SPEECH RECOGNITION EXPERIMENTS USING MULTI-SPAN STATISTICAL LANGUAGE MODELS

Jerome R. Bellegarda

Spoken Language Group
Apple Computer, Inc.
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 performance of the resulting multi-span language models, as measured by perplexity, has been shown to compare favorably with the corresponding n-gram performance. This paper reports on actual speech recognition experiments, and shows that word error rate is also substantially reduced. On a subset of the Wall Street Journal speaker-independent, 20,000-word vocabulary, continuous speech task, the multi-span framework resulted in a reduction in average word error rate of up to 17%.

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. This has motivated further investigation into alternative ways to extract suitable long distance information, other than resorting to a formal parsing mechanism.

One such attempt was based on the concept of word triggers [1]. Unfortunately, trigger pair selection is a complex issue: different pairs display markedly different behavior, which limits the potential of low frequency triggers [2]. Still, self-triggers seem to be particularly powerful and robust [1], which underscores the desirability of exploiting correlations between the current word and features of the document history.

This observation led the author to explore the use of latent semantic analysis (LSA) for such purpose [3] – [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 [3], LSA was used for word clustering, and in [4], 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 perplexity [5].

In this paper, we are primarily interested in the behavior of such multi-span language modeling in actual recognition. 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 address some of the implementation issues involved in using the resulting multi-span models for large vocabulary recognition. Section 4 illustrates some of the benefits associated with multi-span modeling on a subset of the Wall Street Journal task. Finally, Section 5 analyzes the influence of the data selected to train the LSA component of the multi-span model.

2. N-GRAM+LSA MODELING

Let $\mathcal{V}$, $|\mathcal{V}| = M$, be some vocabulary of interest and $\mathcal{T}$ a training text corpus, comprising N articles (documents) from a variety of sources. (Note that this implies that the training data is tagged at the document level, i.e., there is a way to identify article boundaries. This is the case, for example, with the ARPA North American Business (NAB) News corpus [6].) 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 [7] 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. [3]):

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

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 e_i is the normalized entropy of w_i in the corpus $\mathcal{T}$, given by $e_i = -\,(1/\log N)\sum (c_{i,j}/t_i)\,\log(c_{i,j}/t_i)$, 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 $(1 \le i \le M)$, S is the $(R \times R)$ diagonal matrix of singular values, V is the $(N \times R)$ matrix of right singular vectors v_j $(1 \le j \le N)$, $R \ll M (\ll N)$ 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). This vector representation for $\tilde{d}_{q-1}$ is adequate under some consistency conditions on the general patterns present in the domain considered; see [7] for a complete discussion.

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 [2]. 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 [7]:

$$\Pr\,(w_q|\bar{H}_{q-1}) = \frac{\Pr\,(w_q|w_{q-1}w_{q-2}\ldots w_{q-n+1})\,\Pr\,(\tilde{d}_{q-1}|w_q)}{\displaystyle\sum_{w_i\in\mathcal{V}} \Pr\,(w_i|w_{q-1}w_{q-2}\ldots w_{q-n+1})\,\Pr\,(\tilde{d}_{q-1}|w_i)}\,. \tag{5}$$

Note that, if $\Pr\,(\tilde{d}_{q-1}|w_q)$ is viewed as a prior probability on the current document history, then (5) 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. IMPLEMENTATION ISSUES

There are two ways to take advantage of multi-span modeling for large vocabulary speech recognition. One is to rescore previously produced N-best lists using the integrated models. (This was the scenario implicitly assumed in [5] and [7].) The other is to use the multi-span models directly in the search itself. The latter is preferable, since it allows incremental pruning based on the best knowledge source available.

Compared to N-best rescoring, however, using multi-span modeling directly in the search entails a much higher computational cost. Of particular concern is the calculation of each pseudo-document vector representation in (4), as well as the computation of the integrated probability (5), both of which require $\mathcal{O}(MR)$ floating point operations. The latter can be classically alleviated through appropriate thresholding and caching of the LSA probabilities. But what about the former?

As it turns out, it can be reduced by exploiting the sequential nature of pseudo-documents. Clearly, as each word context is expanded, the document context remains largely unchanged, with only the most recent candidate word added. Assume further that the training corpus $\mathcal{T}$ is large enough, so that the normalized entropy e_i $(1 \le i \le M)$ does not change appreciably with the addition of each pseudo-document. Then it is possible to express the new pseudo-document vector directly in terms of the old pseudo-document vector, instead of each time re-computing the entire mapping from scratch.

To see that, consider $\tilde{d}_q$, and assume, without loss of generality, that word w_i is observed at time q. Then, from (1), we will have, for $k = i$:

$$w_{i,q} = (1 - e_i)\,\frac{c_{i,q-1} + 1}{n_q} = \frac{n_q - 1}{n_q}\,w_{i,q-1} \;+\; \frac{1 - e_i}{n_q}\,, \tag{6}$$

while, for $1 \le k \le M$, $k \ne i$:

$$w_{k,q} = w_{k,q-1}\,. \tag{7}$$

Hence, with the shorthand notation $g_{i,q} = (1 - e_i)/n_q$, we can express $\tilde{d}_q$ as:

$$\tilde{d}_q = \frac{n_q - 1}{n_q}\,\tilde{d}_{q-1} \;+\; [0\ldots g_{i,q}\ldots 0]^{\,T}\,, \tag{8}$$

which is turn implies, from (4):

$$\tilde{v}_q = \frac{n_q - 1}{n_q}\,\tilde{v}_{q-1} \;+\; g_{i,q}\,u_i\,S^{-1}\,. \qquad (9)$$

It is easily verified that (9) requires only $\mathcal{O}(R)$ floating point operations. Thus, we can update the pseudo-document vector directly in the LSA space at a fraction of the cost previously required to map the sparse representation to the space $\mathcal{S}$. This allows multi-span language modeling to be taken advantage of in early stages of the search.

4. RECOGNITION RESULTS

Following [7], we have trained the LSA framework on the WSJ0 part of the NAB News corpus. This was convenient for comparison purposes since conventional n-gram language models are readily available, trained on exactly the same data [6]. 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 News 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]. Over the course of this decomposition, 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 the LSA model (3) and combined it with the standard bigram, as in (5).

The resulting multi-span language model, dubbed bi-LSA model, was then 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.

It is important to note that the task chosen represents a severe test of the LSA component implemented above. 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, which means that 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. (In situations like these, it is beneficial to implement a mechanism to consistently forget the context, to avoid relying on an obsolete representation; details will be presented in [9].)

Speaker	Reduction in Perplexity	Reduction in Word Error Rate
001	22.8 %	8.4 %
002	28.5 %	21.5 %
00a	30.6 %	17.5 %
00b	27.4 %	10.1 %
00c	33.6 %	10.0 %
00d	26.2 %	17.3 %
00f	33.3 %	11.5 %
203	35.3 %	16.1 %
400	15.4 %	14.8 %
430	19.7 %	19.3 %
431	20.0 %	12.2 %
432	24.7 %	7.8 %
Overall	24.7 %	13.7 %

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

Table 1 summarizes the performance achieved using the bi-LSA language model, as compared with that achieved using the baseline bigram. The comparison is made in terms of both reduction in test data perplexity (first column) and reduction in actual word error rate (second column). It can be seen that all speakers substantially benefit from multi-span modeling. Overall, we observed a reduction in perplexity of about 25%, and a reduction in average error rate on the order of 15%.

As usual, the reduction in average error rate is less than the corresponding reduction in perplexity, due to the influence of the acoustic component in actual recognition, and the resulting "ripple effect" of each recognition error. Note that in the case of n-LSA language modeling, this effect can be expected to be more pronounced than in the standard n-gram case. This is because recognition errors are potentially able to affect the LSA context well into the future, through the estimation of a flawed representation of the pseudo-document in the LSA space. This lingering behavior, which can obviously degrade the effectiveness of the LSA component, is an unfortunate by-product of large-span modeling. Clearly, the more accurate the recognition system, the less problematic this unsupervised context construction becomes.

5. INFLUENCE OF LSA TRAINING

In the above, the LSA component of the multi-span language model was trained on exactly the same data as its n-gram component. This is not a requirement, however, which raises the question of how critical the selection of the LSA training data is to the performance of the recognizer. This is particularly interesting since LSA is known to be weaker on heterogeneous corpora (cf., e.g., [10]).

To ascertain the matter, we left the bigram component unchanged, and repeated the LSA training on non-Wall

Speaker	AP 84 K Docs	AP 155 K Docs	AP 224 K Docs	WSJ Test Docs
001	0.0 %	6.3 %	7.0 %	13.3 %
002	0.0 %	4.0 %	5.1 %	28.2 %
00a	8.4 %	9.5 %	11.3 %	21.5 %
00b	-3.1 %	-3.1 %	-3.1 %	15.2 %
00c	2.1 %	2.0 %	2.4 %	14.1 %
00d	2.6 %	2.4 %	2.9 %	15.5 %
00f	2.7 %	2.7 %	3.8 %	18.0 %
203	3.4 %	3.1 %	4.7 %	17.4 %
400	7.1 %	7.3 %	7.1 %	14.8 %
430	5.0 %	3.4 %	0.0 %	23.5 %
431	-0.5 %	4.2 %	3.3 %	17.4 %
432	1.7 %	2.2 %	4.5 %	10.6 %
Overall	2.4 %	3.3 %	4.0 %	17.1 %

Table 2. Multi-Span Sensitivity to LSA Training for Bi-LSA Language Modeling.

Street Journal data from the same general period, using the same underlying vocabulary $\mathcal{V}$. Three corpora of increasing size were considered, all corresponding to Associated Press (AP) data: (i) $\mathcal{T}_1$, composed of $N_1 = 84,000$ documents from 1989, comprising approximately 44 million words; (ii) $\mathcal{T}_2$, composed of $N_2 = 155,000$ documents from 1988 and 1989, comprising approximately 80 million words; and (iii) $\mathcal{T}_3$, composed of $N_3 = 224,000$ documents from 1988-1990, comprising approximately 117 million words. In each case we proceeded with the LSA training as described in Section 2. The resulting word error rate reductions are reported in Table 2 in the three columns labelled "AP."

Two things are immediately apparent. First, the performance improvement in all cases is much smaller than in Table 1, which seems to underscore the sensitivity of the LSA framework to the domain considered. And second, the overall performance does not improve appreciably with more training data, a fact already observed in [7] using a perplexity measure. This bodes well for rapid adaptation to cross-domain data, provided a suitable adaptation framework can be derived.

To establish an upper bound on multi-span performance, we then went the other way and re-trained the LSA parameters on just the test set. This time the corpus $\mathcal{T}_4$ was composed of only $N_4 = 140$ documents, comprising approximately 8500 words, which effectively reduced the vocabulary $\mathcal{V}$ to about 2500 words. The resulting error rate reductions are presented in the right-most column of Table 2.

Again, two points can be made. First, the overall performance improvement is only marginally better than that observed in Table 1, suggesting that within-domain adaptation may not generally be compelling. And second, for this task, 17% is the maximum that can be gained by applying LSA constraints. Note, however, that this improvement may not be indicative of the best possible achievable with the multi-span language model, due again to the atypical document fragmentation existing in the test data.

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 perplexity of about 25%, and a reduction in average error rate of about 15%.

We have also looked at the influence of the LSA training data on performance improvement. The multi-span approach showed much more sensitivity to the training domain than to the size of the training data. These results suggest that cross-domain adaptation has greater potential than within-domain adaptation for adaptive multi-span language modeling.

7. REFERENCES

[1] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger–Based Language Models: A Maximum Entropy Approach," in *Proc.* 1993 *Int. Conf. Acoust., Speech, Sig. Proc.*, Minneapolis, MN, pp. II45–48, March 1993.

[2] 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.

[3] J.R. Bellegarda *et al.*, "A Novel Word Clustering Algorithm Based on Latent Semantic Analysis," in *Proc.* 1996 *Int. Conf. Acoust., Speech, Sig. Proc.*, Atlanta, GA, pp. I172–I175, May 1996.

[4] J.R. Bellegarda, "A Latent Semantic Analysis Framework for Large-Span Language Modeling," in *Proc. EuroSpeech*'97, Rhodes, Greece, Vol. 3, pp. 1451–1454, September 1997.

[5] J.R. Bellegarda, "Exploiting Both Local and Global Constraints for Multi-Span Statistical Language Modeling," in *Proc.* 1998 *Int. Conf. Acoust., Speech, Sig. Proc.*, Seattle, WA, Vol. 2, pp. 677–680, May 1998.

[6] 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.

[7] 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.

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

[9] J.R. Bellegarda, "Large Vocabulary Speech Recognition With Multi-Span Statistical Language Models," *IEEE Trans. Speech Audio Proc.*, in preparation.

[10] Y. Gotoh and S. Renals, "Document Space Models Using Latent Semantic Analysis," in *Proc. EuroSpeech*'97, Rhodes, Greece, Vol. 3, pp. 1443–1448, September 1997.