Proceedings of HLT-NAACL 2003
Main Papers , pp. 56-63
Edmonton, May-June 2003

Latent Semantic Information in Maximum Entropy Language Models for Conversational Speech Recognition

Yonggang Deng and Sanjeev Khudanpur
Center for Language and Speech Processing
The Johns Hopkins University
3400 North Charles Street, Baltimore, MD 21218, U.S.A.
{dengyg, khudanpur}@jhu.edu

Abstract

Latent semantic analysis (LSA), first exploited in indexing documents for information retrieval, has since been used by several researchers to demonstrate impressive reductions in the perplexity of statistical language models on text corpora such as the Wall Street Journal. In this paper we present an investigation into the use of LSA in language modeling for conversational speech recognition. We find that previously proposed methods of combining an LSA-based unigram model with an N-gram model yield much smaller reductions in perplexity on speech transcriptions than has been reported on written text. We next present a family of exponential models in which LSA similarity is a *feature* of a word-history pair. The maximum entropy model in this family yields a greater reduction in perplexity, and statistically significant improvements in recognition accuracy over a trigram model on the Switchboard corpus. We conclude with a comparison of this LSA-featured model with a previously proposed topic-dependent maximum entropy model.

1 Introduction

Statistical language modeling benefits greatly from the augmentation of standard N-gram statistics with information from the syntactic structure of the sentence and the semantic context of the segment or story being processed, as witnessed by improved performance of automatic speech recognition systems that use such models. In highly constrained settings such as a telephone-based interactive voice-response system, sometimes called a dialogue system, it may be reasonable to limit the notion of syntax to finite state grammars, while the notion of semantics may be adequately captured by a dialogue-state variable representing the type of sentence that may be spoken next by a user. In less constrained speech recognition tasks, e.g. transcription of Broadcast News or conversational telephone speech, the incorporation of syntactic information is usually via a statistical left-to-right parser, while semantic information is usually brought in through some notion of topicality or "aboutness" of the sentence being processed. It is this latter notion of semantics in statistical language modeling that is the subject of this paper.

Collocation or N-gram statistics prove to be one of the best predictors of words in a sentence, and all attempts to augment a language model (LM) with semantic information aim to simultaneously conform to N-gram statistics in one form or another. The straightforward technique (Iyer and Ostendorf, 1999; Clarkson and Robinson, 1998) is to

1. group documents or stories from a putatively large LM training corpus into semantically cohesive clusters using an information retrieval based notion of document similarity,
2. estimate N-gram LMs for each cluster, and
3. interpolate the topic-specific N-gram model with an N-gram model estimated from the undivided LM training corpus.

Alternatives to this method fall into two broad categories, one based on latent semantic analysis (LSA), *e.g.*, Coccaro and Jurafsky (1998) and Bellegarda (2000), and another based on maximum entropy, *e.g.*, Chen and Rosenfeld (1998) and Khudanpur and Wu (1999). In this paper, we attempt to find a bridge between these two techniques.

The starting point of LSA is the construction of a matrix describing word-document co-occurrence.

By performing singular value decomposition of this matrix, a short vector representation is derived for each word and document. One advantage of the resulting word and document representations is that they all live in the same low-dimensional continuous vector space, enabling one to quantitatively measure closeness or similarity between words and documents. The cosine of the angle between two vectors is a standard measure of similarity in this framework.

For language modeling, a pseudo-document is constructed from (possibly all) the words preceding a particular position in an utterance and the resulting vector is projected into the abovementioned low-dimensional vector space, sometimes referred to as the LSA-space. Intuition suggests that words with vectors close to the pseudo-document vector are more likely to follow than those far away from it. This is used to construct a conditional probability on the task-vocabulary. This probability, which depends on a long span of "history" is then suitably combined with an N-gram probability.

An alternative to first constructing a conditional probability on the task-vocabulary independently of the N-gram model and then seeking ways to combine the two probabilities, is directly modeling the pseudo-document as yet another conditioning event — on par with the preceding $N-1$ words — and finding a single probability distribution conditioned on the entire "history." Note that the co-occurrence of the predicted word with, say, the immediately preceding word in the history is a discrete event and amenable to simple counting. By contrast, the pseudo-document is a continuous-valued vector and simply counting how often a word follows a particular vector in a training corpus is meaningless; we must employ a parametric model for word-history co-occurrence, possibly together with discretization of the pseudo-document vector.

The remainder of the paper explores these main themes as follows. For completeness, we briefly describe in Section 2 the standard LSA language modeling techniques we implemented. We then describe the maximum entropy alternative for combining N-gram and latent semantic information in Section 3. We present experimental results on the Switchboard corpus of conversational speech in Section 4 and conclude in Section 5.

2 LSA-Based Language Models

LSA requires a corpus separated into semantically coherent documents, and a vocabulary to cover words found in these documents. It is assumed that the co-occurrence of any two words within a document at a rate much greater than chance is an indication of their semantic similarity. This similarity is then used for language modeling, as explained below. The notation and exposition in this section closely follows that of Bellegarda (2000).

2.1 Word-Document Frequency Matrix W

The first step in LSA is to represent co-occurrence information by a large spare matrix. Let $\mathcal{V}$, $|\mathcal{V}| = M$, be the underlying task vocabulary, and $\mathcal{T}$ a text corpus, with document boundaries marked, comprising N documents relevant to some domain of interest. Typically, M and N are of the order of 10^4 and 10^5, respectively. $\mathcal{T}$, the language model training corpus, may thus have hundreds of millions of words. Unlike N-gram models, the construction of the $M \times N$ matrix W of co-occurrences between words and documents ignores word order within the document; it is accumulated from $\mathcal{T}$ by simply counting how many times a word appears in a document.

In constructing the word-document co-occurrence matrix W, the raw count c_{ij} of a word $w_i \in \mathcal{V}$ in a document $d_j \in \mathcal{T}$ is weighted by

- the "relevance" of a word in the vocabulary to the topic of a document, function words being given less weight than content words, and
- the size of the document, a word with a given count in a longer document being given less weight than in a shorter one.

To accomplish the former, pretend that a unique (unknown) document in our collection $\mathcal{T}$ is relevant for some task and our goal is to guess which one it is. Let the *a priori* probability of a document being relevant be uniform $\left(\frac{1}{N}\right)$ on the collection and, further, let an oracle draw a single word at random from the relevant document and reveals it to us. The conditional probability of d_j being the relevant document, given that the relevant document contains the word w_i, is clearly $\frac{c_{ij}}{c_i}$, where $c_i = \sum_{j=1}^{N} c_{ij}$. The ratio of the average *conditional entropy* of the relevant document's identity, given w_i, and its *a priori entropy* is thus a measure of the (un)informativeness of w_i. Highly informative words w_i have small values of

$$\epsilon_i = \epsilon_{w_i} = -\frac{1}{\log N} \sum_{j=1}^{N} \frac{c_{ij}}{c_i} \log \frac{c_{ij}}{c_i} \,. \qquad (1)$$

Since $0 \leq \epsilon_i \leq 1$, the raw counts in the i-th row of W are weighted by $(1 - \epsilon_i)$.

To achieve the latter effect, the counts in the j-th column of W are weighted by the total length $c_j = \sum_{i=1}^{M} c_{ij}$ of the document d_j. In summary,

$$[W]_{ij} = (1 - \epsilon_i) \frac{c_{ij}}{c_j} \qquad (2)$$

is the resulting ij-th matrix entry.

2.2 Singular Value Decomposition of W

Each column of the matrix W represents a document and each row represents a word. Typically, W is very sparse. To obtain a compact representation, singular value decomposition (SVD) is employed (cf. Berry et al (1993)) to yield

$$W \approx \hat{W} = U \times S \times V^T , \tag{3}$$

where, for some order $R \ll \min(M, N)$ of the decomposition, U is a $M \times R$ left singular matrix with rows u_i, $i = 1, \ldots, M$, S is a $R \times R$ diagonal matrix of singular values $s_1 \geq s_2 \geq \ldots \geq s_R \gg 0$, and V is a $N \times R$ right singular matrix with rows v_j, $j = 1, \ldots, N$. For each i, the scaled R-vector $u_i S$ may be viewed as representing w_i, the i-th word in the vocabulary, and similarly the scaled R-vector $v_j S$ as representing d_j, the j-th document in the corpus. Note that the $u_i S$'s and $v_j S$'s both belong to $\mathbb{R}^R$, the so called LSA-space.

The following similarity measure between the i-th and i'-th words w_i and $w_{i'}$ is frequently used:

$$\begin{aligned} K(w_i, w_{i'}) &= \frac{u_i S \cdot u_{i'} S}{\|u_i S\| \times \|u_{i'} S\|} \\ &= \frac{u_i S^2 u_{i'}^T}{\|u_i S\| \times \|u_{i'} S\|} . \end{aligned} \tag{4}$$

Note that $K(w_i, w_{i'})$ is nothing but the cosine of the angle between the vectors $u_i S$ and $u_{i'} S$. Algorithms such as K-means clustering have been applied to the vocabulary using (4) as a measure of similarity.

Replacing the u_i's with v_j's in the definition above, a corresponding measure $K(d_j, d_{j'})$

$$\begin{aligned} K(d_j, d_{j'}) &= \frac{v_j S^2 v_{j'}^T}{\|v_j S\| \times \|v_{j'} S\|} . \end{aligned} \tag{5}$$

of similarity between the j-th and j'-th documents is obtained and has been used for document clustering, filtering and topic detection.

2.3 Calculating Word-Probabilities via LSA

Given a sequence $w_1, w_2, \ldots, w_T$ of words in a sentence, the semantic coherence between w_t, the word in the t-th position, and $\tilde{d}_{t-1} \equiv \{w_1, \ldots, w_{t-1}\}$, all its predecessors, is used to construct a conditional probability on the vocabulary. Specifically, for a word w_i in a training document d_j, it is true by virtue of (3) that $[W]_{ij} \approx u_i S v_j^T$. However, since the word-document similarity function

$$K(w_i, d_j) = \frac{u_i S v_j^T}{\|u_i S^{\frac{1}{2}}\| \times \|v_j S^{\frac{1}{2}}\|} , \tag{6}$$

by itself is not a *bona fide* probability mass function, a $M \times 1$ pseudo-document vector $\tilde{d}_{t-1}$ is constructed by weighting the frequency of the preceding words in accordance with (2), and its scaled R-vector representation $\tilde{v}_{t-1} S = \tilde{d}_{t-1}^T U$ is used in (6) to obtain

$$\begin{aligned} & P_{\text{LSA}}(w_t | \tilde{d}_{t-1}) \\ &= \frac{\left[K(w_t, \tilde{d}_{t-1}) - K_{\min}(\tilde{d}_{t-1})\right]^\gamma}{\sum_w \left[K(w, \tilde{d}_{t-1}) - K_{\min}(\tilde{d}_{t-1})\right]^\gamma} , \end{aligned} \tag{7}$$

where $K_{\min}(\tilde{d}) = \min_w K(w, \tilde{d})$ is an offset to make the resulting probabilities nonnegative. The coefficient $\gamma \gg 1$, as noted by Coccaro and Jurafsky (1998), is chosen experimentally to increase the otherwise small dynamic range of K as w varies over the vocabulary.

As one processes successive words in a sentence, the pseudo-document $\tilde{d}_{t-1}$ is updated incrementally:

$$\tilde{d}_t = \frac{t-1}{t} \tilde{d}_{t-1} + \frac{1 - \epsilon_{w_t}}{t} \mathbf{e}_{w_t} , \tag{8}$$

where $\mathbf{e}_{w_t}$ is a $M \times 1$ vector with a 1 in the position corresponding to w_t and 0 elsewhere. Consequently, the vector $\tilde{v}_{t-1} S$ needed for the similarity computation of (6) towards the probability calculation of (7) is also incrementally updated:

$$\tilde{v}_t S = \lambda \frac{t-1}{t} (\tilde{v}_{t-1} S) + \frac{1 - \epsilon_{w_t}}{t} u_{w_t} , \tag{9}$$

where a positive "decay" coefficient $\lambda < 1$ is thrown in to accommodate dynamic shifts in topic.

2.4 Combining P_{LSA} with N-grams

Several strategies have been proposed (Coccaro and Jurafsky, 1998; Bellegarda, 2000) for combining the LSA-based probability (7) with standard N-gram probabilities, and we list those which we have investigated for conversational speech.

Linear Interpolation: For some experimentally determined constants α, and $\overline{\alpha} = 1 - \alpha$,

$$\begin{aligned} & P(w_t | w_{t-1}, w_{t-2}, \tilde{d}_{t-1}) = \\ & \alpha P_{\text{LSA}}(w_t | \tilde{d}_{t-1}) + \overline{\alpha} P_{N\text{-gram}}(w_t | w_{t-1}, w_{t-2}). \end{aligned} \tag{10}$$

Similarity Modulated N-gram: With the similarity (6) offset to be nonnegative, as done in (7),

$$\begin{aligned} & P(w_t | w_{t-1}, w_{t-2}, \tilde{d}_{t-1}) = \\ & \frac{K(w_t, \tilde{d}_{t-1}) P_{N\text{-gram}}(w_t | w_{t-1}, w_{t-2})}{\sum_w K(w, \tilde{d}_{t-1}) P_{N\text{-gram}}(w | w_{t-1}, w_{t-2})}. \end{aligned} \tag{11}$$

Information Weighted Arithmetic Mean: Setting $\lambda_w = \frac{1-\epsilon_w}{2}$ to account for the informativeness of a word w about its document, cf (1), and $\overline{\lambda}_w = 1 - \lambda_w$,

$$P(w_t|w_{t-1}, w_{t-2}, \tilde{d}_{t-1}) = \tag{12}$$

$$\frac{\lambda_{w_t} P_{\text{LSA}}(w_t|\tilde{d}_{t-1}) + \overline{\lambda}_{w_t} P_{N\text{-gram}}(w_t|w_{t-1}, w_{t-2})}{\sum_w \lambda_w P_{\text{LSA}}(w|\tilde{d}_{t-1}) + \overline{\lambda}_w P_{N\text{-gram}}(w|w_{t-1}, w_{t-2})},$$

Information Weighted Geometric Mean: With the same λ_w and $\overline{\lambda}_w$ as above,

$$P(w_t|w_{t-1}, w_{t-2}, \tilde{d}_{t-1}) = \tag{13}$$

$$\frac{P_{\text{LSA}}^{\lambda_{w_t}}(w_t|\tilde{d}_{t-1}) \cdot P_{N\text{-gram}}^{\overline{\lambda}_{w_t}}(w_t|w_{t-1}, w_{t-2})}{\sum_w P_{\text{LSA}}^{\lambda_w}(w|\tilde{d}_{t-1}) \cdot P_{N\text{-gram}}^{\overline{\lambda}_w}(w|w_{t-1}, w_{t-2})}.$$

We compute language model perplexities for the Switchboard corpus using each of these methods and discuss the results in Section 4.1.

3 Exponential Models with Latent Semantic Features

The ad hoc construction of $P_{\text{LSA}}(w|\tilde{d}_{t-1})$ to somehow capture $K(w, \tilde{d}_{t-1})$, and its combination with N-gram statistics described above are a somewhat unsatisfactory aspect of the LSA-based models. We propose, following Khudanpur (2000), an alternative family of exponential models

$$P_{\underline{\alpha}}(w_t|\tilde{d}_{t-1}, w_{t-2}, w_{t-1}) \tag{14}$$

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

where $f_1(w_t)$, $f_2(w_{t-1}, w_t)$ and $f_3(w_{t-2}, w_{t-1}, w_t)$ are usually, but not necessarily, $\{0,1\}$-valued indicator functions of N-gram features and α_{w_t}, α_{w_{t-1},w_t} and $\alpha_{w_{t-2},w_{t-1},w_t}$ are their corresponding feature weights, and where the semantic coherence between a word w_t and its long-span history $\tilde{d}_{t-1}$ has been thrown in as a feature, on par with the standard N-gram features. *E.g.*, one could have

$$f_{\text{LSA}}(\tilde{d}_{t-1}, w_t) = K(w_t, \tilde{d}_{t-1}). \tag{15}$$

We then find the maximum likelihood estimate of the model parameters $\underline{\alpha}$ given the training data. Recall that the resulting model is also the *maximum entropy* (ME) model among models which satisfy constraints on the marginal probabilities or expected values of these features (Rosenfeld, 1996).

An important decision that needs to be made in a model such as (14) is the parameterization $\underline{\alpha}$. In a traditional ME language model, in the absence of LSA-based features, each N-gram feature function is a $\{0,1\}$-valued indicator function, and there is a parameter associated with each feature: an α_w for each unigram constraint, an $\alpha_{w',w}$ for each bigram constraint, *etc.* In extending this methodology to the LSA features, we note that $K(w_t, \tilde{d}_{t-1})$ is continuous-valued. That in itself is not a problem; the ME framework does not require the $f(\cdot)$'s to be binary. What is problematic, however, is the fact that, almost surely, no two pseudo-documents $\tilde{d}_t$ and $\tilde{d}_{t'}$ will ever be identical. Therefore, assigning a distinct parameter $\alpha_{\tilde{d},w}$ for each pseudo-document – word pair $(\tilde{d}, w)$ is counterproductive, and some tying of parameters for similarly valued $\tilde{d}$ is necessary.

If we tie all the LSA parameters together, i.e., set

$$\alpha_{\tilde{d},w} = \alpha_{\text{LSA}} \qquad \forall\, w \in \mathcal{V} \text{ and } \tilde{d} \in \mathbb{R}^R, \tag{16}$$

then (14) becomes directly comparable to the similarity modulated N-gram model (11), except that the choice of α_{LSA} here is made *jointly* with the N-gram α's to maximize training data likelihood. If we let each vocabulary item to have its own α, i.e.

$$\alpha_{\tilde{d},w} = \alpha_{\text{LSA},w} \qquad \forall\, \tilde{d} \in \mathbb{R}^R, \tag{17}$$

then (14) becomes directly comparable to the geometric interpolation method (13), again except that unlike λ_w, the $\alpha_{\text{LSA},w}$ parameters are determined *jointly* with the N-gram α's to maximize a likelihood criterion.

Since the goal of parameter tying, however, is to deal with the continuous nature of the pseudo-document $\tilde{d}$, another alternative, as suggested by Khudanpur (2000), is

$$\alpha_{\tilde{d},w} = \alpha_{\hat{d},w} \qquad \forall\, \tilde{d} \in \Phi(\hat{d}) \subset \mathbb{R}^R, \tag{18}$$

where $\Phi(\hat{d})$ represents a finite partition of $\mathbb{R}^R$ indexed by $\hat{d}$. We choose to pursue this alternative.

We use a standard K-means clustering of the representations $v_j S$ of the training documents d_j, with (5) in the role of distance, to obtain a modest number of clusters. We then pool documents in each cluster together to form *topic-centroids* $\hat{d}$, and the partition $\Phi(\cdot)$ of $\mathbb{R}^R$ is defined by the *Voronoi regions* around the topic-centroids:

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

We also make two approximations to the feature function of (15). First, we *approximate* the pseudo-document $\tilde{d}_{t-1}$ in $K(\cdot)$ with its nearest topic-centroid

$\hat{d}_{t-1} = \hat{d}$ whenever $\tilde{d}_{t-1} \in \Phi(\hat{d})$. This is motivated by the fact that we often deal with very small pseudo-documents $\tilde{d}_{t-1}$ in speech recognition, and $\hat{d}$ provides a more robust estimate of semantic coherence with w_t than $\tilde{d}_{t-1}$. Furthermore, keeping in mind the small dynamic range of the similarity measure of (6), as well as the interpretation (1) of ϵ_w, we *approximate* the feature function of (15) with

$$\hat{f}_{\rm LSA}(\tilde{d}_{t-1}, w_t) = \begin{cases} 1 & \text{if } K(w_t, \hat{d}_{t-1}) > \eta \\ & \text{and } \epsilon_w < \tau, \\ 0 & \text{otherwise.} \end{cases} \quad (19)$$

This pragmatic approximation results in a simplified implementation, particularly for the computation of feature-expectations during parameter estimation. More importantly, when there is a free parameter α for each $(\hat{d}, w)$ pair, as is the case in (18), $\hat{f}_{\rm LSA}(\hat{d}, w) = 1$ and $\hat{f}_{\rm LSA}(\hat{d}, w) = K(w, \hat{d})$ yield equivalent model families. Therefore, using

$$\alpha^{1 \text{ or } 0}_{\hat{d}_{t-1}, w_t} \quad \text{instead of} \quad \alpha^{K(w_t, \hat{d}_{t-1})}_{\hat{d}_{t-1}, w_t} \quad (20)$$

in (14) simply amounts to doing feature selection.

For all pairs $(\hat{d}, w)$ with $\hat{f}_{\rm LSA}(\hat{d}, w) = 1$ in (19), the model-expectation of $\hat{f}$ is constrained to be the relative frequency of w within the cluster of training documents whose centroid is $\hat{d}$. By virtue of their semantic coherence, it is usually higher than the relative frequency of w in the entire corpus.

Another interesting way of tying the LSA parameters, which we have *not* investigated here, is

$$\alpha_{\tilde{d}, w} = \alpha_{\hat{d}, \hat{w}} \quad \forall\, w \in \Psi(\hat{w})\,,\ \forall\, \tilde{d} \in \Phi(\hat{d})\,, \quad (21)$$

where $\Psi(\hat{w})$ is a finite, possibly $\hat{d}$-dependent, partition of the vocabulary. This parameterization may be particularly beneficial when, due to a very large vocabulary or a small training corpus, we do not have sufficient counts to constrain the model-expectations of $\hat{f}_{\rm LSA}(\hat{d}, w)$ for all words w bearing high semantic similarity with a topic-centroid $\hat{d}$. An automatically derived or knowledge-based semantic classification of words, e.g. from WordNet, may be used as $\Psi(\cdot)$.

3.1 A Similar ME Model from the Past

An interesting consequence of (19) is that it makes the model of (14) identical in form to the model described by Khudanpur and Wu (1999). Two significant ways in which (14) is novel are that

- clustering of documents d_j to obtain topic-centroids $\hat{d}$ during training, and assignment of pseudo-documents $\tilde{d}_{t-1}$ to topic-centroids $\hat{d}_{t-1}$ during recognition, is based on similarity in LSA-space $\mathbb{R}^R$, not document-space $\mathbb{R}^M$, and
- the set of words with active semantic features (19) for any particular topic-centroid $\hat{d}$ is determined by a threshold η on LSA similarity, not by a difference in within-topic v/s corpus-wide relative frequency.

The former results in some computational savings both during clustering and on-line topic assignment. The latter may result in a different choice of topic-dependent features. We present a comparison of LM performance between these two ME models in Section 4.5 following our main results.

4 Switchboard Experiments

We conducted experiments on the Switchboard corpus of conversational telephone speech (Godfrey et al, 1992), dividing the corpus into a LM training set of approximately 1500 conversations (2.2M words) and a test set of 19 conversations (20K words). The task vocabulary was fixed to 22K words, with an out-of-vocabulary rate under 0.5% on the test set. Acoustic models trained on roughly 60 hours of Switchboard speech and a bigram LM were used to generate lattices for the test utterances, and a 100-best list was generated by rescoring the lattice using a trigram model. All the results in this paper are based on rescoring this 100-best list with different language models.

We treated each conversation-side as a separate document and created W of (2) with $M \approx 22{,}000$ and $N \approx 3000$. Guided by the fact that one of 70-odd topics was prescribed to a caller when the Switchboard corpus was collected, we computed the SVD of (3) with R=73 singular values. We implemented the LSA model of (7) with $\gamma = 20$, and the four LSA + N-gram combinations of Section 2.4.

To obtain the document clusters and topic-centroids $\hat{d}$ required for creating the partition $\Phi(\cdot)$ of (18), we randomly assigned the training documents to one of 50 clusters and used a K-means algorithm to iteratively

(i) compute the topic-centroid $\hat{d}$ of each cluster by pooling together all the documents in the cluster, and

(ii) reassigning each document d_j to a cluster to whose centroid the document bore the greatest LSA similarity $K(d_j, \hat{d})$.

Each cluster was required to have a minimum number of 10 documents in it, and if the number of documents in a cluster fell below this threshold following step (ii), then the cluster was eliminated and each of its documents reassigned to the nearest of the remaining centroids. The iteration stopped when no

reassignments resulted in step (ii). This procedure resulted in 25 surviving centroids, and we checked by a cursory examination of documents that the clusters were reasonably coherent.

For each topic-centroid $\hat{d}$, we chose, according to (19), a set of words that activate an LSA feature. We used $\tau = 0.4$ to first eliminate stop-words and then set a $\hat{d}$-specific η to yield $\sim$800 vocabulary-words above threshold per $\hat{d}$. However, not all these words actually appeared in training documents in $\Phi(\hat{d})$. Only the seen words were chosen, obtaining an average of 750 topic-dependent features for each topic-centroid. The resulting model had 19K $\alpha_{\hat{d},w}$ parameters associated with the semantic features in addition to about 22K unigram α_w's, 300K bigram $\alpha_{w',w}$'s and 170K trigram $\alpha_{w'',w',w}$'s. A ME language model was trained with these parameters using the toolkit developed by Wu (2002), and readers interested in the computational issues pertaining to maximum entropy model estimation are referred to his doctoral dissertation for details.

4.1 Perplexity: LSA + N-gram Models

We used the CMU-CU LM toolkit to implement a baseline trigram model with Good-Turing discounting and Katz back-off. We then measured the perplexity of the reference transcription of the test conversations for the trigram and the four LSA + N-gram models of Section 2.4. The pseudo-document $\tilde{d}_{t-1}$ was updated according to (9) with $\lambda = 0.97$ for all four models. We used $\alpha = 0.1$ for the linear interpolation of the LSA and N-gram models. The other three combination techniques require no additional parameters.

Language Model	Perplexity
CMU-CU Standard Trigram	81.1
LSA+Trigram Linear Interpolation	81.8
Similarity Modulated Trigram	79.1
Info Weighted Arithmetic Mean	81.8
Info Weighted Geometric Mean	75.8

Table 1: Perplexities: N-gram + LSA Combination

The trends in the performance of the four schemes, reported in Table 1, are consistent with those reported by Coccaro and Jurafsky (1998), with the information weighted geometric interpolation showing the greatest reduction in perplexity. However, the reduction in perplexity is much smaller on this corpus than, *e.g.*, the 19% reduction reported by Bellegarda (2000) using similar models on a corpus of newspaper text.

4.2 Effect of Replacing $\tilde{d}_{t-1}$ with $\hat{d}_{t-1}$

We next describe our attempt to gain some understanding of the effect of replacing the pseudo-document $\tilde{d}_{t-1}$ with the closest topic-centroid $\hat{d}_{t-1}$ before the similarity computation in (19). For several of our test conversation-sides, we computed $K(w_t, \tilde{d}_{t-1})$ and $K(w_t, \hat{d}_{t-1})$, $t = 1, \ldots, T$, where w_t denotes the word in the t-th position and T denotes the number of words in the conversation-side. For a *typical* conversation side in our test set, these

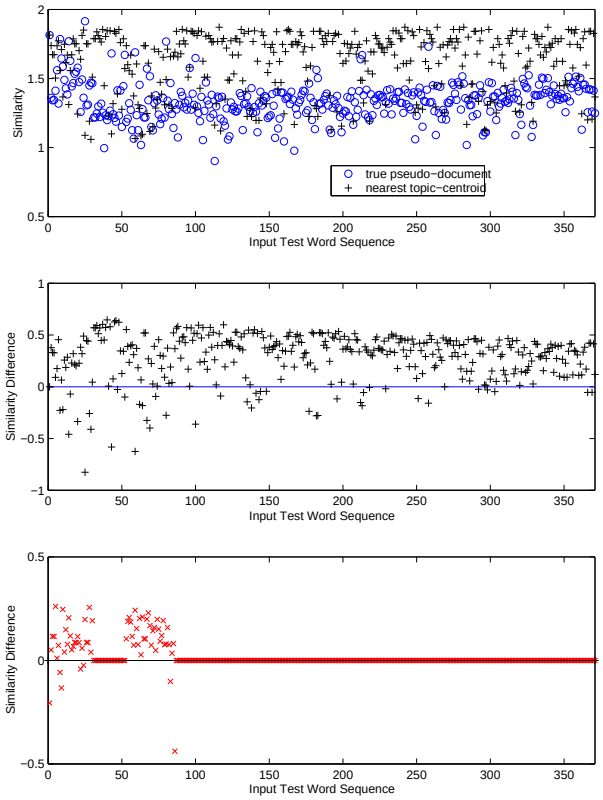

Figure 1: $K(w_t, \hat{d}_{t-1})$ and $K(w_t, \tilde{d}_{t-1})$ through a conversation (TOP), $K(w_t, \hat{d}_{t-1}) - K(w_t, \tilde{d}_{t-1})$ (MIDDLE), and $K(w_t, \hat{d}_T) - K(w_t, \hat{d}_{t-1})$ (BOTTOM).

similarities are plotted as a function of t in the box at the top of Figure 1. The second box shows the difference $K(w_t, \hat{d}_{t-1}) - K(w_t, \tilde{d}_{t-1})$. It is clear from the second box that $\hat{d}_{t-1}$ bears a greater similarity to the next word than $\tilde{d}_{t-1}$, confirming the beneficial effect of replacing $\tilde{d}_{t-1}$ with $\hat{d}_{t-1}$. We also computed $K(w_t, \hat{d}_T)$, the similarity of w_t with the topic-centroid most similar to the entire conversation side, and the box at the bottom of Figure 1 depicts the difference $K(w_t, \hat{d}_T) - K(w_t, \hat{d}_{t-1})$. We note with some satisfaction that as the conversation proceeds, the dynamically computed topic-centroid $\hat{d}_{t-1}$ converges to $\hat{d}_T$. Our conversation-sides are 470 words long

on average, and we observe convergence roughly 110 words into the conversation side.

4.3 Perplexity: ME Model with LSA Features

In the process of comparing our ME model of (14) with the one described by Khudanpur and Wu (1999), we noticed that they built a baseline trigram model using the SRI LM toolkit. Other than this, our experimental setup – training and test set definitions, vocabulary, etc. – matches theirs exactly. We report the perplexity of our ME model against their baseline in Table 2, where the figures in the first two lines are quoted directly from Khudanpur and Wu (1999). A single topic-centroid $\hat{d}_T$ selected

Language Model	Perplexity
SRI Trigram	78.8
ME Trigram	78.9
ME + LSA Features (Closest $\hat{d}_T$)	73.6
ME + LSA Features (Oracle $\hat{d}_T$)	73.0

Table 2: Perplexities: Maximum Entropy Models

for an entire test conversation-side was used in these experiments. The last line of Table 2 shows the best perplexity obtainable by any topic-centroid, suggesting that the automatically chosen, Voronoi region based topic-centroids are quite adequate.

A comparison of Tables 1 and 2 also shows that the maximum entropy model is more effective in capturing semantic information than the information weighted geometric mean of the LSA-based unigram model and the trigram model. The correspondence of information weighted geometric mean with the parameterization of (17) and the corresponding richer parameterization of (18) are perhaps adequate to explain this improvement.

4.4 Word Error Rates for the ME Model

We rescored the 100-best hypotheses generated by the baseline trigram model using the ME model with LSA features. In order to assign a topic-centroid $\hat{d}$ to a test utterance in the absence of its correct transcription, we investigated using a concatenation of the 1-best, 10-best or 100-best first-pass hypotheses of utterances in the test set, computed $\hat{d}$ once per test utterance, and found the performance of the 10-best hypotheses to yield a slightly lower word error rate (WER). This is perhaps the optimal trade-off between robustness in topic assignment resulting from considering additional word hypotheses, and noise introduced by considering erroneous words. We also

Language Model ($\hat{d}_T$ Assignment)	WER
SRI Trigram	38.47%
ME Trigram	38.32%
ME+LSA (per utterance via 10-best)	37.94%
ME+LSA (per conv-side via 10-best)	37.86%

Table 3: Error Rates: Maximum Entropy Models

investigated assigning topic for the entire conversation side based on the first-pass output and found it to yield a further reduction in WER. We report the results in Table 3 where the top two lines are, again, quoted directly from Khudanpur and Wu (1999).

We performed the standard NIST MAPSSWE statistical significance test (Pallett et al, 1990) and found that

- the WER improvement of the ME trigram model over the baseline SRI trigram model is not significant (p=0.529),
- that of the ME model with LSA features and utterance-level topic assignment over the ME trigram model is significant (p=0.008), and
- that of the ME model with LSA features and conversation-level topic assignment over the ME trigram model is also significant (p=0.002).

The difference between the WER obtained by utterance-level v/s conversation-level topic assignment is not significant (p=0.395); nor are other WER differences (not reported here) between using the 1- v/s 10- v/s 100-best hypotheses for topic assignment.

4.5 Benefits of Dimensionality Reduction

It was pointed out in Section 3.1 that the model proposed here differs from the model of Khudanpur and Wu (1999) mainly in the use of the R-dimensional LSA-space for similarity comparison rather than direct comparison in M-dimensional document-space. We present in Table 4 a summary comparison of the two modeling techniques. While, due to the sparse nature of the vectors, the 22K-dimensional space does not entail a proportional growth in similarity computation relative to the 73-dimensional space, the LSA similarities are still expected to be faster to compute. Furthermore, the LSA based model yields comparable perplexity and WER performance with considerably fewer topic-centroids, resulting in fewer comparisons during run time for determining the nearest centroid. Of lesser note is the observation that the η-threshold based topic-feature selection of (19) results in a content word being an active feature for fewer topics than it does when topic-features are selected based on differences in within-topic and overall relative frequencies.

Attribute	Model A	Model B
Similarity measure	cosine	
Document clustering	K-means	
Vector-space dimension	22K	73
Num. topic-centroids	67	25
Avg. # topics/topic-word	1.8	1.3
Total # topic-parameters	15500	19000
ME + topic perplexity	73.5	73.6
ME + topic WER	37.9%	

Table 4: A comparison between the model (A) of Khudanpur and Wu (1999) and our model (B).

5 Summary and Conclusion

We have presented a framework for incorporating latent semantic information together with standard N-gram statistics in a unified exponential model for statistical language modeling. This framework permits varying degrees of parameter tying depending on the amount of training data available. We have drawn parallels between some conventional ways of combining LSA-based models with N-grams and the parameter-tying decisions in our exponential models, and our results suggest that incorporating semantic information using maximum entropy principles is more effective than the ad hoc techniques.

We have presented perplexity and speech recognition accuracy results on the Switchboard corpus which suggest that LSA-based features, while not as effective on conversational speech as on newspaper text, produce modest but statistically significant improvements in speech recognition performance.

Finally, we have shown that the maximum entropy model presented here performs as well as a previously proposed maximum entropy model for incorporating topic-dependencies, but it is computationally more economical.

6 Acknowledgments

We would like to thank Jun Wu of Google Inc. for assistance in the use of his tools for maximum entropy model estimation and application, and Woosung Kim of Johns Hopkins University for assistance in the use of other software. We also thank the anonymous referees for comments that helped improve this manuscript. This research was partially supported by the National Science Foundation via MLIAM Grant No IIS 9982329.

References

Bellegarda, Jerome. 2000. Exploiting latent semantic information in statistical language modeling. *Proc. IEEE*, 88:1279-1296.

Berry, Michael et al. 1993. SVDPACKC (version 1.0) user's guide. *Tech. Report CS-93-194*, University of Tennessee, Knoxville, TN.

Chen, Stanley and Roni Rosenfeld. 1998. Topic adaptation for language modeling using unnormalized exponential models. *Proc. ICASSP*, pages 681-684, Seattle, WA.

Clarkson, Philip and Anthony Robinson 1997. Language model adaptation using mixtures and an exponentially decaying cache. *Proc. ICASSP*, pages 799-802, Munich, Germany.

Coccaro, Noah and Daniel Jurafsky. 1998. Towards better integration of semantic predictors in statistical language modeling. *Proc. ICSLP*, pages 2403-2406, Sydney, Australia.

Godfrey, John et al. 1992. Switchboard: telephone speech corpus for research and development. *Proc. ICASSP*, pages 517-520, San Francisco, CA.

Gotoh, Yoshihiko and Steve Renals. 1997. Document space models using latent semantic analysis. *Proc. of Eurospeech*, pages 1443-1446, Patras, Greece.

Iyer, Rukmini and Mari Ostendorf. 1999. Modeling Long Distance Dependence in Language: Topic Mixtures vs. Dynamic Cache Models. *IEEE Trans Speech and Audio Processing*, 7:30-39.

Khudanpur, Sanjeev and Jun Wu. 1999. A maximum entropy language model to integrate n-grams and topic dependencies for conversational speech recognition. *Proc. ICASSP*, pages 553-556, Phoenix, USA.

Khudanpur, Sanjeev. 2000. Putting language back into language modeling. Presented at the *DARPA-Lucent Workshop on Spoken Language Recognition and Understanding*, Summit, NJ, Feb 6-9.

Pallett, David et al. 1990. Tools for the analysis of benchmark speech recognition tests. *Proc. ICASSP*, 1:97-100, Alburquerque, NM.

Rosenfeld, Roni. 1996. A maximum entropy approach to adaptive statistical language modeling. *Computer Speech and Language*, 10:187-228.

Wu, Jun. 2002. Maximum entropy language modeling with nonlocal dependencies. *PhD Dissertation*, Johns Hopkins University CS Department, Baltimore, MD.