Interpolated Distanced Bigram Language Models for Robust Word Clustering

Nikoletta Bassiou* and Constantine Kotropoulos ‡
Department of Informatics, Aristotle University of Thessaloniki
Box 451 Thessaloniki 541 24, GREECE
* E-mail:nbassiou@zeus.csd.auth.gr
‡ E-mail: costas@zeus.csd.auth.gr

***Abstract*— Two methods for interpolating the distanced bigram language model are examined which take into account pairs of words that appear at varying distances within a context. The language models under study yield a lower perplexity than the baseline bigram model. A word clustering algorithm based on mutual information with robust estimates of the mean vector and the covariance matrix is employed in the proposed interpolated language model. The word clusters obtained by using the aforementioned language model are proved more meaningful than the word clusters derived using the baseline bigram.**

I. Introduction

Statistical Language Models (LM) have been used in a wide range of natural language processing tasks including speech recognition [1], word clustering [2], machine translation [3], and information retrieval [4], [5], [6].

In speech recognition systems, the N-gram language models [1] are mainly used, particularly in large vocabulary speech transcription tasks [7], [8]. However, since the main notion of these models is that the probability of a word depends only upon the last $N-1$ words, the number of parameters in N-gram models increases considerably as N increases, resulting in an increase in the size of the model and the data required for training. Therefore, low values of N, usually $2-4$, are employed causing the loss of long-term context information.

This drawback has motivated further research in alternative ways for extracting suitable long distance information. Trigger language models derive trigger word pairs using mutual information whose statistics are combined with the N-gram statistics in a maximum entropy framework [9]. Trigger word pairs provide long-distance information since the triggering word and the triggered one can be separated by several words. However, their selection is a complex issue, since different trigger word pairs display a markedly different behavior, which limits the potential of low frequency triggers. An extension of the trigger concept in the latent semantic analysis (LSA) where a more systematic framework is exploited to handle trigger pair selection is proposed in [10]. Cache language models boost the probability of a word or a word class seen in a long-term window over the history [11]. In [12], distance-2 bigrams are combined with standard bigrams and trigrams using maximum entropy, while in [13], a pruning strategy is proposed in order to obtain a variable-length model. In the latter approach, a full N-gram model with a large N is successively discarded from those N-grams having the smallest impact on the model. A tree-based language model that uses a longer context with a limited number of parameters is proposed in [14].

In this paper, we examine two different interpolation methods for distanced bigram models at varying distances within a context H. The first method conducts interpolation over the component probabilities of the models [7] whereas the second method interpolates the full models. The latter model and the baseline bigram are employed for word clustering using an algorithm similar to that proposed in [2], but including now robust estimators of the mean vector and the covariance matrix. The effectiveness of the interpolated models is compared with that of the baseline bigram by measuring the perplexity in leave-one-out experiments. In addition, the validity of the word clusters derived by employing the aforementioned language models in the clustering algorithm is studied. The Jaccard coefficient [15] has been computed in order to test if the clustering is not a matter of chance but it has been created systematically capturing a similarity inherent in the data. Moreover, we have computed the average mutual information between the classic bigram and the distance bigram for $d=2$. The experimental results indicate that the clustering employing robust distanced bigrams is less predictable by the clustering obtained using the classic bigram than that based on the distance bigram without robust estimators of the mean vector and the covariance matrix. The experiments were conducted on the Reuters corpus collection.

The outline of the paper is as follows. Language modeling is described in Section II. A discounting method suitable for the interpolated language models under study is outlined in Section III. The word clustering approach is explained in Section IV and experimental results are demonstrated in Section V. Finally, conclusions are drawn in Section VI.

II. Language Modeling

Given a sequence of M words $\mathbf{W} = w_1, w_2, \ldots, w_M$ a language model estimates the *a priori* probability $P(\mathbf{W}) = P(w_1, w_2, \ldots, w_M)$. In the following, the N-gram, the distanced bigram, and the interpolation of language models are reviewed briefly.

A. The N-Gram Model

In this traditional stochastic language model, the current word is predicted based on the preceding word (bigram) or the preceding $N-1$ words (N-gram) expecting that most of the relevant syntactic information lies in the immediate past. Its *a priori* probability $P(\mathbf{W})$ is expressed using conditional probabilities by

$$P(\mathbf{W}) = P(w_1) \prod_{i=2}^{M} P(w_i|w_1, \ldots, w_{i-1}). \quad (1)$$

For simplicity reasons, by invoking Markov chain assumptions, $P(w_i|w_1, \ldots, w_{i-1})$ is approximated by $P(w_i|w_{i-N+1}, ..., w_{i-1})$ with $N = 2, 3$ or 4 at the expense of preserving the syntactic and semantic information from the more distant words.

B. The Distanced Bigram Model

In an attempt to reduce the number of free parameters of the N-Gram model and to maintain the modeling capacity, long-distance bigrams are proposed in [7], [8]. In this model, the notion of distance d is added to the bigrams of the simple N-gram model. A word w_i lies at distance d from the word w_j, when $i = j - d$. That is, when w_i is the dth word before w_j. For $d = 1$, we get the baseline bigram. The probability $P_d(\mathbf{W})$ of a word sequence in the proposed model is expressed by

$$P_d(\mathbf{W}) = \prod_{j=1}^{d} P(w_j) \prod_{j=d+1}^{M} P_d(w_j|w_i). \quad (2)$$

The probability $P_d(w_j|w_i)$ is expressed with the help of the relative frequency approach as follows:

$$P_d(w_j|w_i) \simeq \hat{P}_d(w_j|w_i) = \frac{N_d(w_i, w_j)}{N(w_i)} \quad (3)$$

where $N_d(w_i, w_j)$ is the number of times the word w_i appears to be the dth word before w_j. Similarly, $N(w_i)$ is the number of times the word w_i is met in the training corpus.

C. Interpolating Language Models

When long-distanced bigrams are met in H different distances, $P^H(\mathbf{W})$ can be approximated by interpolating the component probabilities of the models using

$$P^H(\mathbf{W}) = \prod_{j=1}^{M} \sum_{\substack{d=1 \\ j \le d}}^{H} \lambda_d \, P_d(w_j) \sum_{\substack{d=1 \\ j > d}}^{H} \lambda_d \, P_d(w_j|w_i) \quad (4)$$

where $\sum_{d=1}^{H} \lambda_d = 1$ and $0 \le \lambda_d \le 1$. The values of λ_d are estimated based on held out data by means of the Expectation Maximization (EM) algorithm [1].

In a second approach, the interpolation is applied on the full language models. That is, $P^H_{int}(\mathbf{W})$ is described by

$$P^H_{int}(\mathbf{W}) = \sum_{d=1}^{H} \xi_d \, P_d(\mathbf{W}) = \sum_{d=1}^{H} \xi_d \left(\prod_{j=1}^{d} P(w_j) \prod_{j=d+1}^{M} P_d(w_j|w_i) \right) \quad (5)$$

where $\sum_{d=1}^{H} \xi_d = 1$ and $0 \le \xi_d \le 1$. The most simple values of ξ_d that could be used are $\xi_d = 1/H$ resulting in a language model averaged over the component language models. In our experiments, we used $\xi_d = \lambda_d$, where λ_d are the weights determined for (4).

III. Absolute Discounting

In discounting models, the relative frequencies of seen events are discounted and the gained probability mass is distributed over the unseen events. To succeed this, *counts* that represent how many times a certain N-gram was found, and *"count-counts"* which represent the number of times a certain count has occurred have to be calculated. Let us denote the "counts-counts" by $n_{r,d}$ where $n_{r,d}$ expresses the total number of distinct joint events that occurred exactly r times at distance d. Events with counts $r = 0, 1$ are characterized as unseen events and singleton ones,respectively and they are the most commonly met. In the following, we briefly describe how absolute discounting can be applied to the proposed language models.

A. Distanced Bigram Language Model

Following similar lines to [16], it can easily be derived that the probabilities required in the model under study can be obtained by

$$P_d(w_j|w_i) = \begin{cases} \frac{N_d(w_i,w_j) - b_d}{N(w_i)} \\ \quad \text{if } N_d(w_i, w_j) > 0 \\ b_d \frac{|V| - n_{0,d}(w_i)}{N(w_i)} \cdot \frac{\beta(w_j|w_i)}{\sum_{w:N_d(w_i,w)=0} \beta(w|w_i)} \\ \quad \text{if } N_d(w_i, w_j) = 0 \end{cases} \quad (6)$$

where b_d is the non-integer count offset for bigrams at distance d, $|V|$ is the total number of words in the vocabulary and $\beta(w_j|w_i)$ is a generalized distribution that serves as a normalization constraint. As in [16], we can derive that

$$b_d = \frac{n_{1,d}}{n_{1,d} + 2n_{2,d}}. \quad (7)$$

Using (7) the absolute discounting model of (6) can be expressed in an interpolated formula as follows:

$$P_d(w_j|w_i) = \max\{0, \frac{N_d(w_i, w_j) - b_d}{N(w_i)}\} + b_d \cdot \frac{|V| - n_{0,d}(w_i)}{N(w_i)} \frac{\beta(w_j|w_i)}{\sum_{w:N_d(w_i,w)=0} \beta(w|w_i)} \quad (8)$$

where $\frac{\beta(w_j|w_i)}{\sum_{w:N(w_i,w)=0} \beta(w|w_i)}$ is usually approximated by $P(w_i)$.

B. Interpolated Distanced N-Gram Model

By substitution of (8) into (5), the absolute discounting model, when an interpolation over the distanced bigram models is wanted, is expressed by

$$P^H_{int}(w_j|w_i) = \sum_{d=1}^{H} \xi_d \left(\max\{0, \frac{N_d(w_i, w_j) - b_d}{N(w_i)}\} + b_d \cdot \frac{|V| - n_{0,d}(w_i)}{N(w_i)} \frac{\beta(w_j|w_i)}{\sum_{w:N_d(w_i,w)=0} \beta(w|w_i)} \right) \quad (9)$$

where $\sum_{d=1}^{H} \xi_d = 1$ and b_d is given by (7).

IV. Automatic Word Clustering

A word clustering algorithm will be described subsequently when an interpolation over the component distanced bigram language models is employed, i.e.,

$$P_{int}^{H}(w_j|w_i) = \sum_{d=1}^{H} \xi_d \, P_d(w_j|w_i). \tag{10}$$

A common assumption in most language models is that classes of functionally equivalent words exist. Given a vocabulary V of size $Q = |V|$, let us assume that L non-overlapping classes C_k, $k = 1, \ldots, L$, can be found such that

$$V = \bigcup_{k=1}^{L} C_k, \qquad C_h \cap C_k = \emptyset \text{ for } h \neq k. \tag{11}$$

The following transition probabilities hold:

$$\forall w_i \in C_h, \forall w_j \in C_k, \quad P_{int}^{H}(w_j|w_i) = P_{int}^{H}(C_k|C_h) \tag{12}$$

where $P_{int}^{H}(C_k|C_h)$ can be estimated using (9).

We statistically characterize the estimate of transition probability $P_{int}^{H}(w_j|w)$ from a given word w to all the other words of the vocabulary, $j = 1, 2, \ldots, Q$, in the same way as estimating the probability of the occurrence of the Q generalized distanced bigrams, in $\sum_{d=1}^{H} N_d(w)$ repeated Bernoulli trials [17]:

$$P(\textstyle\sum_{d=1}^{H} N_d(w, w_1), \ldots, \sum_{d=1}^{H} N_d(w, w_Q)) = \\ \textstyle\sum_{d=1}^{H} \xi_d(N_w)! \; \prod_{j=1}^{Q} \frac{\sum_{d=1}^{H} \xi_d P_d(w,w_j) \sum_{d=1}^{H} N_d(w,w_j)}{(\sum_{d=1}^{H} N_d(w,w_j))!} \tag{13}$$

where $P\left(\sum_{d=1}^{H} N_d(w, w_1), \ldots, \sum_{d=1}^{H} N_d(w, w_Q)\right)$ denotes the probability of having $\sum_{d=1}^{H} N_d(w,\ w_j), j = 1, 2, \ldots, Q$ occurrences of the corresponding distanced bigrams in the training set.

In (13), since $\sum_{d=1}^{H} \xi_d N_w$ is sufficiently large and $\sum_{d=1}^{H} N_d(w, w_j)$ is in the $\sqrt{\sum_{d=1}^{H} \xi_d N(w)}$ neighborhood of $\sum_{d=1}^{H} \xi_d N(w)$, according to De Moivre-Laplace theorem each term in the right-hand side is approximated by a Q-dimensional Gaussian probability density function (pdf) [17]. That is,

$$P(\sum_{d=1}^{H} \xi_d \hat{P}_d(w, w_1), \ldots, \sum_{d=1}^{H} \xi_d \hat{P}_d(w, w_Q)) = N(\boldsymbol{\mu}_w^H, \mathbf{U}_w^H) \tag{14}$$

where the mean vector is given by

$$\boldsymbol{\mu}_w^H = (\sum_{d=1}^{H} \xi_d \hat{P}_d(w_1|w), \ldots, \sum_{d=1}^{H} \xi_d \hat{P}_d(w_Q|w))^T \tag{15}$$

and, to a first degree, the covariance matrix $\mathbf{U}_w^H$ is approximated by a diagonal matrix, i.e.,

$$\mathbf{U}_w^H = \frac{1}{HN(w)} \cdot \\ \text{diag}[\textstyle\sum_{d=1}^{H} \xi_d \hat{P}_d(w_1|w), \ldots, \sum_{d=1}^{H} \xi_d \hat{P}_d(w_Q|w)] \tag{16}$$

where $\text{diag}[\ ]$ denotes the diagonal matrix having the indicated arguments as elements on the main diagonal. Let us assume that every word w_i comprises a single class and it is represented by an estimated transition probability vector $\mathbf{v}_i$ whose elements are the estimated transition probabilities. That is, its kth component is the transition probability from the word w_i to word w_k, $\hat{P}_d(w_k|w_i)$.

The probability of the hypothesis H_{pq} that two classes C_p and C_q form a single class is given by

$$P(H_{pq}) = \prod_{\forall i \to w_i \in C_p \cup C_q} \frac{1}{(2\pi)^{Q/2} (\det(\mathbf{U}_{pq}^H))^{1/2}} \\ \exp\Big\{ -\frac{1}{2} (\mathbf{v}_i - \boldsymbol{\mu}_{pq}^H)^T [\mathbf{U}_{pq}^H]^{-1} (\mathbf{v}_i - \boldsymbol{\mu}_{pq}^H) \Big\} \tag{17}$$

where $\boldsymbol{\mu}_{pq}^H$ and $\mathbf{U}_{pq}^H$ are the mean vector and covariance matrix of the class formed by merging classes C_p and C_q, respectively. In (17), $\det()$ denotes the determinant of a matrix and T is the transposition operator. Classes to be merged correspond to the hypothesis that maximizes (17).

By taking the logarithm of (17) and dropping the normalization term we get the following hypothesis

$$H_{pq}^{*} = \arg\min_{pq} \\ \textstyle\sum_{\forall i \to w_i \in C_p \cup C_q} (\mathbf{v}_i - \boldsymbol{\mu}_{pq}^H)^T [\mathbf{U}_i^H]^{-1} (\mathbf{v}_i - \boldsymbol{\mu}_{pq}^H)\}. \tag{18}$$

An estimate of μ_{pq}^H is given by

$$\boldsymbol{\mu}_{pq}^H = \frac{1}{|C_p| + |C_q|} (\sum_{\forall i \to w_i \in C_p \cup C_q} \mathbf{v}_i) \tag{19}$$

where $|C_p|$ and $|C_q|$ are the numbers of the elements that belong to the corresponding classes. In (18), we assume that the covariance matrix is diagonal and its kkth element is given by

$$[\mathbf{U}_i]_{kk}^H = \frac{1}{\sum_{\forall i \to w_i \in C_p \cup C_q} N(w_i)^2} \\ \cdot \sum_{\forall i \to w_i \in C_p \cup C_q} N(w_i)^2 [U_{w_i}]_{kk}^H. \tag{20}$$

For the mean vector (19) and the covariance matrix (20) we perform robust M-estimation in order to weigh more the observations that come from the typical distribution and to weigh less the observations that come from the contaminating distribution [18], [19].

The criterion (18) is used for determining the best pair of classes that should be merged as in [2]. To summarize, the clustering algorithm works as follows.

- Step 1: Each word of the vocabulary comprises a class on its own. Thus the algorithm starts with Q number of classes.
- Step 2: The two classes that minimize (18) are merged in a single class.
- Step 3: If the number of remaining classes equals a predetermined number of classes L, the algorithm stops. Otherwise a new iteration starts at Step 2.

In the described algorithm, there are approximately $(Q-k)^2/2$ class pairs that have to be examined for merging in each iteration k. In order to avoid the exhaustive computational needs of the algorithm, we sort the words of the vocabulary in decreasing order of frequency and we assign the first $L+1$ words to $L+1$ distinct classes. At each iteration, we try to find the class pair for which the loss in pointwise mutual information is minimal, we perform the merging acquiring L classes and we insert the next word of the vocabulary in a distinct class resulting again in $L+1$ classes. So at iteration k, we assign the $(L+k)$th most probable word of the vocabulary in a distinct class and we continue in the same way until no vocabulary words are left. After $Q-L$ steps the words of the vocabulary are assigned in L classes. Using this approach at iteration k we have to investigate $((L+1)-k)^2/2 < (Q-k)^2/2$ class candidates for merging [2]. If we do not stop the algorithm in L classes, but we continue for $Q-1$ merges, we obtain a single class containing all the vocabulary words. The order in which clusters are merged determines a binary tree having a single cluster as root, the vocabulary words as leaves, and the intermediate clusters as nodes.

V. Results

The following language models were implemented: the classical bigram model, the distanced bigram with $d = 2, \ldots, 5$ and the two versions of the interpolated distanced bigram model described in Section II-C. The models were tested on a subset of the Reuters corpus that yields a vocabulary of 2200 words.

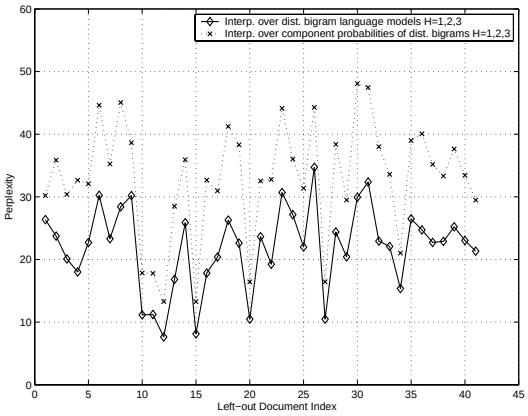

Fig. 1. Perplexity comparison between the interpolated models over (a) the component probabilities of distanced bigram models with $H = 1, 2, 3$ and (b) the distanced language models with $H = 1, 2, 3$.

For the interpolation model (4), EM with a two-way cross validation was used in order to estimate λ_d. For the model (5) the interpolation weights ξ_d were selected to be equal with λ_d. To evaluate the model effectiveness, perplexity was calculated holding out one document from the collection for test document while all the remaining documents are used in the training phase. For the bigram model with distance values $d = 2, \ldots, 5$, the perplexity slightly increased, when the distance increased, as expected [20], [21]. However, the interpolated models with $H = 2, 3$, when either the component probabilities are interpolated or the full language models are interpolated, outperformed the baseline bigram model exhibiting a perplexity reduction of 5% and 20%, respectively. It is worth mentioning that the interpolation over the whole language models gave better perplexity results than that over the component probabilities as depicted in Fig. 1.

Subsequently, the baseline bigram model and the language models derived by interpolating over the distanced bigram language models (with $H = 1, 2$ and $H = 1, 2, 3$) were used for word clustering using the algorithm described in Section IV. The clustering procedure started with a number of classes equal to the size of vocabulary and produced at the end 166 clusters for the first two language models and 171 clusters for the latter one.

Clusters were evaluated using the Jaccard coefficient and the average mutual information. Let $C = \{C_1, C_2, \ldots, C_m\}$ be a clustering structure on the vocabulary V and $C' = \{C'_1, C'_2, \ldots, C'_s\}$ another partition of V. We refer to a pair of words $(w, u) \in C \times C'$ and we define the following terms:

- a: number of word pairs that are in the same cluster in both partitions
- b: number of word pairs that are in the same cluster in C but in different clusters in C'
- c: number of word pairs that are in different clusters in C but in the same cluster in C',and
- d: number of word pairs that are in different clusters in both partitions.

The total number of pairs of objects is $M = a + b + c + d = Q(Q-1)/2$, where Q is the cardinality of V. The Jaccard index is then given by [15]:

$$J = \frac{a}{a+b+c}. \tag{21}$$

J admits values between 0 and 1. Large values of J imply a close agreement between the two partitions. The values of Jaccard index between a randomly generated clustering and the clusterings obtained using procedure described in Section IV for the several language models under study are shown in Table I. As can be seen, the resulting clusterings are not created by chance, because the Jaccard index values are very close to 0.

TABLE I
JACCARD INDEX VALUES.

Language Model	*Jaccard Index*
Classical Bigram	0.003172
Interpolation over the distanced bigram language models for $H = 1, 2$	0.001663
Interpolation over the distanced bigram language models for $H = 1, 2, 3$	0.002934
Interpolation over the distanced bigram language models for $H = 1, 2$ using robust mean estimates	0.001925

The amount of information contained in each clustering and the amount of information the one clustering predicts

about the other were also studied by estimating the entropy of each clustering and the average mutual information between the different clusterings, respectively. Given a clustering C containing K clusters ($C = \{C_1, \ldots, C_K\}$), the entropy of the clustering is

$$H(C) = -\sum_{k=1}^{K} P(k) \log P(k) \tag{22}$$

where $P(k)$ is the probability that a word belongs to the cluster C_k of the clustering C [22]. The average mutual information between two clusterings C and C' given the probabilities $P(k), k = 1, \ldots, K$ for clustering C, $P'(k'), k' = 1, \ldots, K'$ for clustering C', and $P(k, k')$ for the intersection of C and C' ($C \cap C'$) is defined as [22]:

$$\bar{M}(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} \log \frac{P(k, k')}{P(k)P'(k')}. \tag{23}$$

The variation of information $VI(C, C')$, proposed in [23] as a figure of merit for the assessment of clusterings, was also estimated. $VI(C, C')$ is given by

$$VI(C, C') = [H(C) - \bar{M}(C, C')] + [H(C') - \bar{M}(C, C')]. \tag{24}$$

The fist term in the above equation measures how well clustering C can be predicted from C', while the second one how well clustering C' can be predicted from C. The values of terms appearing in (22)-(24) for clusterings obtained using different language models are collected in Table II.

As it can be seen from the Table II, most clusterings predict approximately the same information for the others. However, the higher value of variation of information for the interpolated distanced bigram with $H = 1, 2$ when employing robust mean estimation implies that the latter differs more from the classic bigram than the interpolated distanced bigram with $H = 1, 2$ without the robust estimation for the mean vectors.

VI. Conclusions

In this paper, two language models based on the distance between word pairs have been employed which differ in the way the interpolation between models is implemented. The experimental results indicate the ability of the interpolated models to capture long-term word dependencies with the additional advantage of a low number of parameters in contrast with the classical N-gram. Our first results suggest that a more efficient language model is obtained when the interpolation is made at the level of the entire language models instead of their component probabilities. The more effective interpolated distance bigram model were used for word clustering in an hierarchical clustering approach where robust estimators of the mean vector and the covariance matrix were also employed.

Acknowledgment

This work has been supported by the FP6 European Union Network of Excellence MUSCLE "Multimedia Understanding through Semantics, Computation and LEarning" (FP6-507752).

References

[1] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, Massachussetts, 1998.

[2] P. F. Brown, V. J. Della Pietra, P. V. De Souza, R. L. Mercer, and J. C. Lai, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[3] P. F. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[4] D. Hiemstra, "A linguistically motivated probabilistic model of information retrieval," in *Proc. European Conf. Digital Libraries*, 1998, pp. 569–584.

[5] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval system," in *Proc. SIGIR 98*, New York, 1998, pp. 275–281, ACM.

[6] J. Lafferty and C. Zhai, "Document language models, query models, and risk minimization for information retrieval," in *Proc. SIGIR 01*, New York, 2001, pp. 111–119, ACM.

[7] X. Huang, F. Alleva, H. Hon, M. Hwang, K. Lee, and R. Rosenfeld, "The SPHINX-II speech recognition sytem: An overview," *Computer Speech and Language*, vol. 2, pp. 137–148, 1993.

[8] DARPA, *Proceedings of the Broadcast News Transcription and Understanding Workshop*, Morgan Kaufmann Publishers, Lansdowne, Virginia, 1998.

[9] R. Rosenfeld, *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*, Ph. D. Thesis, School of Computer Science, Pittsburgh, PA, April 1994.

[10] J. R. Bellegarda, "A latent semantic analysis framework for large-span language modeling," in *Proc. Eurospeech'97*, September 1997, vol. 3, pp. 1451–1454.

[11] R. Kuhn and R. De Mori, "A cache-based natural language model for speech recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 6, pp. 570–583, June 1990.

[12] M. Simons, H. Ney, and S. Martin, "Distant bigram language modeling using maximum entropy," in *Proc. IEEE ICASSP 97*, 1997, pp. 787–790.

[13] R. Kneser, "Statistical language modeling using a variable context length," in *Proc. Fourth Int. Conf. Spoken Language Processing*, 1996, vol. 1, pp. 494–497.

[14] L. Bahl, P. Brown, P. De Souza, and R. Mercer, "A tree-based statistical language model for natural speech recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 1001–1008, July 1989.

[15] J. A. Kain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Inc., 1988.

[16] H. Ney, S. Martin, and F. Wessel, "Statistical language modeling using leaving-one-out," in *Corpus-Based Methods in Language and Speech Processing*, S. Young and G. Bloothooft, Eds., pp. 174–207. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[17] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw Hill, New-York, third edition, 1991.

[18] N. A. Campbell, "Robust procedures in multivariate analysis. I: Robust variance estimation," *Applied Statistics*, vol. 29, no. 3, pp. 231–237, 1980.

[19] R. C. Hardie and G. R. Arce, "Ranking in R^p and its use in multivariate image estimation," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 1, no. 2, pp. 197–209, 1991.

[20] J. Goodman, "A bit of progress in language modeling," *Computer Speech and Language*, vol. 15, pp. 403–434, 2001.

[21] J. Goodman, "A bit of progress in language modeling," Tech. Rep. MSR-TR-2001-72, Microsoft, July 2004.

[22] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, John Wiley and Sons, Inc., 2000.

[23] M. Meila, "Comparing clusterings," Tech. Rep. 418, Department of Statistics, University of Washington, October 2002.

TABLE II

ENTROPY, AVERAGE MUTUAL INFORMATION AND VARIATION OF INFORMATION VALUES FOR CLUSTERINGS USING DIFFERENT LANGUAGE MODELS

Language model used for clustering C	*Language model used for clustering C'*	$H(C)$	$H(C')$	$M(C,C')$	$H(C) - M(C,C')$	$H(C') - M(C,C')$	$VI(C,C')$
Classical Bigram (166 Clusters)	Interpolation over the distanced bigram language models for $H = 1,2$ (166 Clusters)	2.207597	2.216013	1.173421	1.034176	1.042593	2.076769
Classical Bigram (166 Clusters)	Interpolation over the distanced bigram language models for $H = 1,2,3$ (166 Clusters)	2.207597	2.216304	1.145027	1.06257	1.071276	2.133846
Classical Bigram (166 Clusters)	Interpolation over the distanced bigram language models for $H = 1,2$ with robust mean estimation (166 Clusters)	2.207597	2.215707	1.170185	1.037412	1.045522	2.082933
Interpolation over the distanced bigram language models for $H = 1,2$ with robust mean estimation (360 Clusters)	Interpolation over the distanced bigram language models for $H = 1,2$ with robust mean and covariance estimation (360 Clusters)	1.310646	2.215707	0.721121	0.589525	1.494586	2.084110