Statistical Language Model Adaptation

Reference:
J. R. Bellegarda, “Statistical language model adaptation: review and perspectives”, 2003

Outline

- Introduction
- Adaptation techniques
- Conclusions

Why adaptation?

- Natural language is highly variable in several aspects
  - Language evolves as does the world it seeks to describe
    - The vocabulary changes dynamically with time
    - “protemics” to the utter demise of “ague”
  - Different domains tend to involve relatively disjoint concepts with markedly different word sequence statistics
    - Subject matter affects the underlying semantic characteristics
    - Interest rate to banking application vs. a general conversation on gaming platforms

Why adaptation? (cont.)

- People naturally adjust their use of the language based on the task at hand
  - Typical syntax or grammatical infrastructure
  - Formal technical papers vs. casual e-mails

- People style of discourse may independently vary due to a variety of factors such as socio-economics, emotional state
  - More pronounced on spoken natural language

Why adaptation? (cont.)

- Lexical, syntactic or semantic characteristics of the discourse in the training and recognition tasks are quite likely differ
- The performance of any statistical approach, i.e. n-gram modeling, always suffers from such mismatch
- SLMs is extremely brittle across domains and even within domain when training and recognition domain involve moderately disjoint time periods
- Adaptation techniques
  - Model adaptation
  - Constraint specification
  - Meta-information extraction

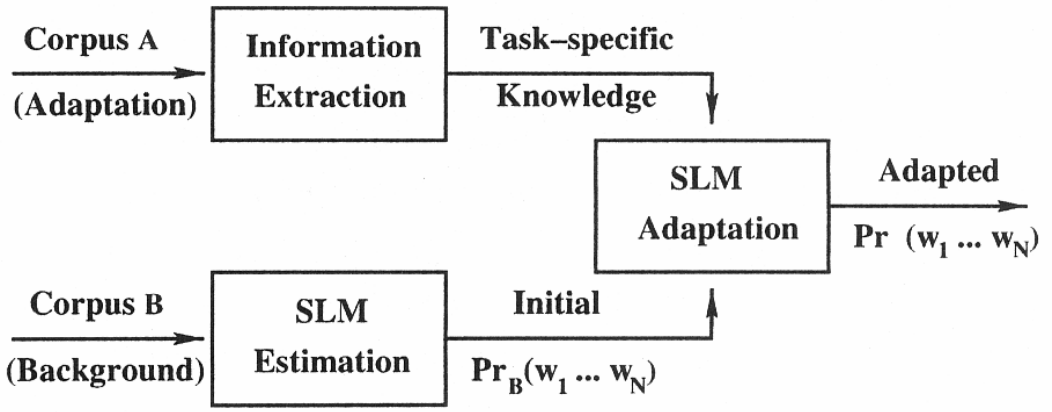
Adaptation framework

- LM estimate:

$$\Pr(w_1,\ldots,w_N)=\prod_{q=1}^{N}\Pr(w_q \mid h_q)$$

  - N-gram model: $h = w_{q-n+1},\ldots,w_{q-1}$

- Framework for SLM adaptation:



  - The general idea is to dynamically modify the background SLM estimate on the basis of what can be extracted from A

Adaptation data

- Corpus A may already be available
  - Domain-specific texts for some purpose
  - Wizard of Oz experiments

- Corpus A is not available
  - Use Monte Carlo methods for choosing rules at the branching points of the grammar and then create an artificial corpus
  - Use small amount of adaptation data to weigh the different rules and retain more realistic corpus (bootstrap SLM)

Adaptation data (cont.)

- Use multiple sentence hypotheses weighted by posteriori likelihood from an n-best list
  - The effect of a recognition error is distributed over several competing hypotheses
- sausages

- Once enough adaptation material has been accumulated to suitable characterize the new domain
  - To gather similar data by Information Retrieval

- Data augmentation
  - Adjust the count of corpus

Data augmentation

- D. Janiszek, et al, “DATA AUGMENTATION AND LANGUAGE MODEL ADAPTATION”, 2001

- represent words as vectors after SVD
  - Augmented count

$$a_{ij} = c_{ij} + \sum_{h_k} c_{ik} \times f\left(d_{jk}^{\alpha}\right) \text{with } h_k \in \Gamma_j^{\alpha}(\theta), f\left(d_{jk}^{\alpha}\right) = e^{-\frac{d_{jk}^{\alpha}}{D}}$$

a_{ij} : augmented count for word i with history j

c_{ij} : orginal count for word i with history j

d_{jk}^{α} : distance of history j and k

$\Gamma_j^{\alpha}(\theta)$: set of histories having distance lower than threshold θ with history j in the space S_{α}

Model interpolation

- Corpus A is used to derive a task-specific (dynamic) SLM which is then combined with the background (static) SLM
  - Provides fertile grounds

- Model interpolation
  - Model merging
  - Dynamic cache models
  - MAP adaptation

Model merging

- Only for certain idiosyncratic word sequences, particularly frequent in the current task, may the dynamic model outperform the initial estimate SLM
  - Take advantage of the new information as appropriate

- Linear interpolation

$$\Pr\left(w_q \mid h_q\right) = \left(1-\lambda\right)\Pr_A\left(w_q \mid h_q\right) + \lambda \Pr_B\left(w_q \mid h_q\right)$$

- Back-off

$$\Pr\left(w_q \mid h_q\right) = \begin{cases} \Pr_A\left(w_q \mid h_q\right) & \text{if } C_A\left(hq, wq\right) \geq T \\ \beta \Pr_B\left(w_q \mid h_q\right) & \text{otherwise} \end{cases}$$

Dynamic Cache models

- A special case of linear interpolation
- Cache models exploit self-triggering words inside the corpus A to capture short-term shifts in word-use frequencies
  - Unigram case of general model merging strategy
- To propagate the power of the method
  - Class n-gram model

$$\Pr(w_q \mid h_q) = \sum_{\{c_q\}} \Pr(w_q \mid c_q) \Pr(c_q \mid h_q)$$

MAP adaptation

- It is argued that combination should be done at the frequency count level rather than the model level

$$P(W \mid \theta)P(\theta) \propto \prod_{k=1}^{|h|}\prod_{i=1}^{|v|}\phi_{h_k,w_i}^{C_{h_k,w_i}+\nu_{h_k,w_i}-1}$$ Note: Multinomial * Dirichlet

$$\log P(W \mid \theta)P(\theta) \propto \sum_{k=1}^{|h|}\sum_{i=1}^{|v|}\left(C_{h_k,w_i}+\nu_{h_k,w_i}-1\right)\log\phi_{h_k,w_i} + \sum_{k=1}^{|h|} l_{h_k}\left(\sum_{i=1}^{|v|}\phi_{h_k,w_i}-1\right)$$

$$\Pr(w_q \mid h_q) = -$$

$$\phi_{h_k,w_i} = \frac{C_{h_k,w_i}+\nu_{h_k,w_i}-1}{\sum_{j=1}^{|v|}\left(C_{h_k,w_j}+\nu_{h_k,w_j}-1\right)} \qquad ,\nu_{h_k,w_i} = \frac{1}{\varepsilon}C_B\left(h_k,w_i\right)+1$$

- Count merging

$$\Pr(wq \mid hq) = \frac{\varepsilon C_A\left(h_q w_q\right)+C_B\left(h_q w_q\right)}{\varepsilon C_A\left(h_q\right)+C_B\left(h_q\right)}$$

Constraint specification

- Extract feature from corpus A that the adapted SLM is constrained to satisfy

- Exponential models
  - MDI adaptation
  - Unigram constraints

MDI adaptation

- The feature extracted from corpus A are considered as important properties

- Minimize KL distance from background distribution

$$\min_{Q(h,w)} \sum_{\{(h,w)\}} Q(h,w) \log \frac{Q(h,w)}{\Pr_B(h,w)}$$

which satisfies

$$\sum_{\{(h,w)\}} I_k(h,w) Q(h,w) = \alpha_A\left(\hat{h}_k, \hat{w}_k\right), 1 \le k \le \mathbf{K}$$

- Take joint background distribution as prior distribution

$$\Pr(h,w) = \frac{\Pr_B(h,w)}{Z(h,w)} \prod_{k=1}^{K} \exp\{\lambda_k I_k(h,w)\}$$

Unigram Constraints

- Special case of MDI adaptation
  - Only unigram feature are reliably estimated

$$\sum_{\{(h,w)\}} I_k(h,w) Q(h,w) = \alpha_A(\hat{w}_k), 1 \le k \le \mathbf{K}$$

  - Similar to dynamic cache model
  - GIS

$$\Pr(h,w) = \Pr_B(h,w) \frac{\alpha_A(w)}{\Pr_B(w)}$$

  - The adapted SLM is simply the background SLM scaled by the scaling factor

Topic Information

- Exploit the information about the underlying subject matter from corpus A

- Mixture models
  - Linear interpolate K n-gram that the resulting mixture best matches the adaptation data A

$$\Pr\left(w_q \mid h_q\right) = \sum_{k=1}^{K} \lambda_{A,k} \Pr_{B,k}\left(w_q \mid h_q\right)$$

  - One of these background model is trained on the entire corpus B
  - count merge from mixture model with different weight of topics
  - Drawback: inherent fragmentation of training data

Explicit topic models

- Mixture modeling includes topic information indirectly
  - Express the topic contribution more directly
  - No assumption that each history belongs to exactly one topic
  - Conditional independence assumption on word and topic

$$\Pr(w_q \mid h_q) = \sum_{k=1}^{K} \Pr(w_q \mid t_k) \Pr(t_k \mid h_q)$$

  - Topic n-gram is assumed to be unaffected
  - The topic assignment is adapted

$$\Pr(w_q \mid t_k) = (1-\lambda) \Pr_A(w_q \mid t_k) + \lambda \Pr_B(w_q \mid t_k)$$

Trigger

- Rosenfeld, “A maximum entropy approach to adaptive statistical language model”, 1996

- If a word sequence A is significantly correlated with another word sequence B
  - (A→B) is considered a “trigger pair”, with A being the trigger and B the triggered sequence
  - If A,B are single words, possible triggers are more than bigram

- Correlation measure
  - Cross product ratio (not enough in determining the utility)
  - Average utility ? (the more common, the higher)
  - Average mutual information

Trigger (cont.)

- Average mutual information

$$I(A_o : B) = P(A_o, B)\log\frac{P(B \mid A_o)}{P(B)} + P(A_o, \overline{B})\log\frac{P(\overline{B} \mid A_o)}{P(\overline{B})}$$

$$+ P(\overline{A}_o, B)\log\frac{P(B \mid \overline{A}_o)}{P(B)} + P(\overline{A}_o, \overline{B})\log\frac{P(\overline{B} \mid \overline{A}_o)}{P(\overline{B})}$$

Or $$I(A_o : B) = \log\frac{P(A_o, B)}{P(A_o)P(B)}$$

  - Implementation :
    - set the window size (Church et al., Word Association Norms, Mutual Information, and Lexicography, 1990)

Trigger (cont.)

- Carlos Troncoso et al, “Trigger-Based Language Model Adaptation for Automatic Meeting Transcription”, 2005

- Focus on the transcription of meetings and lectures rather than newspaper articles
  - the subject is more homogeneous in meeting session (document)
  - the topic of newspaper is too general to extract task-specific triggers

- Extract trigger from initial speech recognition results

Trigger (cont.)

- Selection methods
  - TD/IDF

$$v_{ik} = \frac{tf_{ik} \log\left(N / df_k\right)}{\sqrt{\sum_{j=1}^{T} \left(tf_{ij}\right)^2 \left[\log\left(N / df_k\right)\right]^2}}$$

  - The TF part is calculated from the K-best hypotheses
  - The IDF part is computed from a fraction of a large corpus

- Create all possible word pairs including self-triggers and then use POS-based filter to discard function words

Trigger (cont.)

- Probability of trigger pairs

$$P_{TP}^{IT}(w_2 \mid w_1) = \frac{N(w_1, w_2)}{\sum_j N(w_1, w_j)}$$

- Proposed trigger-based language model

$$P_{LM}(w_i \mid H) = \frac{1}{|H|} \sum_{h \in H} P_{LM}(w_i \mid h)$$

$$P_{LM}(w_i \mid H) = \begin{cases} P_{NG}(w_i \mid H), \textbf{ if } P_{TP}^{IT}(w_j \mid h) = 0, \forall j \\ \lambda P_{NG}(w_i \mid H) + (1-\lambda) P_{TP}^{IT}(w_j \mid h), \textbf{ else} \end{cases}$$

- Proposed back-off method

$$P_{LM}(w_i \mid H) = \lambda P_{NG}(w_i \mid H) + (1-\lambda)\left(\delta P_{TP}^{LC}(w_i \mid h) + (1-\delta) P_{TP}^{IT}(w_i \mid h)\right)$$

Latent Semantic Analysis

- Bellegarda, “Exploiting latent semantic information in statistical language modeling”, 2000
- The paradigm relies on the concept of a document, i.e., a “bag-of-words” entity forming a semantically homogeneous unit
- The resulting semantic knowledge is encapsulated in a continuous vector space

Latent Semantic Analysis

- Step 1: Build a matrix of co-occurrences between words and documents
  - Normalized entropy

$$\varepsilon_i = -\frac{1}{\log N}\sum_{j=1}^{N}\frac{c_{i,j}}{t_i}\log\frac{c_{i,j}}{t_i}$$

t_i : **the total number of tmes of** w_i **occurs in Corpus**

$c_{i,j}$: **the total number of tmes of** w_i **occurs in document** j

  - Cell of Matrix W

$$w_{i,j} = (1-\varepsilon_i)\frac{c_{i,j}}{n_j}$$

Latent Semantic Analysis (cont.)

- Step 2: Singular Value Decomposition

$$W \approx \hat{W} = USV^T$$

U : **left singular matrix,** $U^T U = I_R$

V : **right singular matrix,** $V^T V = I_R$

S : **diagonal marix,** $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_R \geq 0$

row vector $u_i S, v_j S$ **characterize**

the position of word w_i **and document** d_j **in the space**

Latent Semantic Analysis (cont.)

- Step 3: Create pseudo document vector
  - SVD expansion: $\tilde{d}_p = US\tilde{v}_p^T$
  - pseudo document vector

$$\tilde{\tilde{v}}_p = \tilde{v}_p S = \tilde{d}_p^T U$$

  - If the new document contains language patterns which are inconsistent with those extracted from *W*, the SVD expansion will no longer apply

Latent Semantic Analysis (cont.)

- N-gram + LSA Language Modeling
  - The closeness between word and pseudo document vector

$$\Pr_{LSA}\left(w_q \mid H_{q-1}, S\right) = \Pr_{LSA}\left(w_q \mid \tilde{d}_{q-1}\right)$$

  - d_{q-1} and d_q differ only in one coordinate and the entropy does not change appreciably

$$\tilde{d}_q = \frac{n_q - 1}{n_q}\tilde{d}_{q-1} + \frac{1-\varepsilon_i}{n_q}\left[0 \ldots 1 \ldots 0\right]^T$$

  - Pseudo document vector can be efficiently updated directly in the LSA space

$$\tilde{\tilde{v}}_q = \tilde{v}_q S = \frac{1}{n_q}\left[\left(n_q - 1\right)\tilde{\tilde{v}}_{q-1} + \left(1-\varepsilon_i\right)u_i\right]$$

Latent Semantic Analysis (cont.)

- LSA probability
  - Closeness

$$K\left(w_q, \tilde{d}_{q-1}\right) = \cos\left(u_q S^{1/2}, \tilde{v}_{q-1} S^{1/2}\right) = \frac{u_q S \tilde{v}_{q-1}^T}{\left\| u_q S^{1/2} \right\| \left\| \tilde{v}_{q-1} S^{1/2} \right\|}$$

  - Probability

$$\Pr_{LSA}\left(w_q \mid \tilde{d}_{q-1}\right) = \frac{\pi - \mathbf{acos}(K)}{\pi}$$

  - Integration with N-grams

$$\Pr\left(w_q \mid H_{q-1}^{(n+l)}\right) = \frac{\Pr_{N-gram}\left(w_q \mid w_{q-n+1}, \ldots, w_{q-1}\right) \frac{\Pr_{LSA}\left(w_q \mid d_{q-1}\right)}{\Pr_{N-gram}\left(w_q\right)}}{\sum_{w_i \in V} \Pr_{N-gram}\left(w_i \mid w_{i-n+1}, \ldots, w_{i-1}\right) \frac{\Pr_{LSA}\left(w_i \mid d_{q-1}\right)}{\Pr_{N-gram}\left(w_i\right)}}$$

Latent Semantic Analysis (cont.)

- Adaptation:

$$\frac{\Pr_{LSA}\left(w_q \mid d_{q-1}\right)}{\Pr_{N-gram}\left(w_q\right)} = (1-\lambda)\frac{\Pr_{A-LSA}\left(w_q \mid d_{q-1}\right)}{\Pr_{A-N-gram}\left(w_q\right)} + \lambda\frac{\Pr_{B-LSA}\left(w_q \mid d_{q-1}\right)}{\Pr_{B-N-gram}\left(w_q\right)}$$

- Other issues
  - History size or importance of history

$$\tilde{\tilde{v}}_q = \tilde{v}_q S = \frac{1}{n_q}\left[(n_q - 1)\tilde{\tilde{v}}_{q-1} + (1-\varepsilon_i)u_i\right]$$

  - Word Smoothing

$$\Pr_{LSA}\left(w_q \mid \tilde{d}_{q-1}\right) = \sum_{k=1}^{K}\Pr\left(w_q \mid C_k\right)\Pr\left(C_k \mid \tilde{d}_{q-1}\right)$$

  - Document Smoothing

$$\Pr_{LSA}\left(w_q \mid \tilde{d}_{q-1}\right) = \sum_{l=1}^{L}\Pr\left(w_q \mid D_l\right)\Pr\left(D_l \mid \tilde{d}_{q-1}\right)$$

Latent Dirichlet Allocation

- Yik-Cheung Tam et al. ,"Dynamic Language Model Adaptation using Variational Bayes Inference", 2005

- In broadcast news, a document usually refers to a piece of news story within which the latent topics are consistent

- LDA is a Bayesian extension of a mixture of unigram models where the topic mixture weight θ is drawn from a prior Dirichlet distribution

$$f(\theta;\alpha) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

- LDA makes an assumption in which the words in a document are conditionally independent given the latent topic sequence

$$f(w_1^n \mid \theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} f(w_i \mid z_i = k) \cdot f(z_i = k \mid \theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \beta_{w_i k} \cdot \theta_k$$

Latent Dirichlet Allocation (cont.)

- A document is generated by firstly sampling a mixture weight θ from its prior distribution and then repeatedly sample a topic k from θ and a word from the k-th latent unigram until all words in the document are generated

$$f\left(w_1^n\right) = \int_\theta f\left(w_1^n \mid \theta\right) \cdot f\left(\theta; \alpha\right) d\theta$$

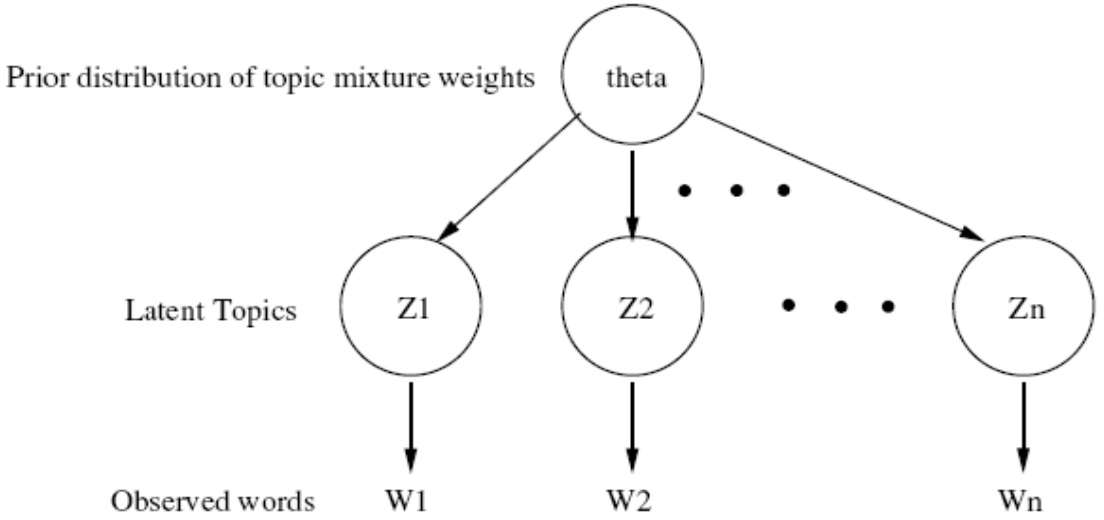


Figure 1: Graphical representation of Latent Dirichlet Allocation.

Variational Bayes approach (cont.)

- Jensen's inequality

$$\log \sum_i q_i \cdot \frac{f_i}{q_i} \geq \sum_i q_i \cdot \log \frac{f_i}{q_i} = E_q\left[\log \frac{f(.)}{q(.)}\right] \textbf{ where } \sum_i q_i = 1$$

- Optimizing the exact likelihood is computationally intractable
  - optimize the lower bound

$$\log f\left(w_1^n\right) = \log \sum_{\theta, z_1^n} q\left(\theta, z_1^n; \Gamma\right) \frac{f\left(\theta, w_1^n, z_1^n; \Lambda\right)}{q\left(\theta, z_1^n; \Gamma\right)}$$

$$\geq \sum_{\theta, z_1^n} q\left(\theta, z_1^n; \Gamma\right) \log \frac{f\left(\theta, w_1^n, z_1^n; \Lambda\right)}{q\left(\theta, z_1^n; \Gamma\right)} = E_q\left[\log \frac{f\left(\theta, w_1^n, z_1^n; \Lambda\right)}{q\left(\theta, z_1^n; \Gamma\right)}\right] = Q(\Lambda, \Gamma)$$

- To choose the tractable variational distribution that the integration can be done in a tractable manner (interpret q an approximation to the posteriori distribution over f)

$$\Lambda = \{\{\alpha_k\}, \{\beta_{vk}\}\}, \quad q\left(\theta, z_1^n; \Gamma\right) = q(\theta) \cdot \prod_{i=1}^{n} q(z_i)$$

Dirichlet * multinomial

Variational Bayes approach (cont.)

- Parameters for a single document

$$\gamma_k = \alpha_k + \sum_{i=1}^{n} q(z_i = k)$$

$$q(z_i = k) \propto \beta_{w_i k} \cdot e^{E_q[\log \theta_k]}$$

$$\beta_{vk} \propto \sum_{i=1}^{n} q(z_i = k)\delta(w_i, v)$$

- Adaptation
  - Compute the Maximum A-posterior likelihood

$$f(w \mid h) \approx \int_\theta \sum_{k=1}^{K} f(w \mid z = k) f(z = k \mid \theta) q(\theta \mid h) \approx \sum_{k=1}^{K} \beta_{wk} \cdot \hat{\theta}_k, \hat{\theta}_k = \frac{\gamma_k}{\sum_{k=1}^{K} \gamma_k}$$

  - $\hat{\theta}$ is the mode of the variational posteriori

Syntactic infrastructure

- Background and recognition tasks share a common grammatical infrastructure
  - Grammatical constraints are portable from corpus B to corpus A
- Structured language models
  - Hierarchical framework by using syntactic information to determine equivalence classes on the n-gram history

$$\Pr\left(w_q \mid \overline{h}_q\right) = \frac{1}{Z\left(\overline{h}_q\right)} \sum_{\{\pi_q\}} \Pr\left(w_q \mid \overline{h}_q, \pi_q\right) \Pr\left(\overline{h}_q, \pi_q\right)$$

  - Each parse gives rise to a unique headword history, and it is expedient to simplify the model by limit to n-1 headwords
- Main caveat in structured language modeling remain the reliance on the parser, and particularly the implicit assumption that the correct parse will be assigned a high probability

Syntactic infrastructure (cont.)

- Syntactic triggers
  - Exploit syntactic structure contained in previous sentences
  - Full parse or the syntactic/semantic tag

- Although not yet implemented in an adaptation context, this concept may ultimately provide the necessary framework to extend the benefits of structured language modeling to a span greater than that of a sentence

Multiple sources

- Whole sentence models
  - It has been argued hindrance to modeling linguistic supra-structure such as person and number agreement, semantic coherence, and even length
  - Adopt a “bag-of-features” approach to each sentence

$$\Pr(\sigma) = \frac{\Pr_0(\sigma)}{Z} \prod_{k=1}^{K} \exp\{\lambda_k I_k(\sigma)\}$$

  - Normalization is infeasible since it involves summation

Conclusions

- Language model adaptation refers to the process of exploiting specific, albeit limited, knowledge about recognition task to compensate for any mismatch between and recognition

- Dynamically modifying the language model statistics according to this information

- Several techniques are surveyed