Document Representations for Topic-Adaptation in Statistical Language Modeling

Sanjeev Khudanpur

Dept of Electrical and Computer Engineering

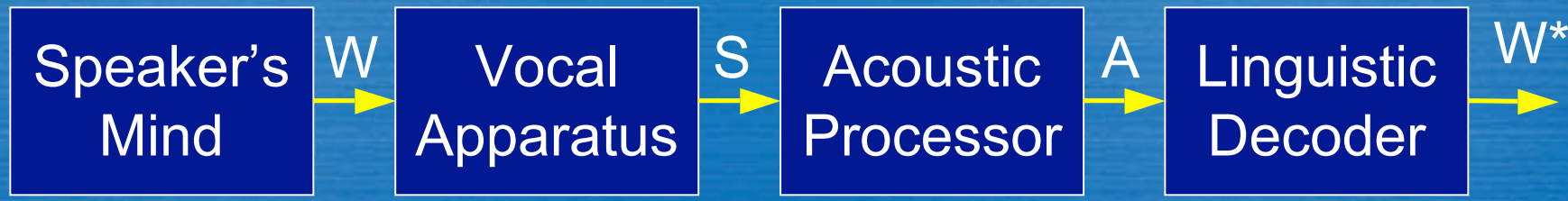
and

Center for Language and Speech Processing

Johns Hopkins University

Statistical Language Modeling

- Key human language technology component
  - Automatic speech recognition (ASR)
  - Statistical machine translation (MT)
  - Information retrieval (IR)
  - Spelling correction



$$W^* = \arg\max_{W \in \mathrm{W}} P(W|A) = \arg\max_{W \in \mathrm{W}} P(A|W)P(W)$$

Commonly Used Language Models: N-grams

- Model word sequences as realizations of a Markov (finite memory) processes
  - $P(w_1,w_2,\ldots,w_n) \approx \prod P(w_t|w_{t-1},w_{t-2})$
- Conditional probabilities estimated from data
  - Nonparametric estimates are popular
- Acute data sparseness is encountered even for very short conditioning history-lengths
  - Vocabulary sizes are typically 10,000 to 100,000
  - Roughly 50% of the trigrams are seen only once or twice in the training text
- Statistical dependence clearly extends beyond trigram or 4-gram range
  - Syntactic dependencies in a sentence and topic-depdencies

Modeling Topic Dependence

- Cluster training text / documents into coherent topics
  - Estimate topic-specific language models
    - $P(w_t|w_{t-1},w_{t-2},w_{t-3},\ldots) \approx P_{topic}(w_t|w_{t-1})$ or $P(w_t|w_{t-1},topic)$
  - Results in further aggravating data sparseness
  - Usual remedy is linear interpolation
    - $P(w_t|w_{t-1}, w_{t-2},w_{t-3},\ldots) = \lambda P_{topic}(w_t|w_{t-1}) + (1-\lambda)P(w_t|w_{t-1})$
  - Topic clusters may be created manually or automatically
- Topic assignment of a test utterance
  - Obtained automatically by "classifying" the history
    - $\Phi(w_{t-1},w_{t-2},w_{t-3},\ldots) = \arg\min_i \text{distance}(\{w_{t-1},w_{t-2},w_{t-3},\ldots\},\text{cluster}(i))$
    - $P(w_t|w_{t-1}, w_{t-2},w_{t-3},\ldots) = \lambda P_{\Phi(wt-1,wt-2,wt-3,\ldots)}(w_t|w_{t-1}) + (1-\lambda)P(w_t|w_{t-1})$
  - Or integrated away using a probabilistic topic assignment
    - $\lambda[\Sigma_{topics}\ P_{topic}(w_t|w_{t-1})P(\text{topic}|w_{t-1},w_{t-2},w_{t-3},\ldots)] + (1-\lambda)P(w_t|w_{t-1})$

Word-Document Co-occurrence Matrix

- A vocabulary $\{w_1, w_2, \ldots, w_M\}$
- A document collection $\{d_1, d_2, \ldots, d_N\}$
- The raw co-occurrence count c_{ij} of w_i in d_j
- Relative uncertainty about the document's identity j given that it contains the word w_i

$$\varepsilon_i = -\frac{1}{\log N} \sum_{j=1}^{N} \frac{c_{ij}}{c_i} \log \frac{c_{ij}}{c_i}$$

- The weighted co-occurrence count of w_i in d_j

$$\omega_{ij} = \left(1 - \varepsilon_i\right) \frac{c_{ij}}{c_j}$$

- Columns of the matrix $[\omega_{ij}]$ represent the training documents

Creating Document Clusters and Assigning Topic at Test-time

- Use manually assigned topic clusters if available, or
- Use K-means or any other clustering scheme to cluster the column-vectors of the matrix $[\omega_{ij}]$
  - Initialize with K arbitrary "centroid" vectors
  - Assign each column/document to nearest centroid
  - Recompute centroids based on vectors assigned to them
- Cosine similarity is preferred over Euclidean distance

$$sim(d_j, d_k) = \frac{\vec{\omega}_j \cdot \vec{\omega}_k}{\|\vec{\omega}_j\| \times \|\vec{\omega}_k\|}$$

- At test-time a pseudo-document d_t is created from the history $w_{t-1}, w_{t-2}, w_{t-3}, \ldots$, and used like an ω vector

Indicative Empirical Results

- The switchboard conversational speech corpus
  - Strangers talking via telephone on assigned topics
  - Spontaneous human-human speech
  - Still the hardest of speech recognition problems
  - Training: 2.1M words, 70 topics; Test 18K words, 22K vocab
- Perplexity: $2^{-\frac{1}{n}\log_2 P(w_1,w_2,\ldots,w_n)} = 2^{-\frac{1}{n}\sum_{t=1}^{n}\log_2 P(w_t|w_{t-1},w_{t-2},\ldots)}$

Language Model	Perplexity
$P(w_t \mid w_{t-1},w_{t-2})$	83.4
$\lambda P_{\Phi(.)}(w_t) + (1-\lambda)P(w_t \mid w_{t-1},w_{t-2})$	83.0
$\lambda P_{\Phi(.)}(w_t \mid w_{t-1}) + (1-\lambda)P(w_t \mid w_{t-1},w_{t-2})$	78.8
$\lambda P_{\Phi(.)}(w_t \mid w_{t-1},w_{t-2}) + (1-\lambda)P(w_t \mid w_{t-1},w_{t-2})$	77.6

A Maximum Entropy Alternative to Linear Interpolation

- The previous model treats content words, whose probability varies with topic, and grammatical function words the same way
  - $P(w_t|w_{t-1}, w_{t-2},w_{t-3},\ldots) = \lambda P_{\Phi(wt-1,wt-2,wt-3,\ldots)}(w_t|w_{t-1}) + (1-\lambda)P(w_t|w_{t-1})$
- For each w word in the vocabulary and each topic τ, if

$$\left|\log\frac{f(w|\tau)}{f(w)}\right| = \left|\log\frac{\#[w,\tau]/\#[\tau]}{\#[w]/N}\right| > \eta$$

  then declare w a topic-unigram
- Collect corpus-wide relative frequencies for N-grams and topic-conditional frequencies of topic-unigrams
  - Construct a family of models that agree with these "marginals"

An Exponential Family

$$P(w,x,y) = f(w,x,y), \quad P(x,y) = f(x,y), \quad P(y) = f(y), \quad P(y,\tau) = f(y,\tau)$$

- Define all joint pmfs on 4-tuples (*w,x,y,τ*) to be admissible that agree with the observed marginals
  - Defines a linear family of pmfs
- Choose from the family the pmf with the highest entropy
  - Entails computing the partition function on 4-tuples
  - But we mainly care for the conditional pmfs P(y|w,x,τ)
  - This makes for a significant simplifying assumption

$$P(y \mid w,x,\tau) = \frac{\exp\left(\sum \lambda_j g_j(w,x,y,\tau)\right)}{Z(w,x,\tau)} = \frac{\alpha_y^{1(w_t = y)} \alpha_{x,y}^{1(w_{t-1}, w_t = x,y)} \alpha_{w,x,y}^{1(w_{t-2}, w_{t-1}, w_t = w,x,y)} \alpha_{y,\tau}^{1(w_t, \Phi_t = y,\tau)}}{Z(w,x,\tau)}$$

Empirical Results: Max-Ent LMs

- Treat each conversation side as a document: N = 3,000
- Use a preexisting vocabulary: M = 22,000
- Perform K-means with 70-odd centroids (originally 70 topics)
- Train a max-ent model with N-gram and topic-unigram features
- Assign a (potentially new) topic-centroid for each test utterance

Language Model	Perplexity	WER
Baseline Trigram (SRI LM)	78.8	38.5%
Maximum Entropy Trigram	78.9	38.3%
Maximum Entropy with Topic-Unigram Features	73.6	37.9%

Language Model	Topic-unigrams	Others	Content-words	Stop-words
ME Trigram	3936	63.9	225	58.8
ME+Topic-unigram	354	68.2	177	57.2

Let Us Return to the Matrix of Word-Document Co-occurrences

- A vocabulary $\{w_1, w_2, \ldots, w_M\}$
- A document collection $\{d_1, d_2, \ldots, d_N\}$
- The raw co-occurrence count c_{ij} of w_i in d_j
- Relative uncertainty about the document's identity j given that it contains the word w_i

$$\varepsilon_i = -\frac{1}{\log N} \sum_{j=1}^{N} \frac{c_{ij}}{c_i} \log \frac{c_{ij}}{c_i}$$

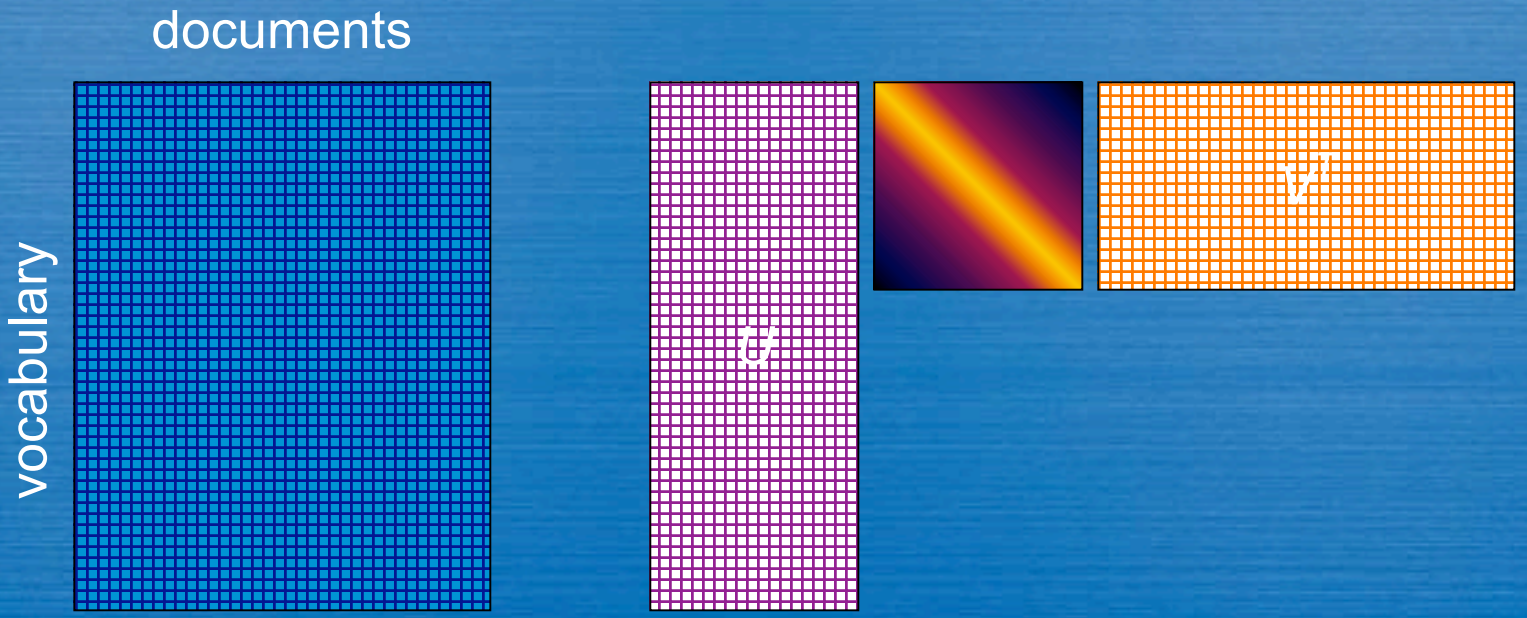
- The weighted co-occurrence count of w_i in d_j

$$\omega_{ij} = \left(1 - \varepsilon_i\right) \frac{c_{ij}}{c_j}$$

Singular Value Decomposition

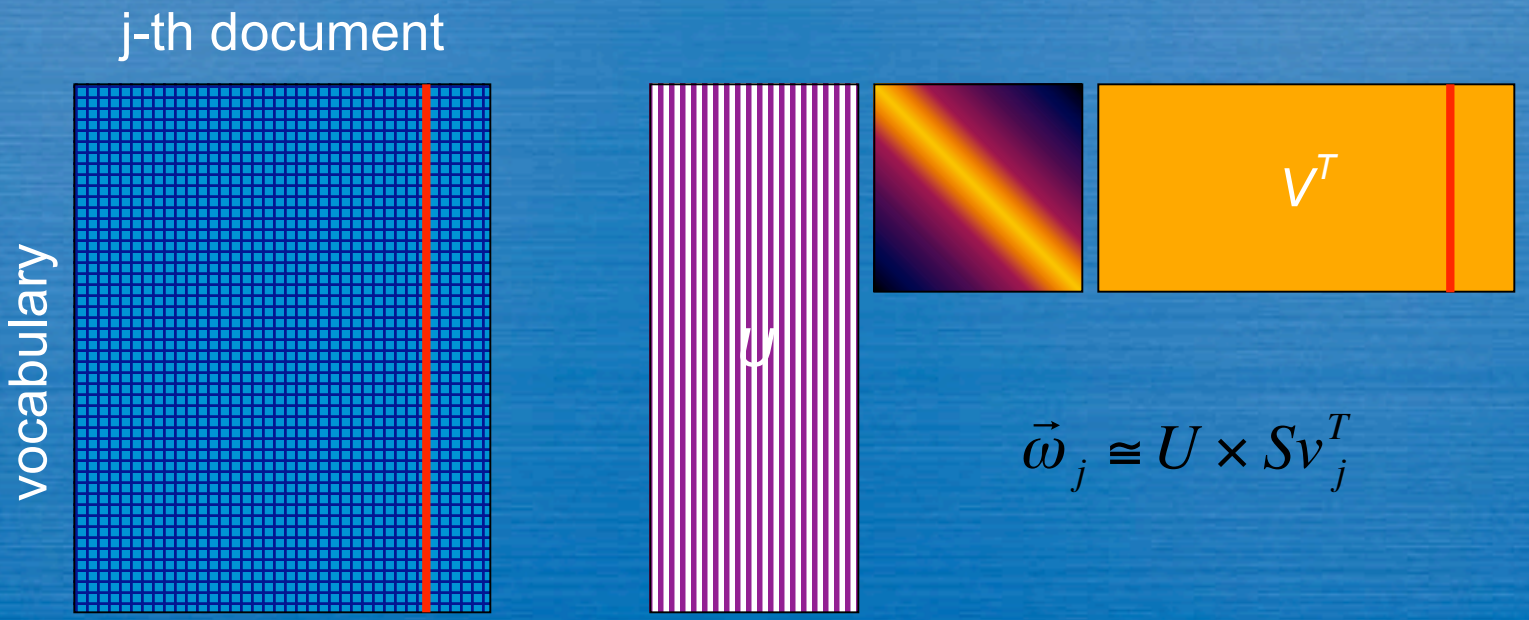
- A rank R approximation of $\Omega_{M\times N}$, for an $R \ll \min\{M,N\}$

$$\Omega_{M\times N} = \left[\omega_{ij}\right] \approx U_{M\times R} \times S_{R\times R} \times V^{T}_{N\times R}$$



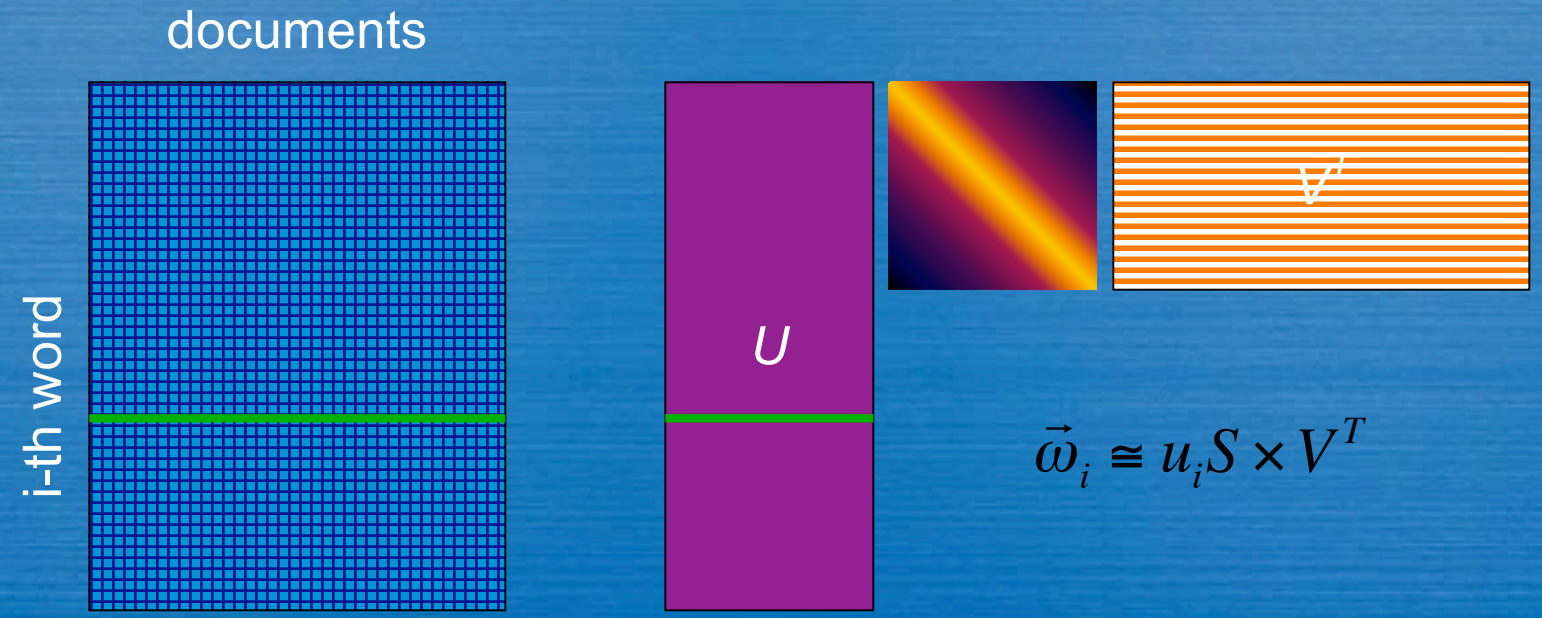
A Low-Dimensional Document Representation: $d_j \rightarrow v_j S$

- Think of each **document** as a linear combination of a few proto- or meta-documents, each representing a single "**concept**" or "**topic**"



A Low-Dimensional Word Representation: $w_i \rightarrow u_i S$

- Think of each **word** as a linear combination of a few proto- or meta-words, each representing a single "**concept**" or "**sense**"



$$\vec{\omega}_i \cong u_i S \times V^T$$

Proximity of Representations and Semantic Similarity

- Similarity between two documents d_j and $d_{j'}$ may be measured by the cosine of the angle between their representative vectors

$$K(d_j, d_{j'}) = \frac{v_j S \cdot v_{j'} S}{\|v_j S\| \times \|v_{j'} S\|} = \frac{v_j S^2 v_{j'}}{\|v_j S\| \times \|v_{j'} S\|}$$

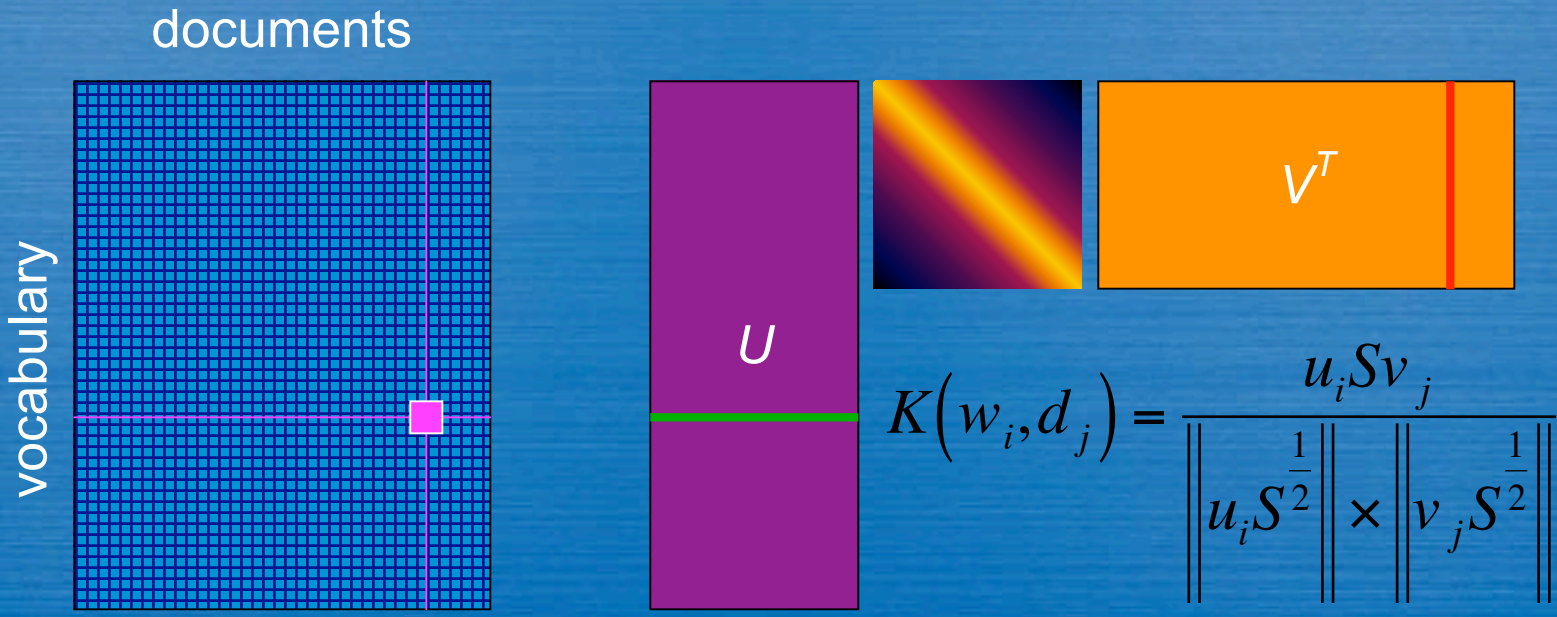
- Similarly, for two words w_i and $w_{i'}$

$$K(w_i, w_{i'}) = \frac{u_i S \cdot u_{i'} S}{\|u_i S\| \times \|u_{i'} S\|} = \frac{u_i S^2 u_{i'}}{\|u_i S\| \times \|u_{i'} S\|}$$

Proximity of Representations and Semantic Similarity (Cont'd)

- The SVD approximates elements of Ω as

$$\omega_{ij} \approx u_i \times S \times v_j^T$$



$$K(w_i, d_j) = \frac{u_i S v_j}{\left\| u_i S^{\frac{1}{2}} \right\| \times \left\| v_j S^{\frac{1}{2}} \right\|}$$

Re-enter, Language Modeling

- Given a word-sequence $w_1, w_2, \ldots, w_{t-1}$,

$$P\left(w_t \middle| w_1, w_2, \cdots, w_{t-1}\right) \cong P\left(w_t \middle| w_{t-2}, w_{t-1}\right)$$

- What may the longer context offer?

$$\tilde{d}_{t-1} \equiv \left\langle w_1, w_2, \cdots, w_{t-1} \right\rangle; \qquad P_{\mathrm{LSA}}\left(w_t \middle| \tilde{d}_{t-1}\right) \stackrel{?}{=} \frac{K\left(w_t, \tilde{d}_{t-1}\right)}{\sum_w K\left(w, \tilde{d}_{t-1}\right)}$$

$$P_{\mathrm{LSA}}\left(w_t \middle| \tilde{d}_{t-1}\right) = \frac{\left[K\left(w_t, \tilde{d}_{t-1}\right) - K_{\min}\left(\tilde{d}_{t-1}\right)\right]^{\gamma}}{\sum_w \left[K\left(w, \tilde{d}_{t-1}\right) - K_{\min}\left(\tilde{d}_{t-1}\right)\right]^{\gamma}}$$

Combining P_{LSA} with N-grams: Part 1A

- Linear Interpolation

$$P\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) = \alpha P_{\text{LSA}}\left(w_t \middle| \tilde{d}_{t-1}\right) + (1-\alpha) P_{\text{N-gram}}\left(w_t \middle| w_{t-1}, w_{t-2}\right)$$

- Information Weighted Arithmetic Mean

$$P\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) = \frac{\lambda_{w_t} P_{\text{LSA}}\left(w_t \middle| \tilde{d}_{t-1}\right) + \left(1-\lambda_{w_t}\right) P_{\text{N-gram}}\left(\text{w}_\text{t} \middle| \text{w}_\text{t-1}, \text{w}_\text{t-2}\right)}{\sum_w \lambda_w P_{\text{LSA}}\left(w \middle| \tilde{d}_{t-1}\right) + \left(1-\lambda_w\right) P_{\text{N-gram}}\left(w \middle| w_{t-1}, w_{t-2}\right)}$$

with $\lambda_w = \frac{1-\varepsilon_w}{2}$

Combining P_{LSA} with N-grams: Part 1B

- Similarity Modulated N-grams

$$P\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) = \frac{\underline{K}\left(w_t, \tilde{d}_{t-1}\right) P_{\text{N-gram}}\left(w_t \middle| w_{t-1}, w_{t-2}\right)}{\sum_w \underline{K}\left(w, \tilde{d}_{t-1}\right) P_{\text{N-gram}}\left(w \middle| w_{t-1}, w_{t-2}\right)}$$

- Information Weighted Geometric Mean

$$P\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) = \frac{P_{\text{LSA}}^{\lambda_{w_t}}\left(w_t \middle| \tilde{d}_{t-1}\right) \times P_{\text{N-gram}}^{\left(1-\lambda_{w_t}\right)}\left(w_t \middle| w_{t-1}, w_{t-2}\right)}{\sum_w P_{\text{LSA}}^{\lambda_w}\left(w \middle| \tilde{d}_{t-1}\right) \times P_{\text{N-gram}}^{\left(1-\lambda_w\right)}\left(w \middle| w_{t-1}, w_{t-2}\right)}$$

with $\lambda_w = \frac{1-\varepsilon_w}{2}$

Combining P_{LSA} with N-grams: Part 2

- Define $f_{\mathrm{LSA}}\left(w_t, \tilde{d}_{t-1}\right) = K\left(w_t, \tilde{d}_{t-1}\right)$
- Define a parametric model family

$$P_{\underline{\alpha}}\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) = \frac{\alpha_{w_t}^{f_1(w_t)} \times \alpha_{w_t, w_{t-1}}^{f_2(w_t, w_{t-1})} \times \alpha_{w_t, w_{t-1}, w_{t-2}}^{f_3(w_t, w_{t-1}, w_{t-2})} \times \alpha_{w_t, \tilde{d}_{t-1}}^{f_{\mathrm{LSA}}(w_t, \tilde{d}_{t-1})}}{Z_{\underline{\alpha}}\left(w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right)}$$

- Estimate model parameters via maximum likelihood

$$\hat{\underline{\alpha}} = \arg\max_{\underline{\alpha}} P_{\underline{\alpha}}\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) \hat{P}\left(w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right)$$

- … but $\tilde{d}_{t-1}$ takes continuous values …

Parameter Tying Option 2A

- Use a single α for all (w,d) pairs

$$\alpha_{w,d} = \alpha \quad \forall w, \forall d$$

$$P_{\underline{\alpha}}\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) = \frac{\alpha_{w_t}^{f_1(w_t)} \times \alpha_{w_t,w_{t-1}}^{f_2(w_t,w_{t-1})} \times \alpha_{w_t,w_{t-1},w_{t-2}}^{f_3(w_t,w_{t-1},w_{t-2})} \times \alpha^{f_{\mathrm{LSA}}\left(w_t,\tilde{d}_{t-1}\right)}}{Z_{\underline{\alpha}}\left(w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right)}$$

- Directly comparable to similarity modulated N-grams

$$P\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) = \frac{\underline{K}\left(w_t, \tilde{d}_{t-1}\right) P_{\mathrm{N-gram}}\left(w_t \middle| w_{t-1}, w_{t-2}\right)}{\sum_w \underline{K}\left(w, \tilde{d}_{t-1}\right) P_{\mathrm{N-gram}}\left(w \middle| w_{t-1}, w_{t-2}\right)}$$

Parameter Tying Option 2B

- Use a separate α per word w

$$\alpha_{w,d} = \alpha_w \quad \forall d$$

$$P_{\underline{\alpha}}\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) = \frac{\alpha_{w_t}^{f_1(w_t)} \times \alpha_{w_t,w_{t-1}}^{f_2(w_t,w_{t-1})} \times \alpha_{w_t,w_{t-1},w_{t-2}}^{f_3(w_t,w_{t-1},w_{t-2})} \times \alpha_{w_t}^{f_{\text{LSA}}(w_t,\tilde{d}_{t-1})}}{Z_{\underline{\alpha}}\left(w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right)}$$

- Comparable to information weighted geometric mean

$$P\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) = \frac{P_{\text{LSA}}^{\lambda_{w_t}}\left(w_t \middle| \tilde{d}_{t-1}\right) \times P_{\text{N-gram}}^{(1-\lambda_{w_t})}\left(w_t \middle| w_{t-1}, w_{t-2}\right)}{\sum_w P_{\text{LSA}}^{\lambda_w}\left(w \middle| \tilde{d}_{t-1}\right) \times P_{\text{N-gram}}^{(1-\lambda_w)}\left(w \middle| w_{t-1}, w_{t-2}\right)}$$

Parameter Tying Option 2C

- Quantize the document representation vector

$$\alpha_{w,d} = \alpha_{w,\hat{d}} \quad \forall d \in \Phi\left(\hat{d}\right)$$

$$P_{\underline{\alpha}}\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) = \frac{\alpha_{w_t}^{f_1(w_t)} \times \alpha_{w_t,w_{t-1}}^{f_2(w_t,w_{t-1})} \times \alpha_{w_t,w_{t-1},w_{t-2}}^{f_3(w_t,w_{t-1},w_{t-2})} \times \alpha_{w_t,\hat{d}}^{f_{\mathrm{LSA}}(w_t,\tilde{d}_{t-1})}}{Z_{\underline{\alpha}}\left(w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right)}$$

- The *Voronoi* regions and topic-centroids may be obtained by K-means clustering, using LSA-similarity as a "metric"

$$\Phi\left(\hat{d}\right) = \left\{\tilde{d} : K\left(\tilde{d}, \hat{d}\right) \leq K\left(\tilde{d}, \hat{d}'\right) \quad \forall \text{ centroids } \hat{d}' \neq \hat{d}\right\}$$

Further Simplifications in 2C

- Use the nearest topic-centroid for robustness

$$f_{\mathrm{LSA}}(w,d) \cong K\left(w,\hat{d}\right) \qquad \forall d \in \Phi\left(\hat{d}\right)$$

- Use LSA features only for words with "high IDF"

$$f_{\mathrm{LSA}}(w,d) \cong \begin{cases} K\left(w,\hat{d}\right) & \text{if } \varepsilon_w < \tau \\ 0 & \text{otherwise} \end{cases}$$

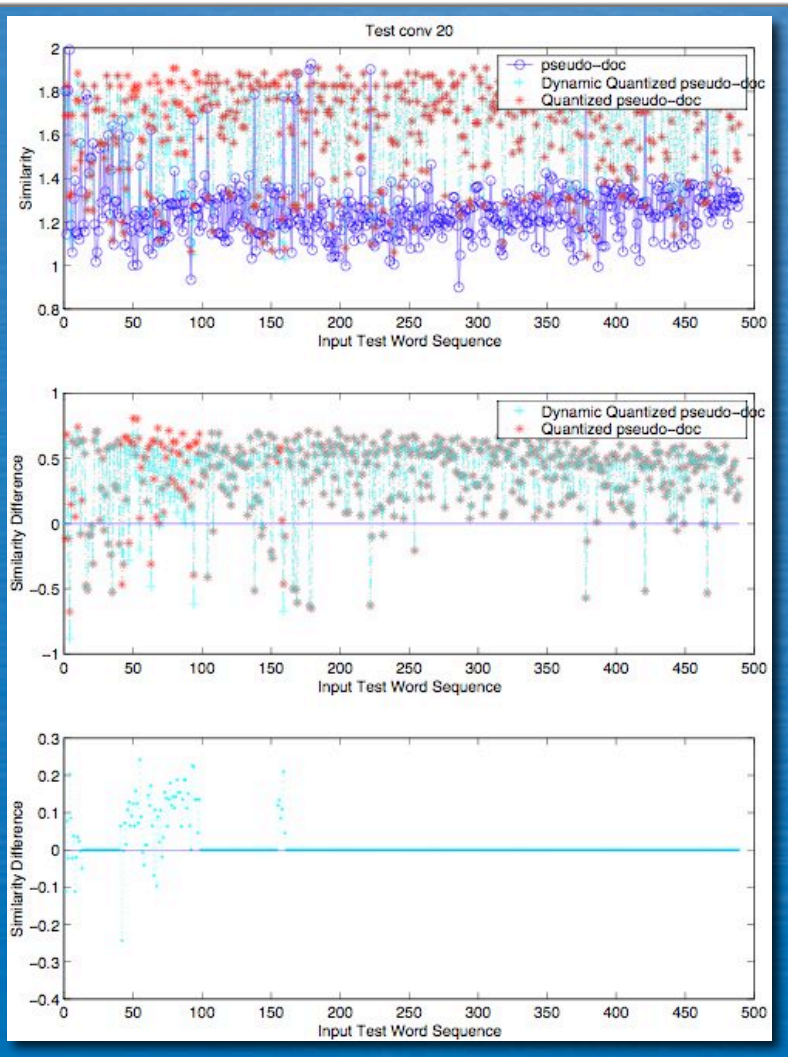
- Quantize the feature function as well, say, to 0 or 1
  - Results in identical model families for the parameter tying of 2C

$$\alpha_{w,\hat{d}}^{0 \text{ or } K\left(w,\hat{d}\right)} = \left(\alpha_{w,\hat{d}}^{K\left(w,\hat{d}\right)}\right)^{0 \text{ or } 1} \equiv \alpha_{w,\hat{d}}^{0 \text{ or } 1}$$

  - Leads to a simpler implementation

Effect of Quantizing $\tilde{d}_{t-1}$

- Similarity between the next word w_t and the
  - true pseudo-document
  - quantized pseudo-document
- Similarity difference
  - Quantized minus true $\tilde{d}_{t-1}$
- Topic assignment on-the-fly vs retrospectively
  - Similarity difference as the conversation evolves



Isn't This a Standard Topic-Dependent Language Model?

- After all that simplification,

$$P_{\underline{\alpha}}\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) = \frac{\alpha_{w_t}^{f_1(w_t)} \times \alpha_{w_t,w_{t-1}}^{f_2(w_t,w_{t-1})} \times \alpha_{w_t,w_{t-1},w_{t-2}}^{f_3(w_t,w_{t-1},w_{t-2})} \times \alpha_{w_t,\hat{d}_{t-1}}^{f_{\text{LSA}}(w_t,\hat{d}_{t-1})}}{Z_{\underline{\alpha}}\left(w_{t-1}, w_{t-2}, \hat{d}_{t-1}\right)}$$

- … looks exactly like a standard topic-dependent maximum entropy language model

$$P_{\underline{\alpha}}\left(w_t \middle| w_{t-1}, w_{t-2}, \mathrm{t}_{t-1}\right) = \frac{\alpha_{w_t}^{f_1(w_t)} \times \alpha_{w_t,w_{t-1}}^{f_2(w_t,w_{t-1})} \times \alpha_{w_t,w_{t-1},w_{t-2}}^{f_3(w_t,w_{t-1},w_{t-2})} \times \alpha_{w_t,\mathrm{t}_{t-1}}^{f_{\text{Topic}}(w_t,\mathrm{t}_{t-1})}}{Z_{\underline{\alpha}}\left(w_{t-1}, w_{t-2}, \mathrm{t}_{t-1}\right)}$$

Parameter Tying Option 2D

- Cluster the vocabulary as well

$$\alpha_{w,d} = \alpha_{\hat{w},\hat{d}} \quad \forall w \in \Psi(\hat{w}), \quad \forall d \in \Phi(\hat{d})$$

$$P_{\underline{\alpha}}\left(w_t \middle| w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right) = \frac{\alpha_{w_t}^{f_1(w_t)} \times \alpha_{w_t,w_{t-1}}^{f_2(w_t,w_{t-1})} \times \alpha_{w_t,w_{t-1},w_{t-2}}^{f_3(w_t,w_{t-1},w_{t-2})} \times \alpha_{\hat{w}_t,\hat{d}}^{f_{\mathrm{LSA}}(w_t,\tilde{d}_{t-1})}}{Z_{\underline{\alpha}}\left(w_{t-1}, w_{t-2}, \tilde{d}_{t-1}\right)}$$

  - The word-clusters may be obtained by a K-means algorithm, again using LSA-similarity as a “metric”
  - Other ontologies, e.g. WordNet, may also be used

- This offers the opportunity to constrain expectations of word-classes given a topic-centroid.

Empirical Results: Reminder

- Switchboard conversational telephone speech corpus
- 2.1M word LM training text; 70-odd “suggested” topics, manually labeled
- 60-odd hours of acoustic training data
- No speaker adaptation, bells or whistles
- 18,000 word test set

Language Model	Perplexity	WER
Baseline Trigram (SRI LM)	78.8	38.5%
Maximum Entropy Trigram	78.9	38.3%

More Empirical Results: N-grams + LSA via Max-Ent

- Treat each conversation side as a document: N = 3,000
- Use a preexisting 20K vocabulary: M = 22,000
- Perform SVD up to 70-odd singular values (originally 70 topics)
- Cluster training documents to obtain 25 topic-clusters $\Phi(d)$
- Train a max-ent model with N-gram and LSA features
- Assign a (potentially new) topic-centroid for each test utterance

Language Model	Perplexity	WER
Baseline Trigram (SRI LM)	78.8	38.5%
Maximum Entropy Trigram	78.9	38.3%
Maximum Entropy with LSA Features	73.6	37.9%

LSA vs Topic-Dependent LMs

Attribute	Topic ME	LSA ME
Similarity measure used	Cosine	
Document clustering algorithm	K-means	
Vector Space Dimension	22,000	73
Number of topic centroids in Φ	67	25
Average number of topics per word	1.8	1.3
Average number of topic-parameters	15,500	19,000
Model perplexity	73.5	73.6
Word Error Rate	37.9%	

Concluding Remarks

- The maximum-entropy models and techniques presented seem to be a reasonable way to incorporate topic-information into a statistical language model
  - Previously proposed models appear to be ad hoc variants
- Only particular special cases have been studied experimentally
  - More interesting options need to be explored
- Word-classes have not been exploited
  - Need to look into WordNet, other ontologies, automatic word-clusters
- A comparison with trigger models may be worthwhile too
  - Found comprable and complementary empirical gains in a closely related problem

Max-Ent vs Previous Models:
A Very Rough Comparison of Perplexity Results

Language Model	SRI LM	CMU LM
Baseline Trigram	78.8	81.1
Maximum Entropy Trigram	78.9	
Maximum Entropy with LSA Features	73.6	
LSA + Trigram Linear Interpolation		81.8
LSA + Trigram Similarity Modulated		79.1
Information Weighted Arithmetic Mean		81.8
Information Weighted Geometric Mean		75.8

Acknowledgements to Coauthors and Coworkers

- Jun Wu (PhD advisee, @ Google)
  - Max-ent implementation, topic-dependent models
- Woosung Kim (PhD advisee, @ Convergys)
  - SVD implementation for language modeling
- Yonggang Deng (GRA, @ Google)
  - LSA + Max-ent combination

- Bill Byrne (Faculty Colleague, @ Cambridge, UK)
  - Acoustic modeling and set-up of rescoring experiments
- Radu Florian (GRA, @ IBM)
  - Topic clustering experiments on Broadcast News documents