A MAXIMUM ENTROPY APPROACH FOR INTEGRATING SEMANTIC INFORMATION IN STATISTICAL LANGUAGE MODELS

Chuang-Hua Chueh [a], *Jen-Tzung Chien* [a] *and Hsin-min Wang* [b]

[a] Department of Computer Science and Information Engineering, Cheng Kung University, Tainan
[b] Institute of Information Science, Academia Sinica, Taipei
E-mail: sgxxx@chien.csie.ncku.edu.tw, jtchien@mail.ncku.edu.tw, whm@iis.sinica.edu.tw

ABSTRACT

In this paper, we propose an adaptive statistical language model, which successfully incorporates the semantic information into an *n*-gram model. Traditional *n*-gram models exploit only the immediate context of history. We first introduce the semantic topic as a new source to extract the long distance information for language modeling, and then adopt the maximum entropy (ME) approach instead of the conventional linear interpolation method to integrate the semantic information with the *n*-gram model. Using the ME approach, each information source gives rise to a set of constraints, which should be satisfied to achieve the hybrid model. In the experiments, the ME language models trained using the China Times newswire corpus achieved 40% perplexity reduction over the baseline bigram model.

1. INTRODUCTION

Language modeling plays an important role in automatic speech recognition (ASR). When observing the speech signal, the corresponding word sequence is obtained to maximize the *posteriori* probability. The classification rule is based on Bayes' theory.

$$\hat{W} = \arg\max_{W} P(W \mid O) = \arg\max_{W} P(O \mid W)P(W) \qquad (1)$$

where O is the observed speech signal, while $P(O|W)$ and $P(W)$ are, respectively, the acoustic model score and the language model score associated with word sequence W. The language model is used to characterize the regularities in natural language. It is also popular to apply language models in machine translation, document classification, information retrieval [11] and other applications.

There are various kinds of language models feasible to extract different linguistic regularities of natural language. The most common and successful one is to use the statistical *n*-gram model [13], which is efficient at capturing local lexical regularities. The structural language model [5] effectively exploited the relevant syntactic regularities based on the predefined grammar rules. Also, the semantic language model [1] can easily explore the document-level semantic regularities. However, none of them can simultaneously take into account the local lexical, hierarchical syntactic and semantic information. Therefore, it is desirable to estimate a hybrid model that integrates multiple information knowledge.

The simplest way to combine different information sources is to apply linear interpolation. Using this approach, each information source is characterized by a separate model. Different information sources are combined using weighted averaging, which minimize the overall perplexity without considering their strengths and weaknesses in particular contexts. In other words, the weights are optimized globally instead of locally. The hybrid model obtained in this way could not guarantee the optimal use of the different information sources.

Another important approach is based on the Jaynes' maximum entropy (ME) principle [10]. This approach presents a procedure to set up the probability distributions on the basis of partial knowledge. Different from linear interpolation, this approach attempts to capture all information provided by various knowledge sources. The maximum entropy principle was first applied to language modeling in [3]. Using the ME approach, the information source provided by trigger pairs was incorporated into the *n*-gram model to integrate long distance dependencies with local lexical regularities [12]. Although trigger pairs can be used to exploit the long distance word relationships, this approach only considers the frequently co-occurred word pairs in the training data, but ignores the ones with a lower frequency. Consequently, some important semantic information could be lost. To reduce the information loss, some topic-based language models were proposed [15]. In this paper, a new long-distance, semantic knowledge source called the semantic topic, is introduced. The latent semantic analysis (LSA) [4][7] is used to exploit the topic information. The relationships between semantic topics and target words are incorporated into the new language model under the ME framework.

2. MAXIMUM ENTROPY PRINCIPLE

The basic ME principle aims to subtly model what we know, and assume nothing about what we do not know. Namely, we choose a model that is consistent with all the information we have; but otherwise, make the distribution in the model as uniform as possible.

In the following, we discuss this method for combining knowledge sources for language modeling [3]. First, each knowledge source provides a set of constraints, which should be satisfied to find the unique ME solution. These constraints are typically expressed as marginal distributions.

Given features $f_1,..., f_N$ specifying the properties extracted from data, the expectation of f_i with respect to the empirical distribution $\tilde{p}(h,w)$ of history *h* and word *w* is calculated by

$$\tilde{p}(f_i) = \sum_{h,w} \tilde{p}(h,w) f_i(h,w) \cdot \qquad (2)$$

where $f_i()$ is a binary-valued feature function. The expectation with respect to the target distribution $p(h,w)$ is calculated by

$$p(f_i) = \sum_{h,w} p(h,w) f_i(h,w) \cdot \qquad (3)$$

Because the target distribution $p(h,w)$ is required to exploit all the information provided by these features, the constraints are accordingly specified as

$$p(f_i) = \tilde{p}(f_i), \quad \text{for } i = 1,..., N \ . \qquad (4)$$

Under these constraints, we maximize the entropy of the distribution $p(h,w)$, which is equivalent to maximize its uniformity. The Lagrange optimization is adopted to solve the constrained optimization problem. For each feature f_i, we introduce a Lagrange multiplier λ_i. The Lagrangian $\Lambda(p,\lambda)$ is defined as

$$\Lambda(p,\lambda) = H(p) + \sum_{i=1}^{N} \lambda_i \left[p(f_i) - \tilde{p}(f_i) \right], \qquad (5)$$

where

$$H(p) = -\sum_{h,w} p(h,w) \log p(h,w) \cdot \qquad (6)$$

Finally, the target distribution $p(h,w)$ is estimated as

$$p(h,w) = \frac{1}{Z_\lambda} \exp\left(\sum_{i=1}^{N} \lambda_i f_i(h,w) \right), \qquad (7)$$

where Z_λ, computed by

$$Z_\lambda = \sum_{h,w} \exp\left(\sum_{i=1}^{N} \lambda_i f_i(h,w) \right), \qquad (8)$$

is a normalization term determined by the probabilistic property that $\sum_{h,w} p(h,w) = 1 \cdot$

The iterative scaling algorithm (GIS/IIS) [3][6][8] can be used to find the Lagrange parameters λ. The IIS algorithm is briefly described as follows.

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

After estimating the optimal parameters λ, we can calculate the ME language model.

In [14], the ME principle was extended to the latent ME (LME) for modeling hidden variables. The probabilistic LSA was successfully incorporated into an *n*-gram model by serving the semantic information as the latent variables. Here, we use the semantic information as the explicit features. The LSA is adopted to build the semantic topics.

3. INTEGRATION OF SEMANTIC INFORMATION AND *N*-GRAMS

In this section, we introduce a new knowledge source, which contains the long distance semantic information. Assuming the occurrence of a word is highly related to the topic of current discourse, we apply LSA to find semantic topics. Furthermore, we combine semantic topics and traditional *n*-grams based on the ME principle.

3.1. Semantic topic

LSA is a popular technique in the areas of information retrieval [4] and semantic inference [1]. The primary assumption is that there exist some latent structures embedded in the words across documents. LSA is able to exploit this structure. The first step of LSA is to construct a $M \times D$ word-by-document matrix, called **W**. Here *M* and *D* represent the vocabulary size and the number of documents respectively. The expression for the (*i*,*j*) entry of the matrix is [1]:

$$w_{i,j} = (1 - \varepsilon_i) \frac{c_{i,j}}{n_j}, \qquad (9)$$

where $c_{i,j}$ is the number of times w_i appears in 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:

$$\varepsilon_i = -\frac{1}{\log D}\sum_{j=1}^{D}\frac{c_{i,j}}{t_i}\log\frac{c_{i,j}}{t_i}, \tag{10}$$

where t_i is the total number of times w_i appears in the corpus.

In the second step, the words and documents are projected to a low dimensional space through the singular value decomposition (SVD):

$$\mathbf{W} \approx \hat{\mathbf{W}} = \mathbf{USV}^T, \tag{11}$$

where **S** is a $R \times R$ diagonal matrix with singular values, **U** is a $M \times R$ matrix whose columns are the eigenvectors derived from the word-by-word correlation matrix, $\mathbf{WW}^T$, **V** is a $D \times R$ matrix whose columns are the eigenvectors derived from the document-by-document correlation matrix, $\mathbf{W}^T\mathbf{W}$, and *R*<<min(*M*,*N*) is the order of the decomposition. After the projection, each column vector of $\mathbf{SV}^T$ characterizes the position of a particular document in the *R*-dimensional semantic space.

We can also perform a document clustering [1][2] based on the LSA projection. Each cluster is composed of some documents similar to each other in the semantic space. In other words, each cluster can reflect a particular topic, namely the semantic topic, which is used to determine the topic of a particular context according to a similarity measure. According to topic assignment of a history, we can obtain $p_{ST}(w \mid h)$ from the training data easily. Before describing the ME approach, we introduce the combination using linear interpolation.

3.2. Linear interpolation

Linear interpolation [12] can be used to combine the information sources from bigrams and semantic topics, too. Given two models, $p_{n-gram}(w \mid h)$ and $p_{ST}(w \mid h)$, the hybrid model can be computed by

$$p_{LI}(w \mid h) = k_1 p_{n-gram}(w \mid h) + k_2 p_{ST}(w \mid h), \tag{12}$$

where $0 < k_1, k_2 \le 1$ and $\sum_{i=1}^{2} k_i = 1$.

Using this method, the *n*-gram model and the semantic topic model are integrated with the fixed weights. The Expectation-Maximization (EM) algorithm [9] can be applied to determine these weights that minimize the overall perplexity.

3.3. ME approach

Using the ME approach, *n*-grams and semantic topics are both viewed as constraints. The features from these two information sources are partitioned into the event space as shown in Table 1.

Table 1. Event space partitioned according to bigrams and semantic topics

$w = w_p$	h ends in w_1 (E_{r1})	h ends in w_2 (E_{r2})	$\cdots$
$h \in T_1$ (E_{l1})	$p(E_{r1}, E_{l1})$	$p(E_{r2}, E_{l1})$	$\cdots$
$h \in T_2$ (E_{l2})	$p(E_{r1}, E_{l2})$	$p(E_{r2}, E_{l2})$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\cdots$

In Table 1, w_p is the predictive word. The columns and rows correspond to different bigram conditions and semantic topic conditions, respectively. The event space is partitioned into different events according to the corresponding bigrams and semantic topics. The feature for each column (bigram) is

$$f_i^B(h,w) = \begin{cases} 1 & \text{if } h \text{ ends in } w_i, w = w_p \\ 0 & \text{otherwise} \end{cases}, \tag{13}$$

and the feature for each row (semantic topic) is

$$f_i^S(h,w) = \begin{cases} 1 & \text{if } h \in T_i, w = w_p \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

The corresponding two constraints are written as follows,

Bigram:

$$\sum_{h,w} p(h,w) f_i^B(h,w) = \sum_{h,w} \tilde{p}(h,w) f_i^B(h,w) = \tilde{p}(w_i, w_p)\cdot \tag{15}$$

Semantic topics:

$$\sum_{h,w} p(h,w) f_i^S(h,w) = \sum_{h,w} \tilde{p}(h,w) f_i^S(h,w) = \tilde{p}(h \in T_i, w_p)\cdot \tag{16}$$

Under these constraints, the ME procedure described in Section 2 can be applied to estimate the hybrid model, which combines the information sources from bigrams and semantic topics. Finally, the solution in the form of Equation (7) can be obtained.

4. EXPERIMENTAL RESULTS

In the following experiments, the training corpus composed of 1,000 newswire articles (343,851 words in total) collected from the China Times. We used a lexicon of 32,909 words, among which 13,516 words occur at least once in the training corpus. The test corpus, which was also collected from China Times, consisted of approximately 14,000 words.

The perplexities for the baseline bigram model and the hybrid models are summarized in Table 2, where *C* represents the number of document clusters. The reduced dimensionality of LSA is 100. We performed the IIS algorithm with 30 iterations. The K-means algorithm was applied to document clustering and semantic topic determination. From Table 2, it was found that the perplexity was reduced from 112.66 to 63.04 when the topic number was set as 30 and the ME approach was used to generate the hybrid model. Though the hybrid models by linear interpolation also have lower complexities than the baseline model, it is obvious that the ME approach outperforms the linear interpolation approach. The best perplexity reduction by the ME approach and the linear interpolation approach is 33.1% and 44%, respectively. The results not only show that the proposed semantic topics are useful for modeling long distance information but also show that the ME approach provides a desirable knowledge integration

Table 2. Experimental results on using China Times News corpus

		Baseline (Bigram)	Linear Interpolation	Maximum Entropy
Perplexity (Reduction %)	*C*=10	112.66 (-)	80.64 (28.2%)	73.45 (34.4%)
	C =30	112.66 (-)	75.34 (33.1%)	63.04 (44%)
	C =50	112.66 (-)	79.54 (29.3%)	68.35 (39.3%)

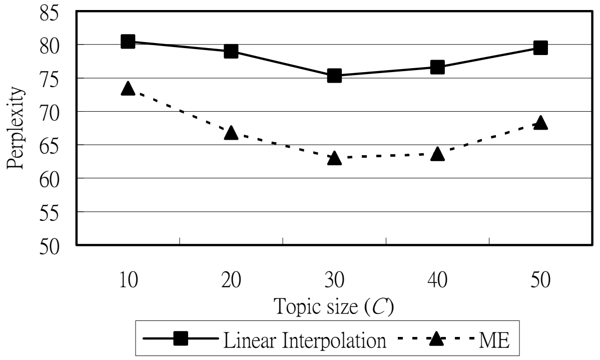

Figure 1. Relationship between topic size and perplexity

Figure 1 depicts the perplexity with respect to the topic size. It was found that the perplexity first decreased as the topic size increased, and then increased with further increase of topic size. The larger topic size produces the higher resolution for the information source. However, the model with higher resolution requires more training data to estimate more parameters. If the training data are insufficient, the overtraining problem will occur. It is a trade-off depending on the training data available.

5. CONCLUSION

Language modeling is used to capture different regularities of natural language. The statistical *n*-gram model is the most popular one. The limitation of the model is that it is not able to exploit long distance dependencies.

In this paper, we introduced a novel long distance, semantic information source called the semantic topic. Different from trigger pairs, the new information source considers the whole context instead of the relationships between highly related words. We applied LSA to extract the semantic topics by performing document clustering. We believe that each document cluster can present a particular topic at the semantic level. Furthermore, the maximum entropy principle is used to combine this semantic knowledge with the conventional *n*-gram model. The hybrid model obtained in this way achieves a significant perplexity reduction compared to the baseline *n*-gram model and the hybrid model obtained by linear interpolation.

REFERENCES

1. J. Bellegarda, “Exploiting latent semantic information in statistical language modeling,” *Proceedings of IEEE*, vol. 88, no. 8, pp. 1279-1296, August 2000.
2. J. Bellegarda, J. Butzberger, Y. Chow, N. Coccaro and D. Naik, “A novel word clustering algorithm based on latent semantic analysis,” In *Proc. of ICASSP96*, pp. I172-I175.
3. A. Berger, S. Della Pietra and V. Della Pietra, “A maximum entropy approach to natural language processing,” *Computational Linguistics*, Vol. 22, No. 1, pp. 39-71, 1996.
4. M. Berry, S. Dumais and G. O’Brien, “Using linear algebra for intelligent information retrieval,” *SIAM Review*, Vol. 37, No. 4, pp. 573-595, 1995.
5. C. Chelba and F. Jelinek, “Structured language modeling,” *Computer Speech and Language*, Vol. 14, No. 4, pp. 283-332, October 2000.
6. J. Darroch and D. Ratcliff. “Generalized iterative scaling for log-linear models,” *The Annals of Mathematical Statistics*, Vol. 43, pp. 1470-1480, 1972.
7. S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society of Information Science*, Vol. 41, pp. 391-407, 1990.
8. S. Della Pietra, V. Della Pietra and J. Lafferty, “Inducing features of random field,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 4, pp. 380-393, 1997.
9. A. Dempster, N. Laird and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, vol. 39, No. 1, pp. 1-38, 1997.
10. E. Jaynes, “Information theory and statistical mechanics,” *Physics Reviews*, Vol. 106, No. 4, pp. 620-630, 1957.
11. J. Pone and W. Croft, “A language modeling approach for information retrieval,” In *Proc. of ACM SIGIR98 on Research and Development in Information Retrieval*, pp. 275-281.
12. R. Rosenfeld, “A maximum entropy approach to adaptive statistical language modeling,” *Computer Speech and Language*, Vol. 10, pp. 187-228, 1996.
13. C. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, Vol. 27, pp. 398-403, 1948.
14. S. Wang, D. Schuurmans, F. Peng and Y. Zhao, “Semantic *n*-gram language modeling with the latent maximum entropy principle,” In *Proc. of ICASSP03*, I376-I379.
15. J. Wu and S. Khudanpur, “Building a topic-dependent maximum entropy model for very large corpora,” In *Proc. of ICASSP02*, I777-I780.