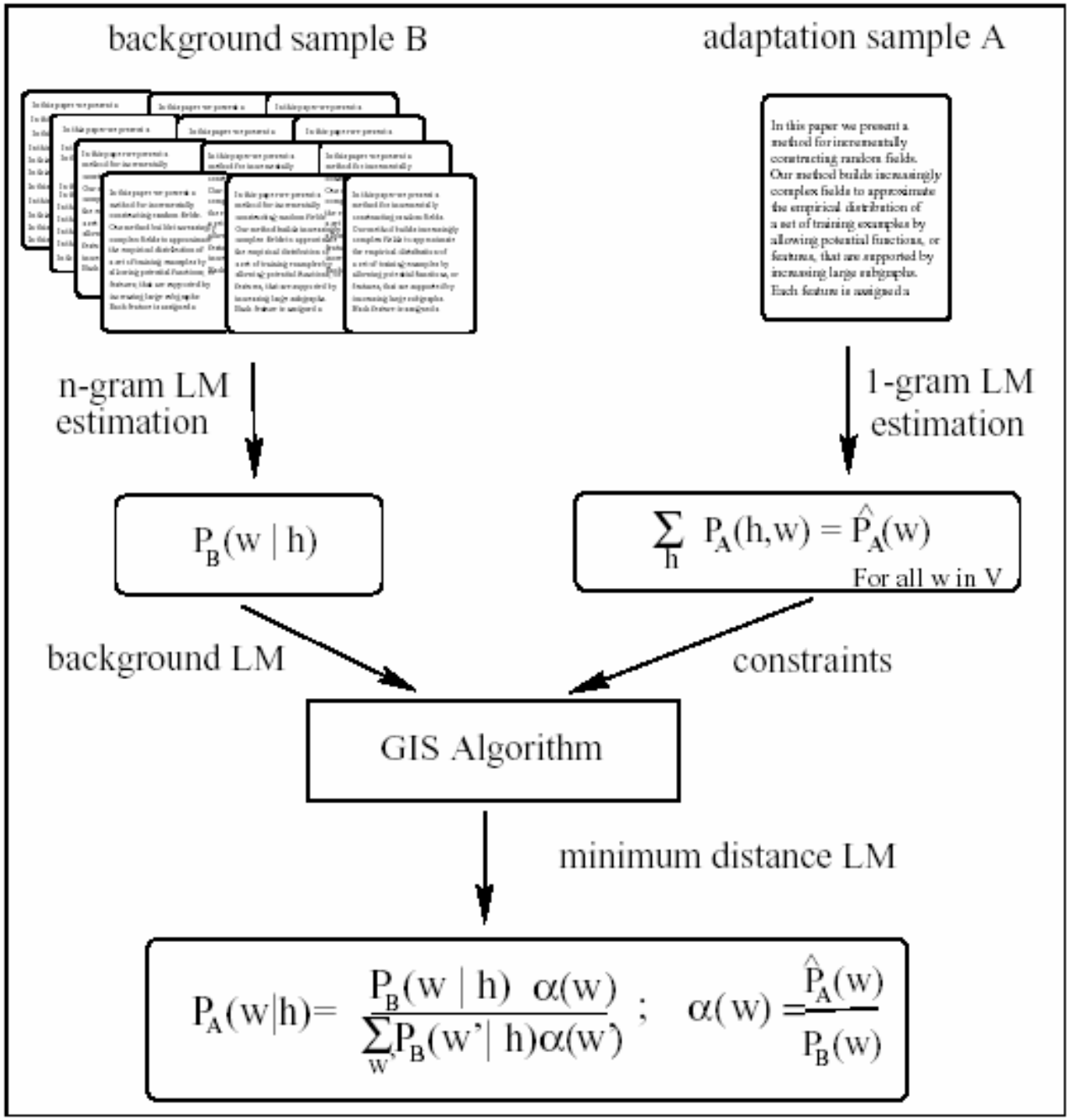

Figure 1: LM adaptation through MDI.

MDI LM Adaptation

- Formally, a set of linear constraints on the joint distribution $P_A(h,w)$ is specified, i.e.:

$$\sum_{hw \in V^n} P_A(h,w)\delta_i(hw) = \hat{P}_A(S_i) \qquad i = 1,\ldots,M$$

$\delta_i(\cdot)$ are indicator functions of subsets $S_i \subset V^n$, also called features, and $\hat{P}_A(S_i)$ are empirical estimates of the features on A

Kullback-Leibler distance

- Minimize the KL distance between background

$$P_A(\cdot) = \arg\min_{Q(\cdot)} \sum_{hw \in V^n} Q(h,w) \log \frac{Q(h,w)}{P_B(h,w)}$$

Generalized Iterative Scaling

- Assuming each $hw \in V^n$ exactly k features

$$P_A^{(0)}(h,w) = P_B(h,w)$$

$$P_A^{(r+1)}(h,w) = P_A^{(r)}(h,w)\prod_{i=1}^{M}\left(\frac{\hat{P}_A(S_i)}{P_A^{(r)}(S_i)}\right)^{\frac{\delta_i(hw)}{k}}$$

where :

$$P_A^{(r)}(S_i) = \sum_{hw \in V^n} P_A^{(r)}(h,w)\delta_i(hw) \qquad i = 1,\ldots,M$$

Generalized Iterative Scaling

- Given that the adaptation sample is typically small, one may assume that only unigram features can be reliably estimated on A. Hence, the following constraints can be set:

$$\sum_{hw \in V^n} P_A(h,w)\delta_{\hat{w}}(hw) = \hat{P}_A(\hat{w}) \qquad \forall \hat{w} \in V$$

$$\text{where } \delta_{\hat{w}}(hw) = \begin{cases} 1, & w = \hat{w} \\ 0, & \text{otherwise} \end{cases}$$

Generalized Iterative Scaling

- k =1. The GIS algorithm reduces to the following closed form:

$$P_A(h,w) = P_B(h,w)\alpha(w)$$

$$\alpha(w) = \frac{\hat{P}_A(w)}{P_B(w)}$$

Generalized Iterative Scaling

$$P_A(w \mid h) = \frac{P_B(w \mid h) P_B(h) \alpha(w)}{\sum_{\hat{w} \in V} P_B(\hat{w} \mid h) P_B(h) \alpha(\hat{w})}$$

$$= \frac{P_B(w \mid h) \alpha(w)}{\sum_{\hat{w} \in V} P_B(\hat{w} \mid h) \alpha(\hat{w})}$$