YY11

TRAINABLE GRAMMARS FOR SPEECH RECOGNITION

James K. Baker, Dialog Systems, Inc., Belmont, MA 02172

Algorithms which are based on modeling speech as a finite-state, hidden Markov process have been very successful in recent years. This paper generalizes these algorithms to certain denumerable-state, hidden Markov processes. This algorithm permits automatic training of the stochastic analog of an arbitrary context-free grammar. In particular, in contrast to many grammatical inference methods, the new algorithm allows the grammar to have an arbitrary degree of ambiguity. Furthermore, allowing ambiguity in the grammar allows errors in the recognition process to be explicitly modeled in the grammar rather than added as an extra component.

THEORY

This section provides a basic introduction to the concept of a hidden Markov process. Let there be two sequences of random variables (X(t)) and (Y(t)), $\underline{1} < \underline{t} < T$. The sequence (X(t)) is known to be generated by a Markov process, but is not directly observed. The sequence (Y(t)), which is observed, depends probabilistically on the sequence (X(t)). The basic task in stochastic pattern analysis is to make inferences about the sequence (X(t)) from the observations (Y(t)).

The assumption that the sequence (X(t)) is generated by a stationary Markov process is expressed by equation (1).

$$\begin{aligned} &\mathrm{Prob}(X(t+1)=j \mid X(1)=x(1),X(2)=x(2), \\ &\qquad\qquad \ldots,X(t-1)=x(t-1),X(t)=i) \qquad (1) \\ &\qquad = \mathrm{Prob}(X(t+1)=j \mid X(t)=i) \\ &\quad = a(i,j) \end{aligned}$$

The quantity a(i,j) is called the transition probability from state i to state j. Equation (1) may be paraphrased as "if the current state is known, then the future of the sequence (X(t)) is conditionally independent of the past." The dependence of the sequence (Y(t)) on the sequence (X(t)) is expressed by equation (2).

$$\begin{aligned} &\mathrm{Prob}(Y(t+1)=k \mid X(1)=x(1),X(2)=x(2), \\ &\qquad\qquad \ldots,X(t-1)=x(t-1),X(t)=i,X(t+1)=j) \qquad (2) \\ &\qquad = \mathrm{Prob}(Y(t+1)=k \mid X(t)=i,X(t+1)=j) \\ &\quad = b(i,j,k) \end{aligned}$$

Equation (2) may be paraphrased as "if the current state transition is known, then the current observation Y(t) is conditionally independent of the past."

A fundamental problem in stochastic pattern analysis is to evaluate the probability of a sequence of observations, conditional on some hypothesis about the underlying pattern. For the purpose of facilitating such evaluations, the quantities defined in equations (3) and (4) are introduced.

$$e(t,j) = \mathrm{Prob}(\ Y(1)=y(1), Y(2)=y(2), \ldots, Y(t)=y(t), X(t)=j\ ) \qquad (3)$$

$$f(t,j) = \mathrm{Prob}(\ Y(t+1)=y(t+1), \ldots, Y(T)=y(T)\ |\ X(t)=j\ ) \qquad (4)$$

ESTIMATION METHOD

The quantities defined in equations (3) and (4) may be evaluated very efficiently by a procedure, introduced by Baum[2], which is computationally very similar to dynamic programming. The essence of this procedure is given in equations (5) and (6).

$$e(t,j) = \sum_{i} e(i,t-1)a(i,j)b(i,j,y(t)) \qquad (5)$$

$$f(t,i) = \sum_{j} f(j,t+1)a(i,j)b(i,j,y(t+1)) \qquad (6)$$

A very important result of Baum's is that if the true matrices a(i,j) and b(i,j,k) are not known, then an iterative procedure exists for making estimates of a(i,j) and b(i,j,k) which satisfy a local optimality condition. The procedure consists of using any estimates of the matrices a(i,j) and b(i,j,k) and performing the evaluations given in equations (5) and (6) as if these were the true probability matrices. Then a new estimate for the matrices a(i,j) and b(i,j,k) is obtained according to equations (7) and (8).

$$a(i,j) = \frac{\sum_{t} e(i,t-1)a(i,j)b(i,j,y(t))f(j,t)}{\sum_{t} e(i,t-1)f(i,t-1)} \qquad (7)$$

$$b(i,j,k) = \frac{\sum_{t,y(t)=k} e(i,t-1)a(i,j)b(i,j,k)f(j,t)}{\sum_{t} e(i,t-1)a(i,j)b(i,j,y(t))f(j,t)} \qquad (8)$$

The re-estimated values of a(i,j) and b(i,j,k) are then used again in equations (5) and (6) to make another step of the iteration.

The procedure for estimating the parameters of a finite-state, hidden Markov process is by now well-established in speech recognition as a procedure for automatic training of the parameters of a model of an acoustic processor. We wish to generalize this procedure to the analysis of the denumerable Markov process which corresponds to a probabilistic context-free grammar. The eventual goal is to obtain a procedure for automatically training a stochastic, grammar-based model of a natural language.

CONTEXT-FREE MODEL

For a stochastic context-free grammar, it is more convenient to consider the hidden stochastic process not as a sequence of random variables, but rather as a family of random variables, with one random variable X(s,t) associated with each "span" (integer interval). Informally, X(s,t) is the non-terminal symbol (or phrase) associated with (producing) the entire sequence of observations Y(s),Y(s+1),...,Y(t). The matrices of parameters which describe this stochastic context-free grammar are a(i,j,k) and b(j,k), where a(i,j,k) is the probability that the non-terminal symbol i will generate the pair of non-terminal symbols j and k, and b(j,k) is the probability that the non-terminal symbol j will generate just the single terminal symbol k. By assumption these probabilities are not dependent on the context. Since any context-free grammar may be reduced to Chomsky normal form, these parameters are sufficient to descibe any stochastic context-free grammar.

The intermediate quantities which facilitate analysis of a hidden, stochastic, context-free grammar are defined in equations (9) and (10).

$$e(s,t,i) = \mathrm{Prob}(\ Y(s)=y(s),\ \ldots\ ,Y(t)=y(t),\ X(s,t)=i\ ) \qquad (9)$$

$$f(s,t,i) = \mathrm{Prob}(Y(1),\ldots,Y(s-1),Y(t+1),\ldots,Y(T)\ |\ X(s,t)=i\ ) \qquad (10)$$

These quantities are more difficult to evaluate than the analogous quantities for a finite-state process. However, the may be computed by the method indicated in equations (11) and (12).

$$e(s,t,i) = \sum_{r,j,k} e(s,r,j)e(r,t,k)a(i,j,k) \qquad (11)$$

$$\begin{aligned} f(s,t,i) = &\sum_{r,j,k} f(r,t,j)e(r,s,k)a(j,k,i) \\ + &\sum_{r,j,k} f(s,r,j)e(t,r,k)a(j,i,k) \end{aligned} \qquad (12)$$

We have assumed that the quantities have been appropriately initialized for the spans of length one.

If the true values of the matrices a(i,j,k) and b(j,k) are not known, then again an iterative re-estimation procedure is possible. This procedure is given in equations (13) and (14).

$$a(i,j,k) = \frac{\sum_{t} f(r,t,i)e(r,s,j)e(s,t,k)a(i,j,k)}{\sum_{t} \sum_{l,m} f(r,t,i)e(r,s,j)e(s,t,k)a(i,l,m)} \quad (13)$$

$$b(j,k) = \frac{\sum_{t,y(t)=k} f(t,t,j)e(t,t,j)}{\sum_{t} f(t,t,j)e(t,t,j)} \quad (14)$$

The crucial idea which has been introduced is that the hidden random variables should be associated with the spans X(s,t) rather than with single sample times X(t) as in the finite state model. Once this idea is introduced, the generalization of the equations is straight-forward.

INTERPRETATION

The abstract equations can, of course, be interpreted in several ways depending on what we choose as the elements of the underlying grammar. For example, the terminal symbols could be words, the non-terminal symbols could be parts-of-speech and phrase types, and the productions could constitute a context-free phrase-structure grammar. Because we use a probability function b(j,k) to associate words and parts-of-speech, any word can be associated with an arbitrary number of different parts-of-speech, and each part-of-speech can be associated with an arbitrary number of different words. This might seem like an obvious property, but many grammatical inferences methods must assume that the words of the vocabulary can be partitioned into equivalence classes, with a unique part of speech for each class.

Another interpretation is to include the words among the (hidden) non-terminal symbols, and to take the terminal symbols to be a sequence of noisy observations of the words (or of the letters or sounds composing the words). This is the interpretation to be used in speech recognition or in optical character or other image recognition. When grammars (even probabilistic grammars) have been used in recognition of noisy sequences an extra criterion, such as minimum Hamming distance, has usually been introduced to measure the discrepancy between the closest string generated by the grammar and the noisy observation string. With the model described above, we can include the noise of the observations with the general stochastic model.

References

[1] J.K. Baker, "Stochastic Modeling for Automatic Speech Understanding," in Speech Recognition, D. Raj Reddy(ed.), Academic Press, 1975.

[2] L.E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of a Markov Process," in Inequalities, Vol. III, 1972.