Computational Linguistics and Chinese Language Processing
Vol. 11, No. 1, March 2006, pp. 37-56

A Maximum Entropy Approach for Semantic Language Modeling

Chuang-Hua Chueh[*], Hsin-Min Wang[+] and Jen-Tzung Chien[*]

Abstract

The conventional *n*-gram language model exploits only the immediate context of historical words without exploring long-distance semantic information. In this paper, we present a new information source extracted from latent semantic analysis (LSA) and adopt the maximum entropy (ME) principle to integrate it into an *n*-gram language model. With the ME approach, each information source serves as a set of constraints, which should be satisfied to estimate a hybrid statistical language model with maximum randomness. For comparative study, we also carry out knowledge integration via linear interpolation (LI). In the experiments on the TDT2 Chinese corpus, we find that the ME language model that combines the features of trigram and semantic information achieves a 17.9% perplexity reduction compared to the conventional trigram language model, and it outperforms the LI language model. Furthermore, in evaluation on a Mandarin speech recognition task, the ME and LI language models reduce the character error rate by 16.9% and 8.5%, respectively, over the bigram language model.

Keywords: Language Modeling, Latent Semantic Analysis, Maximum Entropy, Speech Recognition

1. Introduction

Language modeling plays an important role in automatic speech recognition (ASR). Given a speech signal O, the most likely word sequence $\hat{W}$ is obtained by maximizing *a posteriori* probability $p(W|O)$, or, equivalently, the product of acoustic likelihood $p(O|W)$ and prior probability of word sequence $p(W)$:

$$\hat{W} = \arg\max_{W} p(W|O) = \arg\max_{W} p(O|W)\,p(W)\,. \tag{1}$$

[*] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R. O. C
E-mail: chchueh@chien.csie.ncku.edu.tw

[+] Institute of Information Science, Academia Sinica, Taipei, Taiwan, R. O. C

[Received June 16, 2005; Revised January 2, 2006; Accepted January 3, 2006]

This prior probability corresponds to the language model that is useful in characterizing regularities in natural language. Also, this language model has been widely employed in optical character recognition, machine translation, document classification, information retrieval [Ponte and Croft 1998], and many other applications. In the literature, there were several approaches have been taken to extract different linguistic regularities in natural language. The structural language model [Chelba and Jelinek 2000] extracted the relevant syntactic regularities based on predefined grammar rules. Also, the large-span language model [Bellegarda 2000] was feasible for exploring the document-level semantic regularities. Nevertheless, the conventional *n*-gram model was effective at capturing local lexical regularities. In this paper, we focus on developing a novel latent semantic *n*-gram language model for continuous Mandarin speech recognition.

When considering an *n*-gram model, the probability of a word sequence W is written as a product of probabilities of individual words conditioned on their preceding *n*-1 words

$$p(W) = p(w_1, w_2, \cdots, w_T) \cong \prod_{i=1}^{T} p(w_i | w_{i-n+1}, ..., w_{i-1}) = \prod_{i=1}^{T} p(w_i | w_{i-n+1}^{i-1}), \quad (2)$$

where w_{i-n+1}^{i-1} represents historical words for word w_i, and the *n*-gram parameter $p(w_i | w_{i-n+1}^{i-1})$ is usually obtained via the maximum likelihood estimation:

$$p(w_i | w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i})}{c(w_{i-n+1}^{i-1})}. \quad (3)$$

Here, $c(w_{i-n+1}^{i})$ is the number of occurrences of word sequence w_{i-n+1}^{i} in the training data. Since the *n*-gram language model is limited by the span of window size *n*, it is difficult to characterize long-distance semantic information in *n*-gram probabilities. To deal with the issue of insufficient long-distance word dependencies, several methods have been developed by incorporating semantic or syntactic regularities in order to achieve long-distance language modeling.

One simple combination approach is performed using the *linear interpolation* of different information sources. With this approach, each information source is characterized by a separate model. Various information sources are combined using weighted averaging, which minimizes overall perplexity without considering the strengths and weaknesses of the sources in particular contexts. In other words, the weights were optimized globally instead of locally. The hybrid model obtained in this way cannot guarantee the optimal use of different information sources [Rosenfeld 1996]. Another important approach is based on Jaynes' maximum entropy (ME) principle [Jaynes 1957]. This approach includes a procedure for setting up probability distributions on the basis of partial knowledge. Different from linear interpolation, this approach determines probability models with the largest randomness and simultaneously captures all information provided by various knowledge sources. The ME framework was first applied to language modeling in [Della Pietra *et al.* 1992]. In the following, we survey several language model algorithms where the idea of information combination is adopted.

In [Kuhn and de Mori 1992], the cache language model was proposed to merge domain information by boosting the probabilities of words in the previously-observed history. In [Zhou and Lua 1999], *n*-gram models were integrated with the mutual information (MI) of trigger words. The MI-Trigram model achieved a significant reduction in perplexity. In [Rosenfeld 1996], the information source provided by trigger pairs was incorporated into an *n*-gram model under the ME framework. Long-distance information was successfully applied in language modeling. This new model achieved a 27% reduction in perplexity and a 10% reduction in the word error rate. Although trigger pairs are feasible for characterizing long-distance word associations, this approach only considers the frequently co-occurring word pairs in the training data. Some important semantic information with low frequency of occurrence is lost. To compensate for this weakness, the information of entire historical contexts should be discovered. Since the words used in different topics are inherently different in probability distribution, topic-dependent language models have been developed accordingly. In [Clarkson and Robinson 1997], the topic language model was built based on a mixture model framework, where topic labels were assigned. Wu and Khudanpur [2002] proposed an ME model by integrating *n*-gram, syntactic and topic information. Topic information was extracted from unsupervised clustering in the original document space. A word error rate reduction of 3.3% was obtained using the combined language model. In [Florian and Yarowsky 1999], a delicate tree framework was developed to represent the topic structure in text articles. Different levels of information were integrated by performing linear interpolation hierarchically. In this paper, we propose a new semantic information source using latent semantic analysis (LSA) [Deerwester *et al.* 1990; Berry *et al.* 1995], which is used for reducing the disambiguity caused by polysemy and synonymy [Deerwester *et al.* 1990]. Also, the relations of semantic topics and target words are incorporated with *n*-gram models under the ME framework. We illustrate the performance of the new ME model by investigating perplexity in language modeling and the character-error rate in continuous Mandarin speech recognition. The paper is organized as follows. In the next section, we introduce an overview of the ME principle and its relations to other methods. In Section 3, the integration of semantic information and n-gram model via linear interpolation and maximum entropy is presented. Section 4 describes the experimental results. The evaluation of perplexity and character-error rate versus different factors is conducted. The final conclusions drawn from this study are discussed in Section 5.

2. Maximum Entropy Principle

2.1 ME Language Modeling

The underlying idea of the ME principle [Jaynes 1957] is to subtly model what we know, and assume nothing about what we do not know. Accordingly, we choose a model that satisfies all the information we have and that makes the model distribution as uniform as possible. Using the ME model, we can combine different knowledge sources for language modeling [Berger *et al*. 1996]. Each knowledge source provides a set of constraints, which must be satisfied to find a unique ME solution. These constraints are typically expressed as marginal distributions. Given features $f_1,\cdots,f_N$, which specify the properties extracted from observed data, the expectation of f_i with respect to empirical distribution $\tilde{p}(h,w)$ of history *h* and word *w* is calculated by

$$\tilde{p}(f_i)=\sum_{h,w}\tilde{p}(h,w)f_i(h,w)\,, \tag{4}$$

where $f_i(\cdot)$ is a binary-valued feature function. Also, using conditional probabilities in language modeling, we yield the expectation with respect to the target conditional distribution $p(w|h)$ by

$$p(f_i)=\sum_{h,w}\tilde{p}(h)p(w|h)f_i(h,w)\,. \tag{5}$$

Because the target distribution is required to contain all the information provided by these features, we specify these constraints

$$p(f_i)=\tilde{p}(f_i),\quad \text{for } i=1,\cdots,N\ . \tag{6}$$

Under these constraints, we maximize the conditional entropy or uniformity of distribution $p(w|h)$. Lagrange optimization is adopted to solve this constrained optimization problem. For each feature f_i, we introduce a Lagrange multiplier λ_i. The Lagrangian function $\Lambda(p,\lambda)$ is extended by

$$\Lambda(p,\lambda)=H(p)+\sum_{i=1}^{N}\lambda_i\left[p(f_i)-\tilde{p}(f_i)\right], \tag{7}$$

with conditional entropy defined by

$$H(p)=-\sum_{h,w}\tilde{p}(h)p(w|h)\log p(w|h)\,. \tag{8}$$

Finally, the target distribution $p(w|h)$ is estimated as a log-linear model distribution

$$p(w|h) = \frac{1}{Z_\lambda(h)} \exp\left(\sum_{i=1}^{N} \lambda_i f_i(h,w) \right), \tag{9}$$

where $Z_\lambda(h)$ is a normalization term in the form of

$$Z_\lambda(h) = \sum_{w} \exp\left(\sum_{i=1}^{N} \lambda_i f_i(h,w) \right), \tag{10}$$

determined by the constraint $\sum_w p(w|h) = 1$. The General Iterative Scaling (GIS) algorithm or Improved Iterative Scaling (IIS) algorithm [Darroch and Ratcliff 1972; Berger *et al.* 1996; Della Pietra *et al.* 1997] can be used to find the Lagrange parameters λ. The IIS algorithm is briefly described as follows.

```
Input:    Feature functions $f_1, f_2, \cdots, f_N$ and empirical distribution $\tilde{p}(h,w)$
Output:   Optimal Lagrange multiplier $\hat{\lambda}_i$
    1. Start with $\lambda_i = 0$ for all $i = 1, 2, \cdots, N$.
    2. For each $i = 1, 2, \cdots, N$ :
        a. Let $\Delta\lambda_i$ be the solution to
              $\sum_{h,w} \tilde{p}(h) p(w|h) f_i(h,w) \exp(\Delta\lambda_i F(h,w)) = \tilde{p}(f_i)$ ,
           where $F(h,w) = \sum_{i=1}^{N} f_i(h,w)$.
        b. Update the value of $\lambda_i$ according to $\lambda_i = \lambda_i + \Delta\lambda_i$.
    3. Go to step 2 if any $\lambda_i$ has not converged.
```

With the parameters $\{\hat{\lambda}_i\}$, we can calculate the ME language model by using Eqs. (9) and (10).

2.2 Relation between ML and ME Modeling

It is interesting to note the relation between maximum likelihood (ML) and ME language models. The purpose of ML estimation is to find a generative model with the maximum likelihood of training data. Generally, the log-likelihood function is adopted in the form of

$$L(p) = \log \prod_{h,w} p(w \mid h)^{\tilde{p}(h,w)} = \sum_{h,w} \tilde{p}(h,w) \log p(w \mid h) . \tag{11}$$

Under the same assumption that the target distribution $p(w|h)$ is log-linear, as shown in Eqs. (9) and (10), the log-likelihood function is extended to

$$L(p_\lambda) = \sum_{h,w} \tilde{p}(h,w) \log \frac{\exp\left(\sum_{i=1}^{N} \lambda_i f_i(h,w)\right)}{\sum_{w'} \exp\left(\sum_{i=1}^{N} \lambda_i f_i(h,w')\right)} . \tag{12}$$

By taking the derivative of the log-likelihood function with respect to λ_i and setting it at zero, we can obtain the same constraints in Eq. (6) by using the following derivations:

$$\begin{aligned}
&\sum_{h,w} \tilde{p}(h,w) f_i(h,w) - \sum_{h,w} \tilde{p}(h,w) \sum_{w''} \frac{\exp\left(\sum_{i=1}^{N} \lambda_i f_i(h,w'')\right)}{\sum_{w'} \exp\left(\sum_{i=1}^{N} \lambda_i f_i(h,w')\right)} f_i(h,w'') = 0, \\
&\Rightarrow \sum_{h,w} \tilde{p}(h,w) f_i(h,w) - \sum_{h,w} \tilde{p}(h,w) \sum_{w''} p(w''|h) f_i(h,w'') = 0, \\
&\Rightarrow \sum_{h,w} \tilde{p}(h,w) f_i(h,w) - \sum_{h} \sum_{w''} \tilde{p}(h) p(w''|h) f_i(h,w'') = 0, \\
&\Rightarrow \tilde{p}(f_i) = p(f_i).
\end{aligned} \tag{13}$$

In other words, the ME model is equivalent to an ML model with a log-linear model. In Table 1, we compare various properties using ML and ME criteria. Under the assumption of log-linear distribution, the optimal parameter λ_{ML} is estimated according to the ML criterion. The corresponding ML model $p_{\lambda_{\text{ML}}}$ is obtained through an unconstrained optimization procedure. On the other hand, ME performs the constrained optimization. The ME constraint allows us to determine the combined model $p_{\lambda_{\text{ML}}}$ with the highest entropy. Interestingly, these two estimation methods achieve the same result.

Table 1. Relation between ML and ME language models

Objective function	$L(p_\lambda)$	$H(p)$
Criterion	Maximum Likelihood	Maximum Entropy
Type of search	Unconstrained optimization	Constrained optimization
Search space	$\lambda \in$ real values	p satisfied with constraints
Solution	λ_{ML}	p_{ME}
$p_{\lambda_{\text{ML}}} = p_{\text{ME}}$		

2.3 Minimum Discrimination Information and Latent ME

The ME principle is a special case of minimum discrimination information (MDI) that has been successfully applied to language model adaptation [Federico 1999]. Let $p_b(h,w)$ be the background model trained from a large corpus of general domain, and $p_a(h,w)$ represents the adapted model estimated from an adaptation corpus of new domain. In the MDI adaptation, the language model is adapted by minimizing the distance between the background model and the adapted model. The non-symmetric Kullback-Leibler distance (KLD)

$$D(p_a(h,w), p_b(h,w)) = \sum_w p_a(h,w) \log \frac{p_a(h,w)}{p_b(h,w)} \tag{14}$$

is used for distance measuring. Obviously, when the background model is a uniform distribution, the MDI adaptation is equivalent to the ME estimation. More recently, the ME principle was extended to latent ME (LME) mixture modeling, where the latent variables representing underlying topics were merged [Wang *et al*. 2004]. To find the LME solution, the modified GIS algorithm, called expectation maximization iterative scaling (EM-IS), was used. The authors also applied the LME principle to incorporate probabilistic latent semantic analysis [Hofmann 1999] into *n*-gram modeling by serving the semantic information as the latent variables [Wang *et al.* 2003]. In this study, we use the semantic information as *explicit features* for ME language modeling. Latent semantic analysis (LSA) is adopted to build semantic topics.

3. Integration of Semantic Information and N-Gram Models

Modeling long-distance information is crucial for language modeling. In [Chien and Chen 2004; Chien *et al.* 2004], we successfully incorporated long-distance association patterns and latent semantic knowledge in language models. In [Wu and Khudanpur 2002], the integration of statistical *n*-gram and topic unigram using the ME approach was presented. Clustering of document vectors in the original document space was performed to extract topic information. However, the original document space was generally sparse and filled with noises caused by polysemy and synonymy [Deerwester *et al.* 1990]. To explore robust and representative topic characteristics, here we introduce a new knowledge source to extract long-distance semantic information for *n*-gram modeling. Our idea is to adopt the LSA approach and extract semantic topic information from the reduced LSA space. The proposed procedure of ME semantic topic modeling is illustrated in Figure 1. Because the occurrence of a word is highly related to the topic of current discourse, we apply LSA to build representative semantic topics. The subspace of semantic topics is constructed via *k*-means clustering of document vectors generated from the LSA model. Furthermore, we combine semantic topics and conventional *n*-grams under the ME framework [Chueh *et al.* 2004].

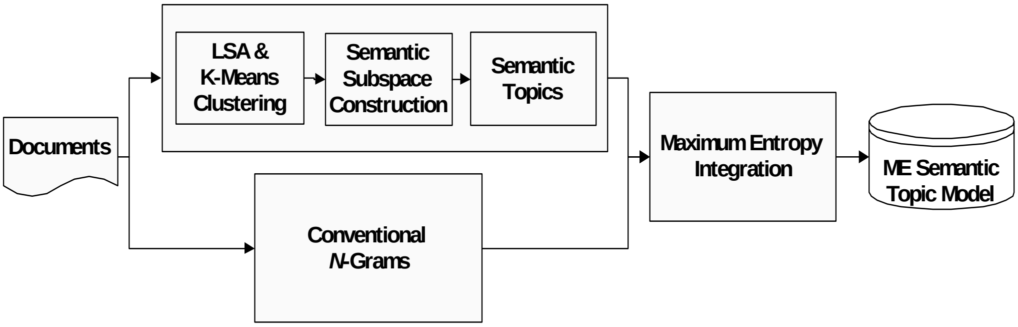

Figure 1. Implementation procedure for ME semantic topic modeling

3.1 Construction of Semantic Topics

Latent semantic analysis (LSA) is popular in the areas of information retrieval [Berry *et al.* 1995] and semantic inference [Bellegarda 2000]. Using LSA, we can extract latent structures embedded in words across documents. LSA is feasible for exploiting these structures. The first stage of LSA is to construct an $M \times D$ word-by-document matrix $\mathbf{A}$. Here, M and D represent the vocabulary size and the number of documents in the training corpus, respectively. The expression for the (i, j) entry of matrix $\mathbf{A}$ is [Bellegarda 2000]

$$a_{i,j} = (1-\varepsilon_i)\frac{c_{i,j}}{n_j}, \tag{15}$$

where $c_{i,j}$ is the number of times word w_i appears in document d_j, n_j is the total number of words in d_j, and ε_i is the normalized entropy of w_i, computed by

$$\varepsilon_i = -\frac{1}{\log D}\sum_{j=1}^{D}\frac{c_{i,j}}{t_i}\log\frac{c_{i,j}}{t_i}, \tag{16}$$

where t_i is the total number of times term w_i appears in the training corpus. In the second stage, we project words and documents into a lower dimensional space by performing singular value decomposition (SVD) for matrix $\mathbf{A}$

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T \approx \mathbf{U}_R\Sigma_R\mathbf{V}_R^T = \mathbf{A}_R, \tag{17}$$

where Σ_R is a reduced $R \times R$ diagonal matrix with singular values, $\mathbf{U}_R$ is an $M \times R$ matrix whose columns are the first R eigenvectors derived from word-by-word correlation matrix $\mathbf{AA}^T$, and $\mathbf{V}_R$ is a $D \times R$ matrix whose columns are the first R eigenvectors derived from the document-by-document correlation matrix $\mathbf{A}^T\mathbf{A}$. The matrices $\mathbf{U}$, Σ, and $\mathbf{V}$ are original full matrices for $\mathbf{U}_R$, Σ_R, and $\mathbf{V}_R$, respectively. The reduced dimension has the property $R < \min(M, D)$. After the projection, each column of $\Sigma_R \mathbf{V}_R^T$ characterizes the location of a particular document in the reduced *R*-dimensional semantic space. Also, we can perform document clustering [Bellegarda 2000; Bellegarda *et al*. 1996] in the common semantic space. Each cluster consists of related documents in the semantic space. In general, each cluster in the semantic space reflects a particular semantic topic, which is helpful for integration in language modeling. During document clustering, the similarity of documents and topics in the common semantic space is determined by a cosine measure

$$\text{sim}(\mathbf{d}_j, \mathbf{t}_k) = \cos(\mathbf{U}_R^T \mathbf{d}_j, \mathbf{U}_R^T \mathbf{t}_k) = \frac{\mathbf{d}_j^T \mathbf{U}_R \mathbf{U}_R^T \mathbf{t}_k}{| \mathbf{U}_R^T \mathbf{d}_j \| \mathbf{U}_R^T \mathbf{t}_k |}, \tag{18}$$

where $\mathbf{d}_j$, $\mathbf{t}_k$ are the vectors constructed by document *j* and document cluster *k*, respectively. $\mathbf{U}_R^T \mathbf{d}_j$ and $\mathbf{U}_R^T \mathbf{t}_k$ are the projected vectors in the semantic space. By assigning topics to different documents, we can estimate the topic-dependent unigram $p(w_i | \mathbf{t}_k)$ and incorporate this information into the *n*-gram model. In what follows, we present two approaches for integrating the LSA information into the semantic language model, namely the linear interpolation approach and the maximum entropy approach.

3.2 Integration via Linear Interpolation

Linear interpolation (LI) [Rosenfeld 1996] is a simple approach to combining information sources from *n*-grams and semantic topics. To find the LI *n*-gram model, we first construct a pseudo document-vector from a particular historical context h. Using the projected document vector, we apply the nearest neighbor rule to detect the closest semantic topic $\mathbf{t}_k$ corresponding to history h. Given *n*-gram model $p_{\text{n}}(w | h)$ and topic-dependent unigram model $p(w | \mathbf{t}_k)$, the hybrid LI language model is computed by

$$p_{\mathbf{LI}}(w | h) = k_{\mathbf{n}} p_{\mathbf{n}}(w | h) + k_{\mathbf{t}} p(w | \mathbf{t}_k), \tag{19}$$

where the interpolation coefficients have the properties $0 < k_{\text{n}}, k_{\text{t}} \le 1$ and $k_{\text{n}} + k_{\text{t}} = 1$. Without the loss of generalization, an *n*-gram model and a topic-dependent model are integrated using fixed weights. Also, the expectation-maximization (EM) algorithm [Dempster *et al*. 1977] can be applied to dynamically determine the value of these weights by minimizing the overall perplexity.

3.3 Integration via Maximum Entropy

More importantly, we present a new ME language model combining information sources of *n*-grams and semantic topics. *N*-grams and semantic topics serve as constraints for the ME estimation. As shown in Table 2, two information sources partition the event space so as to obtain feature functions. Here, the trigram model is considered. Let w_i denote the current word to be predicted by its historical words. The columns and rows represent different constraints that are due to trigrams and semantic topics, respectively. The event space is partitioned into events $E_\mathbf{n}$ and $E_\mathbf{t}$ for different cases of *n*-grams and semantic topics, respectively. It comes out of the probability of the joint event $p(E_\mathbf{n}, E_\mathbf{t})$ to be estimated.

Table 2. Event space partitioned according to trigrams and semantic topics

$w = w_i$	h ends in w_1 $(E_{\text{n}1})$	h ends in w_1, w_2 $(E_{\text{n}2})$	h ends in w_2, w_3 $(E_{\text{n}3})$	$\cdots$
$h \in \mathbf{t}_1$ $(E_{\text{t}1})$	$p(E_{\text{n}1}, E_{\text{t}1})$	$p(E_{\text{n}2}, E_{\text{t}1})$	$p(E_{\text{n}3}, E_{\text{t}1})$	$\cdots$
$h \in \mathbf{t}_2$ $(E_{\text{t}2})$	$p(E_{\text{n}1}, E_{\text{t}2})$	$p(E_{\text{n}2}, E_{\text{t}2})$	$p(E_{\text{n}3}, E_{\text{t}2})$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$

Accordingly, the feature function for each column or *n*-gram event is given by

$$f_i^{\text{n}}(h,w) = \begin{cases} 1 & \text{if } h \text{ ends in } w_{i-1}, w_{i-2} \text{ and } w = w_i \\ 0 & \text{otherwise} \end{cases} . \tag{20}$$

In addition, the feature function for each row or semantic topic event has the form

$$f_i^{\text{t}}(h,w) = \begin{cases} 1 & \text{if } h \in \mathbf{t}_k \text{ and } w = w_i \\ 0 & \text{otherwise} \end{cases} . \tag{21}$$

We can build constraints corresponding to the trigrams and semantic topics as follows:

Trigram:

$$\sum_{h,w} \tilde{p}(h) p(w|h) f_i^{\text{n}}(h,w) = \sum_{h,w} \tilde{p}(h,w) f_i^{\text{n}}(h,w) = \tilde{p}(w_{i-2}, w_{i-1}, w_i) . \tag{22}$$

Semantic topics:

$$\sum_{h,w} \tilde{p}(h) p(w|h) f_i^{\text{t}}(h,w) = \sum_{h,w} \tilde{p}(h,w) f_i^{\text{t}}(h,w) = \tilde{p}(h \in \mathbf{t}_k, w_i) . \tag{23}$$

Under these constraints, we apply the IIS procedure described in Section 2.1 to estimate feature parameters λ_i^{n} and λ_i^{t}, used for combining information sources from trigrams and semantic topics, respectively. Finally, the solution provided by the ME semantic language modeling $p_{\text{ME}}(w|h)$ is computed by substituting λ_i^{n} and λ_i^{t} into Eqs. (9) and (10). We will compare the performance of LI language model $p_{\text{LI}}(w|h)$ and ME language model $p_{\text{ME}}(w|h)$ in the following experiments.

4. Experimental Results

In this study, we evaluate the proposed ME language model by measuring the model perplexity and the character-error rate in continuous speech recognition. The conventional *n*-gram language model is used as the baseline, while the ME language model proposed by Wu and Khudanpur [2002] is also employed for comparison. In addition, we also compare the maximum-entropy-based (ME) hybrid language model with the linear-interpolation-based (LI) hybrid language model. In the experiments, the training corpus for language modeling was composed of 5,500 Chinese articles (1,746,978 words in total) of the TDT2 Corpus, which were collected from the XinHua News Agency [Cieri *et al.* 1999] from January to June in 1998. The TDT2 corpus contained the recordings of broadcasted news audio developed for the tasks of cross-lingual cross-media Topic Detection and Tracking (TDT) and speech recognition. The audio files were recorded in single channel at 16 KHz in 16-bit linear SPHERE files. We used a dictionary of 32,909 words provided by Academic Sinica, Taiwan. 18,539 words in this dictionary occurred at least once in the training corpus. When carrying out the LSA procedure, we built a $32,909 \times 5,500$ word by document matrix $\mathbf{A}$ from the training data. We used MATLAB to implement SVD and *k*-means operations and, accordingly, performed document clustering and determined semantic topic vectors. The topic-dependent unigram was interpolated with the general unigram for model smoothing. The dimensionality of the LSA model was reduced to $R = 100$. We performed the IIS algorithm with 30 iterations. All language models were smoothed using Jelinek-Mercer smoothing [Jelinek and Mercer 1980], which is calculated based on the interpolation of estimated distribution and lower order *n*-grams.

4.1 Convergence of the IIS Algorithm

First of all, we examine the convergence property of the IIS algorithm. Figure 2 shows the log-likelihood of the training data using the ME language model versus different IIS iterations. In this evaluation, the number of semantic topics was set at 30. The ME model that combines the features of trigram and semantic topic information was considered. Typically, the log-likelihood increases consistently with the IIS iterations. The IIS procedure for the ME integration converged after five or six iterations.

4.2 Evaluation of Perplexity

One popular evaluation metric for language models for speech recognition is the *perplexity* of test data. Perplexity can be interpreted as the average number of branches in the text. The higher the perplexity, the more branches the speech recognition system should consider. Generally speaking, a language model with lower perplexity implies less confusion in recognition and achieves higher speech-recognition accuracy. To evaluate the perplexity, we selected an additional 734 Chinese documents from the XinHua News Agency, which consisted of 244,573 words, as the test data. First, we evaluated the effect of the length of history h for topic identification. The perplexities of LI and ME models are shown in Figures 3 and 4, respectively. Here, *C* represents the number of document clusters or semantic topics. In the LI implementation, for each length of history h, the interpolation weight with the lowest perplexity was empirically selected. It is obvious that the proposed ME language model outperforms Wu's ME language model [Wu and Khudanpur 2002] and the ME language model outperforms the LI language model. Furthermore, a larger *C* produces lower perplexity and the case that considering 50 historical words obtains the lowest perplexity. Accordingly, we fixed the length of h at 50 in the subsequent experiments. Table 3 details the perplexities for bigram and semantic language models based on LI and ME. We found that the perplexity was reduced from 451.4 (for the baseline bigram) to 444.7 by using Wu's method and to 441 by using the proposed method when the combination was based on linear interpolation (LI) and the topic number was 30. With the maximum entropy (ME) estimation, the perplexity was further reduced to 399 and 393.7 by using Wu's method and the proposed method, respectively. No matter whether Wu's method or the proposed method was used, the ME language model consistently outperformed the LI language model with different numbers of semantic topics. We also evaluated these models based on the trigram features. The results are summarized in Table 4. We can see that, by integrating latent semantic information into the trigram model, the perplexity is reduced from 376.6 (for the baseline trigram) to 345.3 by using the LI model and to 309.3 by using the ME model, for the case of *C*=100. The experimental results again demonstrate that the performance improves with the number of semantic topics and that the proposed method consistently outperforms Wu's method, though the improvement is not very significant.

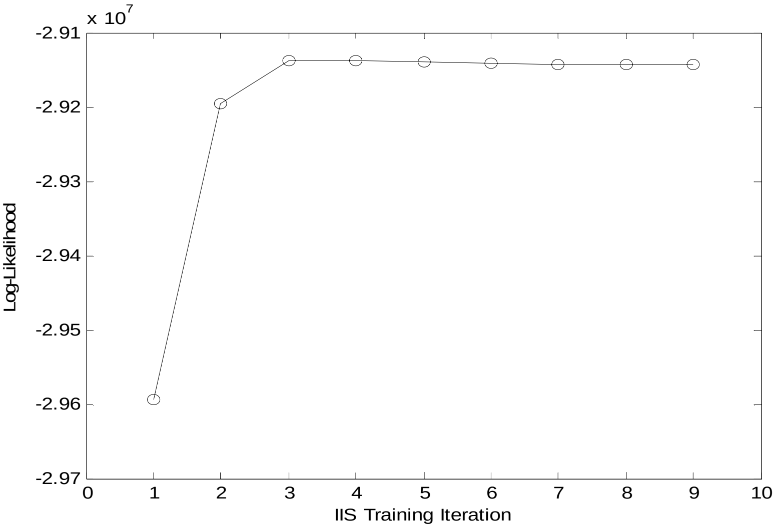

Figure 2. Log-Likelihood of training data versus the number of IIS iterations

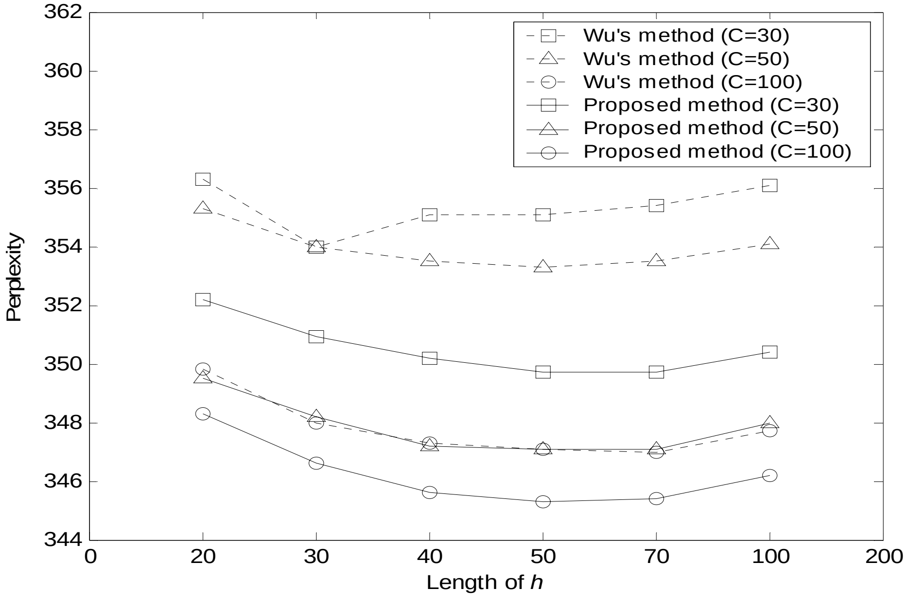

Figure 3. Perplexity of the LI model versus the length of history

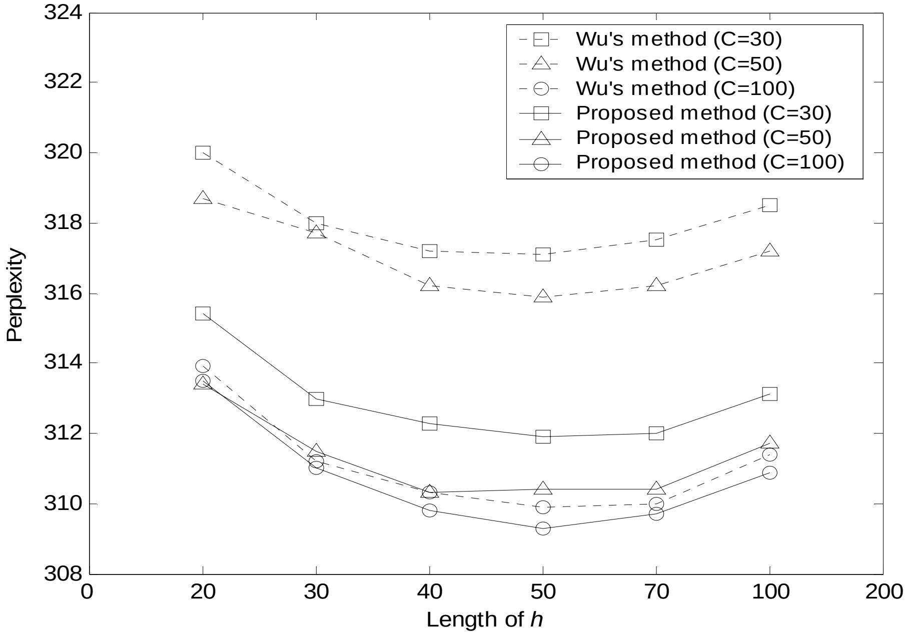

Figure 4. Perplexity of the ME model versus the length of history

Table 3. Comparison of perplexity for bigram, LI and ME semantic language models

	Bigram	Wu's method		Proposed method	
		LI	ME	LI	ME
C=30	451.4	444.7	399	441	393.7
C=50		442.9	402	438	394.8
C=100		437	397.2	435.7	401.2

Table 4. Comparison of perplexity for trigram, LI and ME semantic language models

	Trigram	Wu's method		Proposed method	
		LI	ME	LI	ME
C=30	376.6	355.1	317.1	349.7	311.9
C=50		353.3	315.9	347.1	310.4
C=100		347.1	309.9	345.3	309.3

4.3 Evaluation of Speech Recognition

In addition to perplexity, we evaluated the proposed language models for a continuous Mandarin speech recognition task. Character-error rates are reported for comparison. The initial speaker-independent, hidden Markov models (HMM’s) were trained by the benchmark Mandarin speech corpus TCC300 [Chien and Huang 2003], which was recorded in office environments using close-talking microphones. We followed the construction of context-dependent sub-syllable HMM’s for Mandarin speech presented in [Chien and Huang 2003]. Each Mandarin syllable was modeled by right context-dependent states where each state had, at most, 32 mixture components. Each feature vector consisted of twelve Mel-frequency cepstral coefficients, one log energy, and their first derivatives. The maximum *a posteriori* (MAP) adaptation [Gauvian and Lee 1994] was performed on the initial HMM’s using 83 training sentences (about 10 minutes long), from Voice of America (VOA) news, in the TDT2 corpus for corrective training. The additional 49 sentences selected from VOA news were used for speech recognition evaluation. This test set contained 1,852 syllables, with a total length of 6.6 minutes. To reduce the complexity of the tree copy search in decoding a test sentence, we assumed each test sentence corresponded to a single topic, which was assigned according to the nearest neighbor rule. Due to the above complexity, in this study we only implemented the language model by combining bigram and semantic information in our recognizer. Figure 5 displays the character-error rate versus the number of topics. We can see that the character-error rate decreases in the beginning and then increases as the number of topics increases. Basically, more topics provide higher resolution for representing the information source. However, the model with higher resolution requires larger training data for parameter estimation. Otherwise, the overtraining problem occurs and the performance degrades accordingly. The character-error rates used in Wu's method and the proposed method are summarized in Table 5. In the case of *C*=50, the proposed LI model can achieve an error-rate reduction of 8.5% compared to the bigram model, while the proposed ME model attains a 16.9% error-rate reduction. The proposed method in general achieves lower error rates compared to Wu's method.

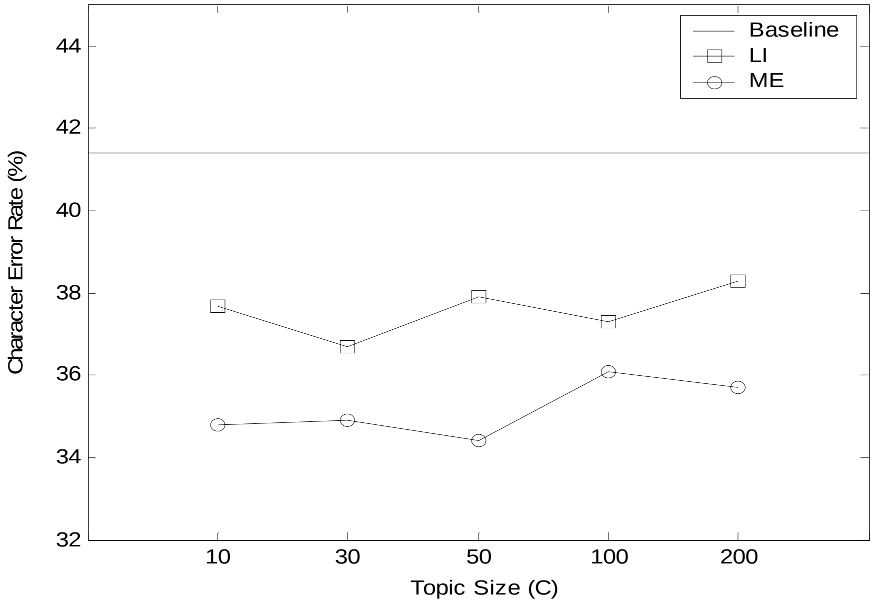

Figure 5. Character error rate (%) versus the number of topics

Table 5. Comparison of character error rate (%) for bigram, LI and ME semantic language models

	Bigram	Wu's method		Proposed method	
		LI	ME	LI	ME
C=30	41.4	38.9	36.4	36.7	34.9
C=50		38.1	36.8	37.9	34.4
C=100		38.3	36.5	37.3	36.1

To evaluate the statistical significance of performance difference between the proposed method and Wu's method, we applied the *matched-pairs* test [Gillick and Cox 1989] to test the hypothesis that the number of recognition errors that occur when using the proposed method is close to that with Wu's method. In the evaluation, we calculated the difference between character errors induced by Wu's method E_a and the proposed method E_t for each utterance. If the mean of variable $z = E_t - E_a$ was zero, we accepted the conclusion that these two methods are not statistically different. To carry out the test, we calculated the sample mean $\bar{\mu}_z$ and sample variance $\bar{\sigma}_z$ from N utterances and determined the test statistic $\omega = \bar{\mu}_z / (\bar{\sigma}_z / \sqrt{N})$. Then, we computed the probability $P = 2\Pr(z \geq |\omega|)$ and compared P with a chosen significance level α . When $P < \alpha$, this hypothesis was rejected or, equivalently, the improvement obtained with the proposed method was statistically significant. In the evaluation, we applied the respective best case of Wu's method and the proposed method (i.e., ME language modeling, and C=30 for Wu's method but C=50 for the proposed method) in the test and obtained a P value of 0.0214. Thus, at the $\alpha = 0.05$ level of significance, the proposed method is better than Wu's method. That is, the proposed LSA based topic extraction is desirable for discovering semantic information for language modeling.

5. Conclusions

We have presented a new language modeling approach to overcome the drawback of lacking long-distance dependencies in a conventional *n*-gram model that is due to the assumption of the Markov chain. We introduced a new long-distance semantic information source, called the semantic topic, for knowledge integration. Instead of extracting the topic information from the original document space, we proposed extracting semantic topics from the LSA space. In the constructed LSA space with reduced dimensionality, the latent relation between words and documents was explored. The *k*-means clustering technique was applied for document clustering. The estimated clusters were representative of semantic topics embedded in general text documents. Accordingly, the topic-dependent unigrams were estimated and combined with the conventional *n*-grams. When performing knowledge integration, both linear interpolation and maximum entropy approaches were carried out for comparison. Generally speaking, linear interpolation was simpler for implementation. LI combined two information sources through a weighting factor, which was estimated by minimizing the overall perplexity. This weight was optimized globally such that we could not localize the use of weights for different sources. To achieve an optimal combination, the ME principle was applied. Each information source served as a set of constrains to be satisfied for model combination. The IIS algorithm was adopted for constrained optimization. From the experimental results of Chinese document modeling and Mandarin speech recognition, we found that ME semantic language modeling achieved a desirable performance in terms of model perplexity and character-error rates. The combined model, through linear interpolation, achieved about an 8.3% perplexity reduction over the trigram model. The proposed semantic language model did compensate the insufficiency of long-distance information in a conventional *n*-gram model. Furthermore, the

ME semantic language model reduced perplexity by 17.9%. The ME approach did provide a delicate mechanism for model combination. Also, in the evaluation of speech recognition, the ME semantic language model obtained a 16.9% character-error rate reduction over the bigram model. The ME model was better than the LI model for speech recognition. In the future, we will validate the coincidence between the semantic topics discovered by the proposed method and the semantic topics labeled manually. We will also extend the evaluation of speech recognition using higher-order *n*-gram models over a larger collection of speech data.

REFERENCES

Bellegarda, J., "Exploiting latent semantic information in statistical language modeling," *Proceedings of the IEEE*, 88(8), 2000, pp. 1279-1296.

Bellegarda, J., J. Butzberger, Y. Chow, N. Coccaro, and D. Naik, "A novel word clustering algorithm based on latent semantic analysis," *IEEE Proceedings of International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 1, 1996, pp. 172-175.

Berger, A., S. Della Pietra, and V. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, 22(1), 1996, pp. 39-71.

Berry, M., S. Dumais, and G. O'Brien, "Using linear algebra for intelligent information retrieval," *SIAM Review*, 37(4), 1995, pp. 573-595.

Chelba, C. and F. Jelinek, "Structured language modeling," *Computer Speech and Language*, 14(4), 2000, pp. 283-332.

Chien, J.-T., and C.-H. Huang, "Bayesian learning of speech duration model," *IEEE Transactions on Speech and Audio Processing*, 11(6), 2003, pp. 558-567.

Chien, J.-T., and H.-Y. Chen, "Mining of association patterns for language modeling," *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2, 2004, pp. 1369-1372.

Chien, J.-T., M.-S. Wu, and H.-J. Peng, "Latent semantic language modeling and smoothing," *International Journal of Computational Linguistics and Chinese Language Processing*, 9(2), 2004, pp. 29-44.

Chueh, C.-H., J.-T. Chien, and H. Wang, "A maximum entropy approach for integrating semantic information in statistical language models," *Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2004, pp. 309-312.

Cieri, C., D. Graff, M. Liberman, N. Martey, and S. Strassel, "The TDT-2 text and speech corpus," *Proc. of the DARPA Broadcast News Workshop*, 28Feb-3Mar 1999.

Clarkson, P., and A. Robinson, "Language model adaptation using mixtures and an exponential decay cache," *IEEE Proceedings of International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2, 1997, pp. 799- 802.

Darroch, J., and D. Ratcliff, "Generalized iterative scaling for log-linear models," *The Annals of Mathematical Statistics*, 43, 1972, pp. 1470-1480.

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society of Information Science*, 41, 1990, pp. 391-407.

Della Pietra, S., V. Della Pietra, and J. Lafferty, "Inducing features of random field," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4), 1997, pp. 380-393.

Della Pietra, S., V. Della Pietra, R. Mercer, and S. Roukos, "Adaptive language modeling using minimum discriminant estimation," *IEEE Proceedings of International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 1, 1992, pp. 633-636.

Dempster, A., N. Laird, and D.Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, 39(1), 1977, pp. 1-38.

Federico, M., "Efficient language model adaptation through MDI estimation," *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, 1999, pp. 1583-1586.

Florian, R., and D. Yarowsky, "Dynamic nonlocal language modeling via hierarchical topic-based adaptation," *Proc. 37th Annual Meeting of ACL*, 1999, pp. 167-174.

Gauvain, J.-L., and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observation of Markov chain," *IEEE Transactions on Speech and Audio Processing*, 2(4), 1994, pp. 291-298.

Gillick, L., and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," *IEEE Proceedings of International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 1989, pp. 532-535.

Hofmann, T., "Probabilistic latent semantic indexing," *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50-57.

Jaynes, E., "Information theory and statistical mechanics," *Physics Reviews*, 106(4), 1957, pp. 620-630.

Jelinek, F., and R. L. Mercer, "Interpolated estimation of Markov source parameters from sparse data," *Proc. Workshop on Pattern Recognition in Practice*, 1980, pp. 381-402.

Khudanpur, S., and J. Wu, "Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling," *Computer Speech and Language*, 14, 2000, pp. 355-372.

Kuhn, R., and R. de Mori, "A cache based natural language model for speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6), 1992, pp. 570-583.

Ponte, J., and W. Croft, "A language modeling approach for information retrieval," *Proc. ACM SIGIR on Research and Development in Information Retrieval*, 1998, pp. 275-281.

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

Wang, S., D. Schuurmans, F. Peng, and Y. Zhao, "Learning mixture models with the regularized latent maximum entropy principle," *IEEE Transactions on Neural Networks,* 15(4), 2004, pp. 903-916.

Wang, S., D. Schuurmans, F. Peng, and Y. Zhao, "Semantic *n*-gram language modeling with the latent maximum entropy principle," *IEEE Proceedings of International Conference on Acoustic, Speech and Signal Processing (ICASSP),* 1, 2003, pp. 376-379.

Wu, J., and S. Khudanpur, "Building a topic-dependent maximum entropy model for very large corpora," *IEEE Proceedings of International Conference on Acoustic, Speech and Signal Processing (ICASSP),* 1, 2002, pp. 777-780.

Zhou, G. D., and K. T. Lua, "Interpolation of *n*-gram and mutual-information based trigger pair language models for Mandarin speech recognition," *Computer Speech and Language*, 13, 1999, pp. 125-141.