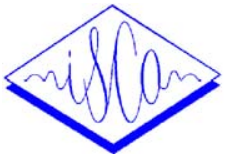

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

CONTEXT SCOPE SELECTION IN MULTI–SPAN STATISTICAL LANGUAGE MODELING

Jerome R. Bellegarda

Spoken Language Group, Apple Computer,
Cupertino, California 95014, USA

ABSTRACT

A multi-span framework was recently proposed to integrate the various constraints, both local and global, that are present in the language. In this approach, local constraints are captured via n-gram language modeling, while global constraints are taken into account through the use of latent semantic analysis. The complementarity between these two paradigms translates into improved modeling performance, as measured by both perplexity and word error rate reduction. This performance improvement is sensitive to the context scope, i.e., the effective length of the document history used in latent semantic analysis during recognition. Context scope selection via exponential forgetting is proposed to discount older utterances as necessary. Experiments on a subset of the Wall Street Journal task led to a reduction in average word error rate of up to 22.5%.

10.21437/Eurospeech.1999-478

1 INTRODUCTION

N-gram language modeling has steadily emerged as the formalism of choice for a wide range of domains. Concerns regarding parameter reliability, however, restrict current implementations to low values of n, which in turn imposes an artificially local horizon to the language model. As a result, n-grams are inherently unable to capture large-span relationships in the language.

Taking more global constraints into account has traditionally involved a paradigm shift toward parsing and rule-based grammars, such as are routinely and successfully employed in small vocabulary recognition applications. This approach solves the locality problem, since it typically operates at the level of an entire sentence. Unfortunately, it is not (yet) practical for large vocabulary recognition.

Among alternative ways to extract suitable long distance information, experiments with word trigger pairs have underscored the desirability of exploiting correlations between the current word and features of the document history [1]. This observation led the author to explore the use of latent semantic analysis (LSA) for such purpose [2]–[5]. In some respect, the LSA paradigm can be viewed as an extension of the trigger concept, where a more systematic framework is used to handle trigger pair selection. In [2], LSA was used for word clustering, and in [3], for language modeling. In both cases, it was found to be suitable to capture some of the global constraints in the language. In fact, multi-span language models, constructed by embedding LSA into the standard n-gram formulation, were shown to result in a substantial reduction in both perplexity [4] and average word error rate [5].

The objective of this paper is to examine how such average word error rate reduction is influenced by the proper selection of the context scope. The paper is organized as follows. In the next section we review the salient properties of n-gram+LSA statistical language modeling. In Section 3, we define context scope and discuss its effect on multi-span performance. Section 4 describes the experimental conditions and illustrates some of the benefits associated with multi-span modeling. Finally, Section 5 analyzes performance variations as the context scope varies.

2 N-GRAM+LSA MODELING

Let $\mathcal{V}$, $|\mathcal{V}| = M$, be some underlying vocabulary and $\mathcal{T}$ a training text corpus, comprising N articles (documents) relevant to some domain of interest, such as, for example, business news. Typically, M and N are on the order of ten thousand and hundred thousand, respectively; $\mathcal{T}$ might comprise a hundred million words or so. The LSA approach defines a mapping between the sets $\mathcal{V}$, $\mathcal{T}$ and a vector space $\mathcal{S}$, whereby each word w_i in $\mathcal{V}$ is represented by a vector u_i in $\mathcal{S}$ and each document d_j in $\mathcal{T}$ is represented by a vector v_j in $\mathcal{S}$. For the sake of brevity, we refer the reader to [6] for further details on the mechanics of LSA and n-gram+LSA language modeling, and just briefly summarize here.

The first step is the construction of a matrix (W) of co-occurrences between words and documents. In marked contrast with n-gram modeling, word order is ignored: the matrix W is accumulated from the available training data by simply keeping track of which word is found in what document. Among other possibilities, a suitable expression for the $(i,j)^{\text{th}}$ element of W is given by (cf. [2]):

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

where $c_{i,j}$ is the number of times w_i occurs in d_j, n_j is the total number of words present in d_j, and ε_i is the normalized entropy of w_i in the corpus $\mathcal{T}$, given by $\varepsilon_i = -(1/\log N)\sum(c_{i,j}/t_i)\log(c_{i,j}/t_i)$, with $t_i = \sum c_{i,j}$.

The second step is to compute the singular value decomposition (SVD) of W as:

$$W \approx \hat{W} = U S V^T, \tag{2}$$

where U is the $(M \times R)$ matrix of left singular vectors u_i $(1 \leq i \leq M)$, S is the $(R \times R)$ diagonal matrix of singular values, V is the $(N \times R)$ matrix of right singular vectors v_j $(1 \leq j \leq N)$, $R \ll M (\ll N)$ is the order of the decomposition, and ^{T} denotes matrix transposition. The left singular vectors represent the words in the given vocabulary, and the right singular vectors represent the documents in the given corpus. Thus, the space $\mathcal{S}$ sought is the one spanned by U and V. An important property of this space is that two words whose representations are "close" (in some suitable metric) tend to appear in the same kind of documents, whether or not they actually occur within identical word contexts in those documents. Conversely, two documents whose representations are "close" tend to convey the same semantic meaning, whether or not they contain the same word constructs. Thus, we can expect that the respective representations of words and documents that are semantically linked would also be "close" in the LSA space $\mathcal{S}$.

The third step is to leverage this property for language modeling purposes. Let w_q denote the word about to be predicted, and H_{q-1} the admissible LSA history (context) for this particular word, i.e., the current document up to word w_{q-1}, denoted by $\tilde{d}_{q-1}$. Then the associated LSA language model probability is given by:

$$\Pr(w_q|H_{q-1},\mathcal{S}) = \Pr(w_q|\tilde{d}_{q-1}), \tag{3}$$

where the conditioning on $\mathcal{S}$ reflects the fact that the probability depends on the particular vector space arising from the SVD representation, and $\tilde{d}_{q-1}$ has a representation in the space $\mathcal{S}$ given by:

$$\tilde{v}_{q-1} = \tilde{d}_{q-1}^T U S^{-1}, \tag{4}$$

through a straightforward extension of (2). The expression (3) is referred to as the direct LSA model. In [4], we have also introduced a number of clustered models with attractive smoothing properties. For instance, if we assume that a set of word clusters C_k, $1 \leq k \leq K$, has been produced in $\mathcal{S}$, then we can expand (3) as:

$$\Pr(w_q|\tilde{d}_{q-1}) = \sum_{k=1}^{K} \Pr(w_q|C_k)\Pr(C_k|\tilde{d}_{q-1}), \tag{5}$$

which is referred to as the word-clustered LSA model. This model has been shown to result in even better performance [4], [6].

Finally, the fourth step is to integrate the above with the conventional n-gram formalism. This integration can occur in a number of ways, such as straightforward interpolation, or within the maximum entropy framework [1]. Alternatively, if we denote by $\bar{H}_{q-1}$ the overall available history (comprising an n-gram component as well as the LSA component mentioned above), then a suitable expression for the integrated probability is given by [6]:

$$\Pr(w_q|\bar{H}_{q-1}) = \frac{\Pr(w_q|w_{q-1}w_{q-2}\ldots w_{q-n+1})\Pr(\tilde{d}_{q-1}|w_q)}{\sum_{w_i \in \mathcal{V}} \Pr(w_i|w_{q-1}w_{q-2}\ldots w_{q-n+1})\Pr(\tilde{d}_{q-1}|w_i)}. \tag{6}$$

Note that, if $\Pr(\tilde{d}_{q-1}|w_q)$ is viewed as a prior probability on the current document history, then (6) simply translates the classical Bayesian estimation of the n-gram (local) probability using a prior distribution obtained from (global) LSA. The end result, in effect, is a modified n-gram language model incorporating large-span semantic information.

3 DEFINITION OF CONTEXT SCOPE

In practice, expressions like (6) are often slightly modified so that a relative weight can be placed on each contribution (here, the n-gram and LSA probabilities). Usually, this is done via empirically determined weighting coefficients. In the present case, such weighting is motivated by the fact that the "prior" probability $\Pr(\tilde{d}_{q-1}|w_q)$ could change substantially as the current document unfolds. Thus, rather than using arbitrary weights, an alternative solution is to dynamically tailor the document history $\tilde{d}_{q-1}$ so that the n-gram and LSA contributions remain empirically balanced. We refer to this procedure as context scope selection.

During training, the context scope is fixed to be the current document. During recognition, however, the concept of "current document" is ill-defined, because (i) its length grows with each new word, and (ii) it is not necessarily clear at which point completion occurs. As a result, a decision has to be made regarding what to consider "current," versus what to consider part of an earlier (presumably less relevant) document. This is especially important when the user utters several "mini-documents" within the same session, as might very well do the average user of a dictation system.

The simplest solution is to postulate that all utterances spoken since the beginning of the session are part of the current document. This is adequate only if the user starts a new session each time s/he wants to work on a new document. (This was the scenario implicitly assumed in [4] and [6].) If, however, the user needs to dictate in an heterogeneous manner, this solution might fail, because the (single, cumulative) "current" document built under this assumption

might not be sufficiently representative of each individual topic.
An alternative solution is to limit the size of the history considered, so as to avoid relying on old, possibly obsolete fragments to construct the current context. The size limit could be expressed in anything from words to paragraphs. The problem here is the difficulty of determining this size limit *a priori*, since it is highly dependent on the kind of documents spoken by the user.
Thus, we will adopt a intermediate solution, which does not require a hard decision to be made on the size of the caching window. This solution uses exponential forgetting to progressively discount older utterances. Assuming $0 < \lambda \leq 1$, this approach corresponds to the following expression for (4):

$$\tilde{v}_q = \frac{1}{n_q} \sum_{p=1}^{q} \lambda^{(n_q - n_p)} \, (1 - \varepsilon_{i_p}) \, u_{i_p} \, S^{-1} \,, \qquad (7)$$

where n_q is the total number of words present in the current document as of time q, and i_p is the index of the word observed at time p. Note that the gap between λ and 1 tracks the expected heterogeneity of the session.

4 EXPERIMENTAL CONDITIONS

Following [6], we have trained the LSA framework on the WSJ0 part of the ARPA North American Business (NAB) News corpus. This was convenient for comparison purposes since conventional n-gram language models are readily available, trained on exactly the same data [7]. The training text corpus $\mathcal{T}$ was composed of about $N = 87,000$ documents spanning the years 1987 to 1989, comprising approximately 42 million words. The vocabulary $\mathcal{V}$ was constructed by taking the 20,000 most frequent words of the NAB corpus, augmented by some words from an earlier release of the Wall Street Journal corpus, for a total of $M = 23,000$ words.
We performed the singular value decomposition of the matrix of co-occurrences between words and documents using the single vector Lanczos method [8]. We experimented with different numbers of singular values retained, and found that $R = 125$ seemed to achieve an adequate balance between reconstruction error (as measured by Frobenius norm differences) and noise suppression (as measured by trace ratios). Using the resulting vector space $\mathcal{S}$ of dimension 125, we constructed a direct LSA model (3), and a word-clustered LSA model (5). Each was then combined with the standard bigram, as in (6).
The resulting multi-span language models, dubbed bi-LSA models, were used in lieu of the standard WSJ0 bigram model in a series of speaker-independent, continuous speech recognition experiments. These experiments were conducted on a subset of the Wall Street Journal 20,000 word-vocabulary task. The acoustic training corpus consisted of 7,200 sentences of data uttered by 84 different native speakers of English (WSJ0 SI-84). The test corpus consisted of 496 sentences uttered by 12 additional native speakers of English.

Speaker	Word Error Rate Reduction, Direct Model	Word Error Rate Reduction, Clustered Model
001	8.4 %	11.2 %
002	21.5 %	35.0 %
00a	17.5 %	25.9 %
00b	10.1 %	7.8 %
00c	10.0 %	17.6 %
00d	17.3 %	35.4 %
00f	11.5 %	16.9 %
203	16.1 %	34.2 %
400	14.8 %	19.8 %
430	19.3 %	20.2 %
431	12.2 %	18.3 %
432	7.8 %	27.9 %
Overall	13.7 %	22.5 %

Table 1. Performance Improvement Using Bi-LSA Language Modeling.

Table 1 summarizes the reduction in word error rate achieved using the bi-LSA language models, as compared to the performance of the baseline bigram. In the first column, the bi-LSA model incorporates the direct model (3), while in the second column, it incorporates the word-clustered model (5) with a word cluster set of size $K = 100$. It can be seen that all speakers substantially benefit from multi-span modeling, displaying a reduction in error rate ranging from about 8% to more than 35%.

5 INFLUENCE OF CONTEXT SCOPE

It is important to note that the task chosen represents a severe test of the LSA paradigm. By design, the test corpus was constructed with no more than 3 or 4 consecutive sentences extracted from a single article. Overall, it comprises 140 distinct document fragments, i.e., each speaker speaks, on the average, about 12 different "mini-documents." As a result, the context effectively changes every 60 words or so, which prevents the multi-span model from building a very accurate pseudo-document representation. This is a situation where it is critical for the multi-span model to appropriately forget the context as it unfolds, to avoid relying on an obsolete representation. In particular, the results of Table 1 were obtained using exponential forgetting (7) with $\lambda = 0.975$. For the sake of illustration, this means that the word which occurred 60 words ago is discounted through a weight of about 0.2.

Speaker	$\lambda = 1.0$	$\lambda = 0.99$	$\lambda = 0.98$	$\lambda = 0.97$	$\lambda = 0.96$	$\lambda = 0.95$
001	7.7 %	11.9 %	11.2 %	4.9 %	−2.1 %	−3.5 %
002	27.7 %	33.3 %	33.9 %	35.0 %	37.9 %	36.2 %
00a	15.7 %	25.2 %	21.2 %	25.9 %	23.0 %	20.8 %
00b	8.2 %	9.7 %	7.8 %	9.7 %	7.8 %	7.8 %
00c	10.3 %	12.9 %	17.6 %	16.5 %	16.5 %	16.2 %
00d	16.1 %	27.8 %	33.6 %	35.4 %	39.2 %	33.0 %
00f	10.7 %	11.1 %	15.3 %	16.9 %	16.5 %	16.9 %
203	15.4 %	21.5 %	32.2 %	34.2 %	33.6 %	28.9 %
400	15.9 %	17.0 %	18.1 %	19.8 %	19.2 %	16.5 %
430	12.6 %	19.3 %	20.2 %	17.6 %	14.3 %	10.9 %
431	8.9 %	15.0 %	18.3 %	18.3 %	17.8 %	13.6 %
432	11.2 %	16.2 %	23.5 %	27.9 %	27.9 %	26.3 %
Overall	13.2 %	18.4 %	21.1 %	21.9 %	21.6 %	19.3 %

Table 2. Influence of Context Scope Selection on Word Error Rate Reduction.

One way to measure the influence of context scope selection is to vary the value of the parameter λ. Table 2 presents recognition results for values ranging from $\lambda = 1$ to $\lambda = 0.95$, in decrements of 0.01. In all cases we used the same clustered model as above, so the results are to be compared to the right-most column of Table 1. As predicted, the performance with no forgetting (13.2% average reduction) is less compelling than that observed in Table 1 (22.5% average reduction). This is consistent with the characteristics of the task, and underscores the role of discounting as a suitable counterbalance to frequent context changes.

Performance rapidly improves as λ decreases from $\lambda = 1$ to $\lambda = 0.97$, presumably because the pseudo-document representation becomes less and less contaminated with obsolete data. If forgetting gets too aggressive, however, the performance starts degrading, as the effective context no longer has an equivalent length which is sufficient for the task at hand. Here, this happens for $\lambda < 0.97$. Note that this degradation is more or less severe depending on the actual article fragments uttered. For example, speaker 00b seems to be considerably less affected than, say, speaker 001.

6 CONCLUSION

We have investigated the behavior of multi-span language models, constructed by embedding latent semantic analysis into the standard n-gram formulation, in actual recognition experiments. When compared to the associated standard n-gram on a subset of the Wall Street Journal large vocabulary task, the multi-span approach resulted in a reduction in average word error rate of about 14% for the direct model and 22.5% for the word-clustered model.

As this experimental task features marked document fragmentation, we have also studied the influence of dynamic context scope selection on multi-span performance. We found that discounting obsolete data via exponential forgetting can make a substantial difference when several "mini-documents" are uttered in quick succession. This is likely to have practical implications in product implementations incorporating multi-span language modeling.

REFERENCES

[1] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer Speech and Language*, Vol. 10, London: Academic Press, pp. 187–228, July 1996.

[2] J.R. Bellegarda *et al.*, "A Novel Word Clustering Algorithm Based on Latent Semantic Analysis," in *Proc. ICASSP*'96, pp. I172–I175.

[3] J.R. Bellegarda, "A Latent Semantic Analysis Framework for Large-Span Language Modeling," in *Proc. EuroSpeech*'97, Vol. 3, pp. 1451–1454.

[4] J.R. Bellegarda, "Exploiting Both Local and Global Constraints for Multi-Span Statistical Language Modeling," in *Proc. ICASSP*'98, Vol. 2, pp. 677–680.

[5] J.R. Bellegarda, "Speech Recognition Experiments Using Multi-Span Statistical Language Modeling," in *Proc. ICASSP*'99, Vol. II, pp. 717–720.

[6] J.R. Bellegarda, "A Multi-Span Language Modeling Framework for Large Vocabulary Speech Recognition," *IEEE Trans. Speech Audio Proc.*, Vol. 6, No. 5, pp. 456–467, September 1998.

[7] F. Kubala *et al.*, "The Hub and Spoke Paradigm for CSR Evaluation", in *Proc.* ARPA *Speech and Natural Language Workshop*, Morgan Kaufmann, pp. 40–44, March 1994.

[8] M.W. Berry, "Large–Scale Sparse Singular Value Computations," *Int. J. Supercomp. Appl.*, Vol. 6, No. 1, pp. 13–49, 1992.