A Latent Semantic Analysis Framework for Large–Span Language Modeling

Jerome R. Bellegarda

Advanced Technology Group, Apple Computer,
Cupertino, California 95014, USA

jerome @ apple.com; +1 (408) 974-7647

ABSTRACT

A new framework is proposed to construct large-span, semantically-derived language models for large vocabulary speech recognition. It is based on the *latent semantic analysis* paradigm, which seeks to automatically uncover the salient semantic relationships between words and documents in a given corpus. Because of its semantic nature, a latent semantic language model is well suited to complement a conventional, more syntactically-oriented n-gram. An integrative formulation is proposed for the combination of the two paradigms. The performance of the resulting integrated language model, as measured by perplexity, compares favorably with the corresponding n-gram performance.

1 INTRODUCTION

Stochastic language modeling plays a central role in large vocabulary speech recognition, where it is usually implemented using the n-gram paradigm. In a typical application, the purpose of an n-gram language model may be to constrain the acoustic analysis, guide the search through various (partial) text hypotheses, and/or contribute to the determination of the final transcription [1]. Success in these endeavors depends on the ability of the language model to suitably discriminate between different strings of n words. This ability is in turn critically influenced by the choice of n, the trade-off being between weak predictive power (low n) and unreliable estimation (higher n).

Many approaches have been developed to improve the robustness of the estimation, see, e.g., [2] – [5]. Still, it remains extremely challenging to go beyond, say $n \leq 4$, with currently available databases and processing power [4]. This imposes an artificially local horizon to the language model and thereby limits its predictive power. Consider, for instance, predicting the word "*fell*" from the word "*stocks*" in the two equivalent phrases:

$$\textit{stocks fell sharply as a result of the call}, \qquad (1)$$

$$\textit{stocks, as a result of the call, sharply fell}. \qquad (2)$$

While a bigram would do fine in (1), a 9-gram language model would be necessary for (2), a rather unrealistic proposition at the present time.

A possible solution might be to take the entire sentence into account. This requires a paradigm shift toward parsing and rule-based grammars, such as are routinely and successfully employed in small vocabulary recognition applications. This solution, unfortunately, is not (yet) practical for large vocabulary recognition, which is precisely the reason why the n-gram framework was so widely adopted in the first place.

What seems to be needed is an intermediate approach, where the effective context is expanded from 3 or 4 words to a larger span, say an entire sentence or even a whole document, without resorting to a formal parsing mechanism. This in turn would allow for the extraction of suitable long distance information.

One approach recently proposed in that direction is based on the concept of word triggers [6]. In the above example, suppose that the training data reveals a significant correlation between "*stocks*" and "*fell*," so that the pair (*stocks*, *fell*) forms a trigger pair. Then the presence of "*stocks*" in the document could automatically trigger "*fell*," causing its probability estimate to change. Because this behavior would occur indifferently in (1) and in (2), the two phrases would be lead to the same result.

Unfortunately, trigger pair selection is a complex issue: different pairs display markedly different behavior, which limits the potential of low frequency word triggers [7]. Still, self-triggers have been shown to be particularly powerful and robust [6], which underscores the desirability of exploiting correlations between the current word and features of the document history.

This paper proposes a different approach along the same lines, based on a paradigm originally formulated in the context of information retrieval, called latent semantic analysis [8]. In some respect, this approach can be viewed as an extension of the word trigger concept, where a more systematic framework is used to handle the trigger pair selection.

The paper is organized as follows. In the next section we review the salient properties of latent semantic analysis. In Section 3, we use this framework to derive a large-span semantic language model and discuss its predictive power. Section 4 addresses the integration of the new framework with conventional n-gram language models. Finally, in Section 5 a series of experimental results illustrates some of the benefits associated with the integrated language model.

2 LATENT SEMANTIC ANALYSIS

The problem is to relate to one another those words which are found to be semantically linked from the evidence presented in the training text database, without regard to the particular syntax used to express that semantic link. Clearly, the trigger approach mentioned earlier provides a solution for those trigger pairs that have been selected by the algorithm [7].

However, trigger pair selection entails a number of practical constraints. First, only word pairs which co-occur in a sufficient number of documents are considered. In addition, a mutual information criterion is typically used to further confine the list of candidate pairs to a manageable size. This may result in too much "filtering" of the data. We would like to use a somewhat more flexible framework to exploit the long distance information present in the history.

This is where the latent semantic paradigm comes into play. In latent semantic indexing [8], co-occurrence analysis takes place across much larger spans than with a traditional n-gram approach (i.e., spans of 3 words as in [3]), and on a much larger scale than with the trigger approach (i.e., about 1.4 million trigger pairs as in [7]). The span of choice is a *document*, which can be defined as a semantically homogeneous set of sentences embodying a given storyline. As for scale, every combination of words from the vocabulary is viewed as a potential trigger combination. An important benefit of this generalization is the systematic integration of long-term dependencies into the analysis.

To take advantage of the concept of document, we of course have to assume that the available 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 [9]. This assumption enables the construction of a matrix of co-occurences between words and documents. This matrix is accumulated from the available training data by simply keeping track of which word is found in what document.

Note that, in marked contrast with n-gram modeling, word order is ignored. This means that the latent semantic paradigm not only does not exploit syntactic information, but effectively throws it away. Thus, it should not be expected to replace conventional n-grams, but rather to complement them.

We have recently used the latent semantic analysis (LSA) approach to derive a novel word clustering algorithm [10]. For the sake of brevity, we refer the reader to [10] for further details on the mechanics of LSA, and just briefly summarize here. After the word-document matrix of co-occurences is constructed, LSA proceeds by computing the singular value decomposition (SVD) of the word-document matrix. The left singular vectors in this SVD represent the words in the given vocabulary, and the right singular vectors represent the documents in the given corpus. Thus, the role of the SVD is to establish a one-to-one mapping between words/documents and some vectors in a space of appropriate dimension. Specifically, this space is spanned by the singular vectors resulting from the SVD.

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 that space. This property is what makes the framework useful for language modeling purposes.

3 LANGUAGE MODELING

Let $\mathcal{V}$, $|\mathcal{V}| = M$, be some vocabulary of interest and $\mathcal{T}$ a training text corpus, i.e., a collection of N articles (documents) from a variety of sources. (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.) As described in [10], 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 $\mathcal{S}$.

Recall that the matrix of co-occurences embodies structural associations between words and documents. Thus, the extent to which word w_i and document d_j co-occur in the training corpus can be inferred from the (i,j) cell of the matrix. From the SVD formalism, it follows that this can be characterized by taking the dot product between the ith row of the matrix $US^{1/2}$ and the jth row of the matrix $VS^{1/2}$, namely $u_iS^{1/2}$ and $v_jS^{1/2}$. In other words, how "close" u_i is to v_j in the space $\mathcal{S}$ can be characterized by the dot product between $u_iS^{1/2}$ and $v_jS^{1/2}$ [10].

Let w_q denote the word about to be predicted, H_{q-1} the admissible history (context) for this particular word, and $\Pr(w_q|H_{q-1})$ the associated language model probability. In the case of an n-gram language model, for example, $\Pr(w_q|H_{q-1}) = \Pr(w_q|w_{q-1}w_{q-2}\dots w_{q-n+1})$. To take the LSA framework into account, we have to consider the slightly modified expression:

$$\Pr(w_q|H_{q-1}) = \Pr(w_q|H_{q-1}, \mathcal{S}), \qquad (3)$$

where the conditioning on $\mathcal{S}$ reflects the fact that in the proposed derivation the probability depends on the particular vector space arising from the SVD representation.

As usual, the quality of the resulting language model can be measured by the perplexity of (3) on some test text. If Q denotes the total number of words in the test text, this measure is given by:

$$PP = \exp\big(-\frac{1}{Q}\sum_{q=1}^{Q}\log \Pr(w_q|H_{q-1},\mathcal{S})\big)\,. \tag{4}$$

Thus, to construct a semantic language model, there are two issues that we have to address: (i) specify what H_{q-1} is in the case of LSA; and (ii) find a way to compute (3).

Since the SVD operates on a matrix of co-occurrences between words and documents, the nominal history is the document in which w_q appears. However, to be admissible, the context must be causal, and therefore be truncated at word w_{q-1}. Thus, in practice, we have to define H_{q-1} to be the current document up to word w_{q-1}.[1]

Obviously, the current document will not (normally) have been seen in $\mathcal{T}$, therefore qualifying as a pseudo-document in the terminology of [10]. If we denote this pseudo-document by $H_{q-1} = \tilde{d}_{q-1}$, then it is possible (cf. [10]) to derive a vector representation $\tilde{v}_{q-1} \in \mathcal{S}$ associated with this pseudo-document. The language model thus becomes:

$$\Pr(w_q|H_{q-1},\mathcal{S}) = \Pr(w_q|\tilde{d}_{q-1})\,, \tag{5}$$

where $\Pr(w_q|\tilde{d}_{q-1})$ is computed directly from the representations of w_q and $\tilde{d}_{q-1}$ in the space S. In other words, this expression can be directly inferred from the "closeness" between u_q and $\tilde{v}_{q-1}$ in $\mathcal{S}$.

From the discussion above, a natural metric to consider for this "closeness" is the cosine of the angle between $u_q S^{1/2}$ and $\tilde{v}_{q-1}S^{1/2}$. Thus:

$$K(u_q,\tilde{v}_{q-1}) = \frac{u_q\, S\, \tilde{v}_{q-1}^T}{\|u_q S^{1/2}\|\,\|\tilde{v}_{q-1}S^{1/2}\|}\,, \tag{6}$$

for any q indexing a word in the text data. A value of $K(u_q,\tilde{v}_{q-1}) = 1$ means that $\tilde{d}_{q-1}$ is a strong semantic predictor of w_q, while a value of $K(u_q,\tilde{v}_{q-1}) < 1$ means that the history carries increasingly less information about the current word. It is straightforward to modify (6) to define a *bona fide* distance in the space $\mathcal{S}$. For a given history (document $\tilde{d}_{q-1}$), this distance then induces an empirical multivariate distribution in the space $\mathcal{S}$, which in turn allows for the computation of $\Pr(w_q|\tilde{d}_{q-1})$.

Note that $\Pr(w_q|\tilde{d}_{q-1})$ reflects the "relevance" of word w_q to the admissible history, as observed through $\tilde{d}_{q-1}$. As such, it will be highest for words whose meaning aligns most closely with the semantic fabric of $\tilde{d}_{q-1}$ (i.e., relevant "content" words), and lowest for words which do not convey any particular information about this fabric (e.g., "function" words like *the*).

Since content words tend to be rare and function words tend to be frequent, this will translate into a relatively high value for (4). Thus, even though this model appears to have the same order as a standard unigram, it will likely exhibit a significantly weaker predictive power.

4 INTEGRATION WITH N-GRAMS

On the other hand, since the language model (5) does not exploit positional information at all, it should not be expected to replace n-grams in terms of predictive power. A more desirable strategy might be to combine global constraints such as provided by LSA with local constraints such as provided by the n-gram paradigm. This amounts to leveraging both syntactic and semantic information to derive an integrated language model with the benefits of both.

This integration can occur in a number of ways, such as straightforward interpolation, or within the maximum entropy framework [7]. In the following, we develop an alternative formulation for the combination of the n-gram and LSA paradigms. The end result, in effect, is a modified n-gram language model incorporating large-span semantic information.

To achieve this goal, we need to compute:

$$\Pr(w_q|H_{q-1}) = \Pr(w_q|H_{q-1}^{(n)},H_{q-1}^{(l)})\,, \tag{7}$$

where the history H_{q-1} now comprises an n-gram component ($H_{q-1}^{(n)} = w_{q-1}w_{q-2}\ldots w_{q-n+1}$) as well as an LSA component ($H_{q-1}^{(l)} = \tilde{d}_{q-1}$). This expression can be rewritten as:

$$\Pr(w_q|H_{q-1}) = \frac{\Pr(w_q,H_{q-1}^{(l)}|H_{q-1}^{(n)})}{\sum\limits_{w_i\in\mathcal{V}}\Pr(w_i,H_{q-1}^{(l)}|H_{q-1}^{(n)})}\,, \tag{8}$$

where the summation in the denominator extends over all words in $\mathcal{V}$. Expanding and re-arranging, the numerator of (8) is seen to be:

$$\begin{aligned}\Pr(w_q,&H_{q-1}^{(l)}|H_{q-1}^{(n)})\\ &= \Pr(w_q|H_{q-1}^{(n)})\Pr(H_{q-1}^{(l)}|w_q,H_{q-1}^{(n)})\\ &= \Pr(w_q|w_{q-1}w_{q-2}\ldots w_{q-n+1})\\ &\qquad\cdot\Pr(\tilde{d}_{q-1}|w_q w_{q-1}w_{q-2}\ldots w_{q-n+1})\,.\end{aligned} \tag{9}$$

Now we make the assumption that the probability of the document history given the current word is not affected by the immediate context preceding it. This reflects the fact that, for a given word, different syntactic constructs (immediate context) can be used to carry the same meaning (document history). This is obviously reasonable for content words, and probably does not matter very much for function words. As a result,

[1] Note, however, that the LSA method could be trivially modified to accommodate other admissible histories. For example, H_{q-1} could be anything from the last $n-1$ words, to the current sentence, to the current document, to the past m documents (the latter three, of course, up to word w_{q-1}). The choice only depends on what information is available on the dynamics of the relevant parameters, to enable the selection of the largest semantically consistent text unit. This is a major benefit of the large-span approach.

the integrated probability becomes:

$$\Pr(w_q|H_{q-1}) = \frac{\Pr(w_q|w_{q-1}w_{q-2}\dots w_{q-n+1})\Pr(\tilde{d}_{q-1}|w_q)}{\sum\limits_{w_i\in\mathcal{V}} \Pr(w_i|w_{q-1}w_{q-2}\dots w_{q-n+1})\Pr(\tilde{d}_{q-1}|w_i)} . \quad (10)$$

Interestingly, this expression has a Bayesian interpretation. If $\Pr(\tilde{d}_{q-1}|w_q)$ is viewed as a prior probability on the current document history, then (10) simply translates the classical Bayesian estimation of the n-gram (local) probability using a prior distribution obtained from (global) LSA.This provides additional evidence to justify the above assumption.

5 PERFORMANCE

To evaluate the performance of the language model proposed in the previous section, we trained it on the so-called WSJ0 part of the NAB News corpus. This was convenient for comparison purposes since conventional bigram and trigram language models are readily available, trained on exactly the same data [9].

Thus, 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. In addition, about 2 million words from 1992 and 1994 were used for test purposes. 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 [11]. 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 direct model (5) and combined it with the standard bigram, as in (10). We then measured the resulting perplexity on the test data, and found a value of 147. This result is to be compared with the baseline results obtained with the standard bigram and trigram language models of [9], found to be 215 and 142, respectively.

Thus, compared to the standard bigram, we obtain a 32% reduction in perplexity with the integrated bigram/LSA language model (10), which brings it to the same level of performance as the standard trigram. We conclude that the new integrated language model is quite effective in combining global semantic prediction with the usual local predictive power of the bigram language model. In addition, we expect that much of the reduction in perplexity observed at the bigram level would carry over to a combined trigram/LSA language model.

6 CONCLUSION

We have described a language modeling approach based on the latent semantic analysis paradigm, which tracks hidden redundancies across documents. The resulting LSA language models capture semantically oriented, large span relationships between words. This stands in marked contrast with conventional n-grams, which inherently rely on syntactically-oriented, short-span relationships. Hence, one paradigm is better suited to account for the local constraints in the language, while the other one is more adept at handling global constraints. To harness the synergy between the two, we have derived an integrative formulation to combine a standard n-gram with a LSA language model. The resulting multi-span language model was shown to substantially outperform the associated standard n-gram on a subset of the NAB News corpus.

REFERENCES

[1] L.R. Bahl, F. Jelinek, and R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Anal. Mach. Intel.*, Vol. PAMI-5, No. 2, pp. 179–190, March 1983.

[2] M. Jardino and G. Adda, "Automatic Word Classification Using Simulated Annealing," in *Proc.* 1993 *ICASSP*, Minneapolis, MN, pp. 41–44, May 1993.

[3] M. Tamoto and T. Kawabata, "Clustering Word Category Based on Binomial Posteriori Co-Occurrence Distribution,"in *Proc.* 1995 *ICASSP*, Detroit, MI, pp. 165–168, May 1995.

[4] T. Niesler and P. Woodland, "A Variable–Length Category–Based N–Gram Language Model," in *Proc.* 1996 *ICASSP*, Atlanta, GA, pp. I164–I167, May 1996.

[5] A. Farhat, J. Isabelle, and D. O'Shaughnessy, "Clustering Words for Statistical Language Models Based on Contextual Word Similarity," in *Proc.* 1996 *ICASSP*, Atlanta, GA, pp. I180–I183, May 1996.

[6] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger–Based Language Models: A Maximum Entropy Approach," in *Proc.* 1993 *ICASSP*, Minneapolis, MN, pp. II45–48, May 1993.

[7] 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.

[8] S. Deerwester *et al.*, "Indexing by Latent Semantic Analysis," *J. Am. Soc. Inform. Science*, Vol. 41, pp. 391–407, 1990.

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

[10] J.R. Bellegarda *et al.*, "A Novel Word Clustering Algorithm Based on Latent Semantic Analysis," in *Proc.* 1996 *ICASSP*, Atlanta, GA, pp. I172–I175, May 1996.

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