Hierarchical Non-Emitting Markov Models

Eric Sven Ristad and **Robert G. Thomas**
Department of Computer Science
Princeton University
Princeton, NJ 08544-2087
{ristad,rgt}@cs.princeton.edu

Abstract

We describe a simple variant of the interpolated Markov model with non-emitting state transitions and prove that it is strictly more powerful than any Markov model. Empirical results demonstrate that the non-emitting model outperforms the interpolated model on the Brown corpus and on the Wall Street Journal under a wide range of experimental conditions. The non-emitting model is also much less prone to overtraining.

1 Introduction

The Markov model has long been the core technology of statistical language modeling. Many other models have been proposed, but none has offered a better combination of predictive performance, computational efficiency, and ease of implementation. Here we add hierarchical non-emitting state transitions to the Markov model. Although the states in our model remain Markovian, the model itself is no longer Markovian because it can represent unbounded dependencies in the state distribution. Consequently, the non-emitting Markov model is strictly more powerful than any Markov model, including the context model (Rissanen, 1983; Rissanen, 1986), the backoff model (Cleary and Witten, 1984; Katz, 1987), and the interpolated Markov model (Jelinek and Mercer, 1980; MacKay and Peto, 1994).

More importantly, the non-emitting model consistently outperforms the interpolated Markov model on natural language texts, under a wide range of experimental conditions. We believe that the superior performance of the non-emitting model is due to its ability to better model conditional independence. Thus, the non-emitting model is better able to represent both conditional independence and long-distance dependence, ie., it is simply a better statistical model. The non-emitting model is also nearly as computationally efficient and easy to implement as the interpolated model.

The remainder of our article consists of four sections. In section 2, we review the interpolated Markov model and briefly demonstrate that all interpolated models are equivalent to some basic Markov model of the same model order. Next, we introduce the hierarchical non-emitting Markov model in section 3, and prove that even a lowly second order non-emitting model is strictly more powerful than any basic Markov model, of any model order. In section 4, we report empirical results for the interpolated model and the non-emitting model on the Brown corpus and Wall Street Journal. Finally, in section 5 we conjecture that the empirical success of the non-emitting model is due to its ability to better model a point of apparent independence, such as may occur at a sentence boundary.

Our notation is as follows. Let A be a finite alphabet of distinct symbols, $|A| = k$, and let $x^T \in A^T$ denote an arbitrary string of length T over the alphabet A. Then x_i^j denotes the substring of x^T that begins at position i and ends at position j. For convenience, we abbreviate the unit length substring x_i^i as x_i and the length t prefix of x^T as x^t.

2 Background

Here we review the basic Markov model and the interpolated Markov model, and establish their equivalence.

A basic Markov model $\phi = \langle A, n, \delta_n \rangle$ consists of an alphabet A, a model order n, $n \geq 0$, and the state transition probabilities $\delta_n : A^n \times A \to [0,1]$. With probability $\delta_n(y|x^n)$, a Markov model in the state x^n will emit the symbol y and transition to the state $x_2^n y$. Therefore, the probability $p_m(x_t|x^{t-1}, \phi)$ assigned by an order n basic Markov model ϕ to a symbol x^t in the history x^{t-1} depends only on the last n symbols of the history.

$$p_m(x_t|x^{t-1}, \phi) = \delta_n(x_t|x_{t-n}^{t-1}) \qquad (1)$$

An interpolated Markov model $\phi = \langle A, n, \lambda, \delta \rangle$ consists of a finite alphabet A, a maximal model order n, the state transition probabilities $\delta = \delta_0 \ldots \delta_n$, $\delta_i : A^i \times A \to [0,1]$, and the state-conditional interpolation parameters $\lambda = \lambda_0 \ldots \lambda_n$, $\lambda_i : A^i \to [0,1]$.

The probability assigned by an interpolated model is a linear combination of the probabilities assigned by all the lower order Markov models.

$$p_c(y|x^i,\phi) = \lambda_i(x^i)\delta_i(y|x^i) + (1-\lambda_i(x^i))p_c(y|x_2^i,\phi) \quad (2)$$

where $\lambda_i(x^i) = 0$ for $i \geq n$, and and therefore $p_c(x_t|x^{t-1},\phi) = p_c(x_t|x_{t-n}^{t-1},\phi)$, ie., the prediction depends only on the last n symbols of the history.

In the interpolated model, the interpolation parameters smooth the conditional probabilities estimated from longer histories with those estimated from shorter histories (Jelinek and Mercer, 1980). Longer histories support stronger predictions, while shorter histories have more accurate statistics. Interpolating the predictions from histories of different lengths results in more accurate predictions than can be obtained from any fixed history length.

A quick glance at the form of (2) and (1) reveals the fundamental simplicity of the interpolated Markov model. Every interpolated model ϕ is equivalent to some basic Markov model ϕ' (lemma 2.1), and every basic Markov model ϕ is equivalent to some interpolated context model ϕ' (lemma 2.2).

Lemma 2.1
$\forall\phi\ \exists\phi'\ \forall x^T \in A^*\ [p_m(x^T|\phi',T) = p_c(x^T|\phi,T)]$

Proof. We may convert the interpolated model ϕ into a basic model ϕ' of the same model order n, simply by setting $\delta'_n(y|x^n)$ equal to $p_c(y|x^n,\phi)$ for all states $x^n \in A^n$ and symbols $y \in A$. □

Lemma 2.2
$\forall\phi\ \exists\phi'\ \forall x^T \in A^*\ [p_c(x^T|\phi',T) = p_m(x^T|\phi,T)]$

Proof. Every basic model is equivalent to an interpolated model whose interpolation values are unity for states of order n. □

The lemmas suffice to establish the following theorem.

Theorem 1 *The class of interpolated Markov models is equivalent to the class of basic Markov models.*

Proof. By lemmas 2.1 and 2.2. □

A similar argument applies to the backoff model. Every backoff model can be converted into an equivalent basic model, and every basic model is a backoff model.

3 Non-Emitting Markov Models

A hierarchical non-emitting Markov model $\phi = \langle A, n, \lambda, \delta\rangle$ consists of an alphabet A, a maximal model order n, the state transition probabilities, $\delta = \delta_0 \ldots \delta_n$, $\delta_i : A^i \times A \rightarrow [0,1]$, and the non-emitting state transition probabilities $\lambda = \lambda_0 \ldots \lambda_n$, $\lambda_i : A^i \rightarrow [0,1]$. With probability $1-\lambda_i(x^i)$, a non-emitting model will transition from the state x^i to the state x_2^i without emitting a symbol. With probability $\lambda_i(x^i)\delta_i(y|x^i)$, a non-emitting model will transition from the state x^i to the state x^iy and emit the symbol y.

Therefore, the probability $p_\epsilon(y^j|x^i,\phi)$ assigned to a string y^j in the history x^i by a non-emitting model ϕ has the recursive form (3),

$$p_\epsilon(y^j|x^i,\phi) = \lambda_i(x^i)\delta_i(y_1|x^i)p_\epsilon(y_2^j|x^iy_1,\phi) + (1-\lambda_i(x^i))p_\epsilon(y^j|x_2^i,\phi) \quad (3)$$

where $\lambda_i(x^i) = 0$ for $i > n$ and $\lambda_0(\epsilon) = 1$. Note that, unlike the basic Markov model, $p_\epsilon(x_t|x^{t-1},\phi) \neq p_\epsilon(x_t|x_{t-n}^{t-1},\phi)$ because the state distribution of the non-emitting model depends on the prefix x^{i-n}. This simple fact will allow us to establish that there exists a non-emitting model that is not equivalent to any basic model.

Lemma 3.1 states that there exists a non-emitting model ϕ that cannot be converted into an equivalent basic model of any order. There will always be a string x^T that distinguishes the non-emitting model ϕ from any given basic model ϕ' because the non-emitting model can encode unbounded dependencies in its state distribution.

Lemma 3.1
$\exists\phi\ \forall\phi'\ \exists x^T \in A^*\ [p_\epsilon(x^T|\phi,T) \neq p_m(x^T|\phi',T)]$

Proof. The idea of the proof is that our non-emitting model will encode the first symbol x_1 of the string x^T in its state distribution, for an unbounded distance. This will allow it to predict the last symbol x_T using its knowledge of the first symbol x_1. The basic model will only be able predict the last symbol x_T using the preceding n symbols, and therefore when T is greater than n, we can arrange for $p_\epsilon(x^T|\phi,T)$ to differ from any $p_m(x^T|\phi',T)$, simply by our choice of x_1.

The smallest non-emitting model capable of exhibiting the required behavior has order 2. The non-emitting transition probabilities λ and the interior of the string x_2^{T-1} will be chosen so that the non-emitting model is either in an order 2 state or an order 0 state, with no way to transition from one to the other. The first symbol x_1 will determine whether the non-emitting model goes to the order 2 state or stays in the order 0 state. No matter what probability the basic model assigns to the final symbol x_T, the non-emitting model can assign a different probability by the appropriate choice of x_1, $\delta_0(x_T)$, and $\delta_2(x_T|x_{T-2}^{T-1})$.

Consider the second order non-emitting model over a binary alphabet with $\lambda(0) = 1$, $\lambda(1) = 0$, and $\lambda(11) = 1$ on strings in $A1^*A$. When $x_1 = 0$, then x_2 will be predicted using the 1st order model $\delta_1(x_2|x_1)$, and all subsequent x_t will be predicted by the second order model $\delta_2(x_t|x_{t-2}^{t-1})$. When $x_1 = 0$, then all subsequent x_t will be predicted by the 0th order model $\delta_0(x_t)$. Thus for all $t > p$, $p_\epsilon(x_t|x^{t-1}) \neq p_\epsilon(x_t|x_{t-p}^{t-1})$ for any fixed p, and no basic model is equivalent to this simple non-emitting model. □

It is obvious that every basic model is also a non-emitting model, with the appropriate choice of non-

emitting transition probabilities.

Lemma 3.2
$\forall \phi \; \exists \phi' \; \forall x^T \in A^* \; [p_\epsilon(x^T|\phi', T) = p_m(x^T|\phi, T)]$

These lemmas suffice to establish the following theorem.

Theorem 2 *The class of non-emitting Markov models is strictly more powerful than the class of basic Markov models, because it is able to represent a larger class of probability distributions on strings.*

Proof. By lemmas 3.1 and 3.2. □

Since interpolated models and backoff models are equivalent to basic Markov models, we have as a corollary that non-emitting Markov models are strictly more powerful than interpolated models and backoff models as well. Note that non-emitting Markov models are considerably less powerful than the full class of stochastic finite state automaton (SFSA) because their states are Markovian. Non-emitting models are also less powerful than the full class of hidden Markov models.

Algorithms to evaluate the probability of a string according to a non-emitting model, and to optimize the non-emitting state transitions on a training corpus are provided in related work (Ristad and Thomas, 1997).

4 Empirical Results

The ultimate measure of a statistical model is its predictive performance in the domain of interest. To take the true measure of non-emitting models for natural language texts, we evaluate their performance as character models on the Brown corpus (Francis and Kucera, 1982) and as word models on the Wall Street Journal. Our results show that the non-emitting Markov model consistently gives better predictions than the traditional interpolated Markov model under equivalent experimental conditions. In all cases we compare non-emitting and interpolated models of identical model orders, with the same number of parameters. Note that the non-emitting bigram and the interpolated bigram are equivalent.

Corpus	*Size*	*Alphabet*	*Blocks*
Brown	6,004,032	90	21
WSJ 1989	6,219,350	20,293	22
WSJ 1987-89	42,373,513	20,092	152

All λ values were initialized uniformly to 0.5 and then optimized using deleted estimation on the first 90% of each corpus (Jelinek and Mercer, 1980).

DELETED-ESTIMATION($\mathbf{B}, \phi$)
1. Until convergence
2. Initialize λ^+, λ^- to zero;
3. For each block B_i in $\mathbf{B}$
4. Initialize δ using $\mathbf{B} - B_i$;
5. EXPECTATION-STEP($B_i, \phi, \lambda^+, \lambda^-$);
6. MAXIMIZATION-STEP($\phi, \lambda^+, \lambda^-$);
7. Initialize δ using $\mathbf{B}$;

Here $\lambda^+(x^i)$ accumulates the expectations of emitting a symbol from state x^i while $\lambda^-(x^i)$ accumulates the expectations of transitioning to the state x^i_2 without emitting a symbol.

The remaining 10% percent of each corpus was used to evaluate model performance. No parameter tying was performed.[1]

4.1 Brown Corpus

Our first set of experiments were with character models on the Brown corpus. The Brown corpus is an eclectic collection of English prose, containing 6,004,032 characters partitioned into 500 files. Deleted estimation used 21 blocks. Results are reported as per-character test message entropies (bits/char), $-\frac{1}{v}\log_2 p(y^v|v)$. The non-emitting model outperforms the interpolated model for all nontrivial model orders, particularly for larger model orders. The non-emitting model is considerably less prone to overtraining. After 10 EM iterations, the order 9 non-emitting model scores 2.0085 bits/char while the order 9 interpolated model scores 2.3338 bits/char after 10 EM iterations.

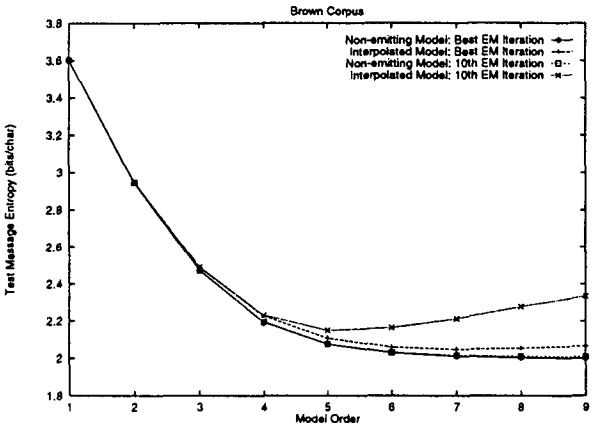

Figure 1: Test message entropies as a function of model order on the Brown corpus.

4.2 WSJ 1989

The second set of experiments was on the 1989 Wall Street Journal corpus, which contains 6,219,350 words. Our vocabulary consisted of the 20,293 words that occurred at least 10 times in the entire WSJ 1989 corpus. All out-of-vocabulary words

[1] In forthcoming work, we compare the performance of the interpolated and non-emitting models on the Brown corpus and Wall Street Journal with ten different parameter tying schemes. Our experiments confirm that some parameter tying schemes improve model performance, although only slightly. The non-emitting model consistently outperformed the interpolated model on all the corpora for all the parameter tying schemes that we evaluated.

were mapped to a unique OOV symbol. Deleted estimation used 22 blocks. Following standard practice in the speech recognition community, results are reported as per-word test message perplexities, $p(y^v|v)^{-\frac{1}{v}}$. Again, the non-emitting model outperforms the interpolated Markov model for all nontrivial model orders.

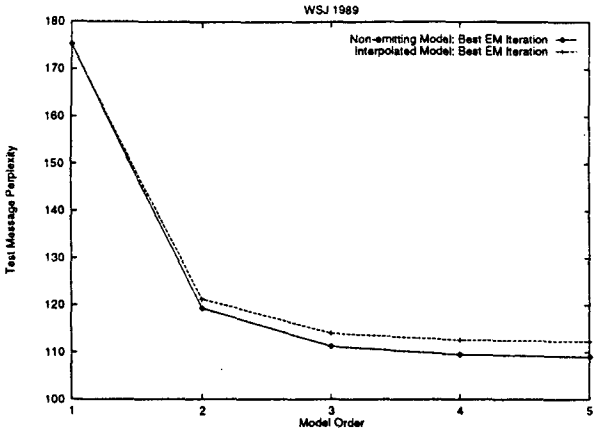

Figure 2: Test message perplexities as a function of model order on WSJ 1989.

4.3 WSJ 1987-89

The third set of experiments was on the 1987-89 Wall Street Journal corpus, which contains 42,373,513 words. Our vocabulary consisted of the 20,092 words that occurred at least 63 times in the entire WSJ 1987-89 corpus. Again, all out-of-vocabulary words were mapped to a unique OOV symbol. Deleted estimation used 152 blocks. Results are reported as test message perplexities. As with the WSJ 1989 corpus, the non-emitting model outperforms the interpolated model for all nontrivial model orders.

5 Conclusion

The power of the non-emitting model comes from its ability to represent additional information in its state distribution. In the proof of lemma 3.1 above, we used the state distribution to represent a long distance dependency. We conjecture, however, that the empirical success of the non-emitting model is due to its ability to remember to ignore (ie., to forget) a misleading history at a point of apparent independence.

A point of apparent independence occurs when we have adequate statistics for two strings x^{n-1} and y^n but not yet for their concatenation $x^{n-1}y^n$. In the most extreme case, the frequencies of x^{n-1} and y^n are high, but the frequency of even the medial bigram $x_{n-1}y_1$ is low. In such a situation, we would like to ignore the entire history x^{n-1} when predicting y^n, because all $\delta(y_j|x_i^{n-1}y_1^{j-1})$ will be close to zero

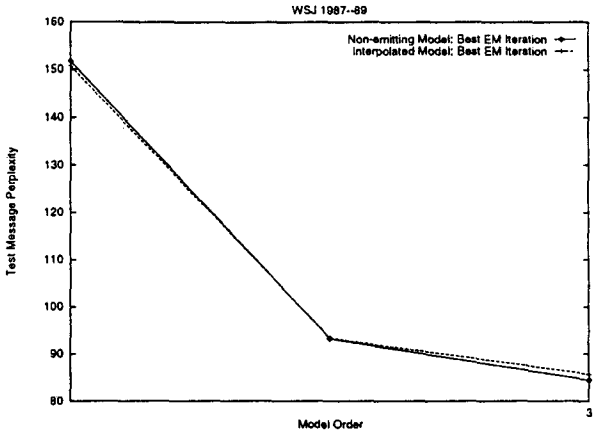

Figure 3: Test message perplexities as a function of model order on WSJ 1987-89.

for $i < n$. To simplify the example, we assume that $\delta(y_j|x_i^{n-1}y_1^{j-1}) = 0$ for $j \geq 1$ and $i < n$.

In such a situation, the interpolated model must repeatedly transition past some suffix of the history x^{n-1} for each of the next $n-1$ predictions, and so the total probability assigned to $p_c(y^n|\epsilon)$ by the interpolated model is a product of $n(n-1)/2$ probabilities.

$$\begin{aligned} & p_c(y^n|x^{n-1}) \\ &= \left[\prod_{i=1}^{n-1}(1-\lambda(x_i^{n-1}))\right] p(y_1|\epsilon) \\ &\quad \left[\prod_{i=2}^{n-1}(1-\lambda(x_i^{n-1}y_1))\right] p(y_2|y_1) \\ &\quad \cdots \\ &\quad (1-\lambda(x_{n-1}y_1^{n-1}))p(y_n|y^{n-1}) \\ &= \left[\prod_{k=1}^{n-1}\prod_{i=k}^{n-1}(1-\lambda(x_i^{n-1}y_1^{i-1}))\right] p_c(y^n|\epsilon) \end{aligned} \quad (4)$$

In contrast, the non-emitting model will immediately transition to the empty context in order to predict the first symbol y_1, and then it need never again transition past any suffix of x^{n-1}. Consequently, the total probability assigned to $p_\epsilon(y^n|\epsilon)$ by the non-emitting model is a product of only $n-1$ probabilities.

$$p_\epsilon(y^n|x^{n-1}) = \left[\prod_{i=1}^{n-1}(1-\lambda(x_i^{n-1}))\right] p_\epsilon(y^n|\epsilon) \quad (5)$$

Given the same state transition probabilities, note that (4) must be considerably less than (5) because probabilities lie in $[0,1]$. Thus, we believe that the empirical success of the non-emitting model comes from its ability to effectively ignore a misleading history rather than from its ability to remember distant events.

Finally, we note the use of hierarchical non-emitting transitions is a general technique that may be employed in any time series model, including context models and backoff models.

Acknowledgments

Both authors are partially supported by Young Investigator Award IRI-0258517 to Eric Ristad from the National Science Foundation.

References

Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, Robert L. Mercer, and David Nahamoo. 1991. A fast algorithm for deleted interpolation. In *Proc. EUROSPEECH '91*, pages 1209–1212, Genoa.

J.G. Cleary and I.H. Witten. 1984. Data compression using adaptive coding and partial string matching. *IEEE Trans. Comm.*, COM-32(4):396–402.

W. Nelson Francis and Henry Kucera. 1982. *Frequency analysis of English usage: lexicon and grammar*. Houghton Mifflin, Boston.

Fred Jelinek and Robert L. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In Edzard S. Gelsema and Laveen N. Kanal, editors, *Pattern Recognition in Practice*, pages 381–397, Amsterdam, May 21–23. North Holland.

Slava Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. ASSP*, 35:400–401.

David J.C. MacKay and Linda C. Bauman Peto. 1994. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(1).

Jorma Rissanen. 1983. A universal data compression system. *IEEE Trans. Information Theory*, IT-29(5):656–664.

Jorma Rissanen. 1986. Complexity of strings in the class of Markov sources. *IEEE Trans. Information Theory*, IT-32(4):526–532.

Eric Sven Ristad and Robert G. Thomas. 1997. Hierarchical non-emitting Markov models. Technical Report CS-TR-544-96, Department of Computer Science, Princeton University, Princeton, NJ, March.

Frans M. J. Willems, Yuri M. Shtarkov, and Tjalling J. Tjalkens. 1995. The context-tree weighting method: basic properties. *IEEE Trans. Inf. Theory*, 41(3):653–664.