A SIMPLIFIED HEURISTIC VERSION OF RAVIV'S ALGORITHM FOR USING CONTEXT IN TEXT RECOGNITIONS

R. Shinghal
School of Computer Science
McGill University

D. Rosenberg
Department of Electrical Engineering
McGill University

G.T. Toussaint
School of Computer Science
McGill University

Abstract

Word-position-independent and word-position dependent n-gram probabilities were estimated from a large English language corpus. A text-recognition problem was simulated, and using the estimated n-gram probabilities, four experiments were conducted by the following methods of classification: without contextual information, Raviv's recursive Bayes algorithm, the modified Viterbi algorithm, and a proposed heuristic approximation to Raviv's algorithm. Based on the estimates of the probabilities of misclassifixation observed in the four experiments, the above methods are compared. The heuristic approximation of Raviv's algorithm performed just as well as Raviv's and required far less computation.

1. Introduction

Some of the contextual text-recognition algorithms invoke the assumption that English language is a Markov source of order r>I[I]. Having made these assumptions these algorithms use either sequential or nonsequential decision theory. This paper describes an experimental investigation of text-recognition by four methods: without contextual information, Raviv's recursive Bayes algorithm [2], the modified Viterbi algorithm 15], and a proposed heuristic approximation to Raviv's algorithm. The objective of the experiments was to compare the above methods.

It was assumed that English language Is a Markov source of order r=I . To conduct the experiments, it was necessary that unigram (single-letter) and first-order transition probabilities be estimated from the English language. The probabilities are collectively referred to as the n-gram probabilities.

2. Estimation of N-gram Probabilities From the English Language

A multi-authored English language corpus written on ten different subject matters [4] comprising over half-a-million words was compiled. Probability estimates were obtained from this corpus for the following:

(i) Word-position-independent unigram probabilities $P(C_1)$.

(ii) Word-position-independent transition probabilities PCcJc.) .

(iii) Word-position-dependent unigram probabilities P (C.) for position m
m l
(l<m<9) .

(iv) Word-position-dependent transition probabilities P (C,|C.) for 0<m<9 .
m i j

P (C,IC) is the probability of occurrence of character C. in position (m+1) of a word given that character C occurs in position m.

Since the corpus was large, it was assumed that PCC^) , $P(C_1$ |Cj), V V ' ?m(CilCj) e8tI" mated from the corpus, reflect closely the true n-gram probabilities of the English language for the 27-character set, tf and the letters A to Z. Each experiment described in Section 4 was repeated twice: once, by using the word-position-independent (WPI) n-gram probabilities and a second time, by using word-position-dependent (WPD) n-gram probabilities.

3• Simulation of the Text-Recognition Problem

Images on 24 x 24 arrays of mixed-font machine printed letters were taken from Ryan's* data set. The images were size-normalized and a 36-dimensional feature-vector X= (x.. ,x_ ,. . . ,x) was extracted from each image [4].

The total number of patterns available was 13,337. These were divided into two sets: a training set of 6,651 patterns and a testing set of 6,686 patterns. The classifier was trained, and P(X|C)—the probability distribution of X conditioned on C_k—was estimated,

A passage, called the Old Passage, was compiled from ten segments, such that each segment was arbitrarily chosen from one of the ten subject matters in the corpus (described in Section 2). Another passage, called the New Passage, was similarly compiled but from sources outside the corpus [4], The Old Passage consisted of 2,521 words and the New Passage of 2,256 words. These two passages were used as text to be recognized in the experiments described in Section 4.

4. The Text-Recognition Problem and the Experiments Conducted

Patterns of the passages were presented sequentially to the classifier. It was assumed that all blanks were perfectly recognizable. Let

$$x * x_Q, x_{\pm 9} x_2 . . . . x_n, x_{n+1}$$

be the pattern-sequence input to the classifier

t This research was supported by the National Research Council of Canada under grant number NRC-A9293.

* Pattern Recognition Data Base No. 1.1.1 A, IEEE Computer Society

such that X_0 and X_{n+1} are blanks. Thus, X_1 to X_n is a word of patterns, and a text is recognized one word at a time.

Let $P(\bar{X}|\bar{\lambda}=\bar{Z})$ denote the probability of $\bar{X}$ conditioned on $\bar{\lambda}=\bar{Z}$, the sequence of classes

$$\bar{\lambda} = \lambda_0, \lambda_1, \ldots, \lambda_n, \lambda_{n+1}$$

taking on the values

$$\bar{Z} = Z_o, Z_1, \ldots, Z_n, Z_{n+1} .$$

Thus Z_0 and Z_{n+1} are $\not{b}$ and Z_1 to Z_n can each take on any of the values of the 26 letters of the English alphabet. Let C_1 to C_{26} refer to the 26 letters of the English alphabet.

4.1 Classification Without Contextual Information

In this case, the class of pattern X_i was decided by a Bayes classifier such that

$$P(X_i|\lambda_i=Z_i)P(Z_i)$$

was maximized over the values of Z_i. Here we use the information present in the feature vector and the apriori knowledge of the probability of occurrence of a particular letter in the English language.

4.2 Raviv's Recursive Bayes Algorithm

In this method pattern $X_m (1 \le m \le n)$ is classified by maximizing the a posteriori probability $P(\lambda_m=C_k|X_1,X_2,\ldots,X_m)$ for $1 \le k \le 26$. Raviv [2] showed that this a posteriori probability can be expressed as follows:

$$\frac{P(X_m|\lambda_m=C_k) \sum_{i=1}^{26} P(\lambda_{m-1}=C_i|X_1,\ldots,X_{m-1})P(\lambda_m=C_k|\lambda_{m-1}=C_i)}{\sum_{j=1}^{26} P(X_m|\lambda_m=C_j) \sum_{i=1}^{26} P(\lambda_{m-1}=C_i|X_1,\ldots,X_{m-1})P(\lambda_m=C_j|\lambda_{m-1}=C_i)}$$

4.3 The Modified Viterbi Algorithm

In this algorithm a decision is made on a whole word by maximizing

$$\prod_{i=1}^{n} P(X_i|\lambda_i=Z_i) \prod_{j=0}^{n+1} P(\lambda_{j+1}=Z_{j+1}|\lambda_j=Z_j) \qquad (4.3\text{-}1)$$

Forney [3] showed that expression (4.3-1) can be maximized by the Viterbi algorithm. Toussaint and Chung [5] proposed a modification where instead of considering all 26 alternatives for every pattern X_i, only the d $(1 \le d \le 26)$ most probable alternatives using $P(X_i|\lambda_i=Z_i)P(Z_i)$ are considered for each letter. In [4] a formal algorithmic implementation of the modified Viterbi algorithm is described. It is shown experimentally in [4] that the modified Viterbi algorithm reduces computation without increasing the error-rate. In the experiment described in this paper, the modified Viterbi algorithm was used to maximize expression (4.3-1). Thus a decision on classifying a word is made after considering all the feature-vectors in the word.

4.4 A Heuristic Approximation to Raviv's Algorithm

In this proposed method, pattern X_i $(1 \le i \le n)$ is classified by maximizing $P(X_i|\lambda_i=Z_i)P(\lambda_i=Z_i|\lambda_{i-1}=Z_{i-1})$ over the values of Z_i. Here we use information present in the feature vector X_i and in the decision that was made on the pattern just before X_i.

METHOD OF CLASSIFICATION	USING WORD-POSITION -INDEPENDENT N-GRAM PROBABILITIES		USING WORD-POSITION -DEPENDENT N-GRAM PROBABILITIES	
	OLD PASSAGES	NEW PASSAGES	OLD PASSAGES	NEW PASSAGES
WITHOUT CONTEXTUAL INFORMATION	15.40	15.50	15.10	15.00
RAVIV'S ALGORITHM	14.06	14.20	13.87	14.08
MODIFIED VITERBI ALGORITHM	12.80	13.10	12.40	12.50
HEURISTIC APPROXIMATION OF RAVIV'S ALGORITHM	14.07	14.25	13.89	14.09

TABLE 1. PERCENTAGE PROBABILITY OF ERROR

5. Conclusions

An unexpected result is the fact that the heuristic approximation of Raviv's algorithm performed just as well as Raviv's recursive Bayes approach and yet requires far less computation as is evident from Section 4.4.

It is, of course, also realized that the different results in Table 1 suggest that context in English is not well-modelled by the Markov assumption.

References

1. Toussaint, G.T., "Recent Progress in Statistical Methods Applied to Pattern Recognition," Proceedings of Second International Conference on Pattern Recognition, Copenhagen, Copenhagen, August 1974, pp. 478-489.

2. Raviv, J., "Decision Making in Markov Chains Applied to the Problem of Pattern Recognition," IEEE Trans. I.T., Vol. IT-13, Oct. 1967, pp. 536-551.

3. Forney, G.D., Jr., "The Viterbi Algorithm," Proc. IEEE, Vol. 61, March 1973, pp.268-278.

4. Shinghal, R., "Using Contextual Information to Improve performance of Character Recognition Machines," Ph.D. Thesis, School of Computer Science, McGill University, 1977.

5. Toussaint, G.T. and Chung, S., "The Modified Viterbi algorithm as an aid to Text Recognition," Manuscript in preparation, School of Computer Science, McGill University, 1977.