제21회 한국정보처리학회 춘계학술발표대회 논문집 제11권 제1호 (2004. 5)

사전 단어 발생 확률을 통해 온라인 필기체 문자의 오인식을 보정하는 후처리 기법

이도곤, 한정훈, 김우생
광운대학교 컴퓨터 공학부
e-mail : {foxhunt, sopure, kwsrain}@cs.kw.ac.kr

The post processing method to reduce the misrecognition of on-line handwritten letters by using an occurrence probability of dictionary words

Do-Gon Lee, Jeong-Hoon Han, Woosaeng Kim
Dept. of Computer Science, Kwangwoon Univ.

요 약

문자들 중에는 매우 비슷한 모양을 갖고 있는 문자가 존재하기 때문에 오인식은 이러한 유사한 문자들 사이에서 일어날 경우가 많다고 볼 수 있다. 즉, 입력된 문자가 유사한 다른 문자에 대응하는 모델에서 발생 확률이 가장 높게 나와 오인식이 되었다고 할지라도, 해당 모델에서는 입력된 문자의 발생 확률도 여전히 높다고 볼 수 있다. 본 논문에서는 사전을 통한 후처리 시, 오인식 된 단어에서 사용된 모델들을 통해 오인식을 보정하는 방법을 제안한다.

1. 서론

전자 펜은 휴대성과 편의성의 장점을 지녀 PDA, palm, pocket PC, 스마트 폰, 테블릿 PC 등의 많은 응용에 사용될 수 있기에 높은 인식률의 온라인 문자 인식 방법이 더욱 필요하다.

온라인 문자 인식 시스템은 전처리를 한 후에 정합을 하게 되는데, 이러한 정합 연구들에 쓰인 방법들 중에서 은닉 마르코프 모델 (HMM)은 음성 인식 분야뿐 아니라 문자 인식에서도 성공적으로 적용되고 있다. HMM 은 잘 정의된 수학적 이론을 배경으로 하며, 전이 확률과 관찰 확률의 이중 확률 구조로 온라인 문자 입력의 특성인 시간에 따른 획의 변이와 공간적인 모양 변이를 동시에 모델링 하기에 적합한 특성을 보인다[1,2].

후처리 기법으로는 Dictionary Look-up 등 문맥적 지식의 구조적 표현에 의한 방법[3], 수정된 비터비 알고리즘 등 문맥적 지식의 확률적 표현에 기초한 방법[4,5], 그리고 예측 교정 알고리즘 등 앞의 두 방식을 결합한 복합적 방법[6]등이 있다.

본 논문에서는 미리 정의된 단어들만이 사용되는 응용에서 사전의 정보를 이용해 오인식을 줄이는 새로운 방법을 제안한다. 문자들 중에는 매우 비슷한 모양을 갖고 있는

문자들이 존재하며, 오인식은 이러한 일부의 유사한 문자들 사이에서 일어날 경우가 많다고 볼 수 있다. 즉, 입력된 문자가 유사한 다른 문자에 대응하는 모델을 통해 오인식이 되었다고 할지라도, 해당 모델에서는 입력된 문자의 발생 확률도 여전히 높다고 볼 수 있다. 따라서 인식된 어떤 단어가 사전에 없는 경우, 사전에 있는 각 단어의 발생 확률을 오인식 된 단어에 사용된 모델들을 통해 계산해서 가장 확률이 높은 경우의 단어로 결정할 수 있다. 제안하는 방법의 타당성을 효율적으로 검증하기 위해서 본 논문에서는 외부 분할 된 영어 알파벳 대문자와 숫자만을 인식의 대상으로 하였다.

2. 은닉 마르코프 모델

은닉 마르코프 모델은 유한개의 상태와 각 상태 사이를 방향성 있게 연결하는 전이의 집합으로 구성된 네트워크로 정의된다. 상태간 전이 확률은 마르코프 프로세스를 따르며 출력 확률은 시간에 독립적으로 각 상태에 지정되어 있다. 일반적으로 HMM 은 $\lambda = (A, B, \pi)$의 기호로 표현되며 A 는 상태전이 (state transition) 확률, B 는 관측심볼 (observation symbol)의 출력확률, π 는 초기 상태전이 (initial state transition) 확률을 나타낸다. 한 문자에 대해 하나의 모델이 구성되고 서로 다른 n 개의 문자에 대해 대응하는 각각의 모델을 λ_i , i =1,2,.., n 으로 표기 할 때, 관측열 $O = O_1, O_2, \cdots, O_T$ 에 대응하는 미지의 문자는 식 (1)을 만족하는 모델 λ_j에 대응하는 문자로 인식된다.

$$\lambda_j = \max_{i=1..n} P(O|\lambda_i) \quad \text{---------- (1)}$$

3. 사전을 통한 오인식 보정 방법

문자들 중에는 매우 비슷한 모양을 갖고 있는 문자가 존재한다. 영문자와 숫자의 경우, 예를 들어 U-V, C-L, a-d, n-h 등과 O-o, I-1, l-1, Z-2, S-5, G-6 등은 매우 유사하다[1]. 따라서 오인식은 이러한 유사한 문자들 사이에서 일어날 경우가 많다고 볼 수 있다.

HMM 시스템에서 어떤 입력된 문자를 x 로 인식하는 것은 그 문자에 대응하는 모델 λ_X 에서의 관측열 발생 확률이 가장 높게 나왔기 때문이다. 일반적으로 오인식의 경우 유사한 모양의 문자로 오인식 할 가능성이 높기 때문에 비록 문자 x 가 문자 y 에 대응하는 모델 λ_Y 에서 발생 확률이 가장 높게 나와 오인식이 되었다고 할지라도, 모델 λ_Y 에서 원래의 문자 x 에 대응하는 관측열의 발생 확률도 여전히 높다고 볼 수 있다. 따라서 후처리기로 들어온 n 개의 문자로 구성된 단어 $Y=y_1y_2..y_n$ 이 사전에 없는 경우, 사전에 있는 단어의 집합으로부터 확률 $P(z_1z_2..z_n|y_1y_2..y_n)$을 최대로 하는 $Z=z_1z_2..z_n$ 을 보정 된 단어로 결정한다. 이때 단어에서의 각 문자는 이전의 문자 발생과 독립적이라고 가정한다면, 사전으로부터 결정하는 단어 Z 은 식 (2)를 최대로 하는 것으로 결정하면 된다. 여기서 λ_Y 는 문자 y 에 대응하는 관측열의 발생 확률이 가장 높은 모델이며, O_Z 는 문자 z 에 대응하는 관측열을 의미한다.

$$P(Z|Y)= P(z_1|y_1)P(z_2|y_2)..P(z_n|y_n) = P(O_{Z1}|\lambda_{Y1}) P(O_{Z2}|\lambda_{Y2})\cdots P(O_{Zn}|\lambda_{Yn}) \quad \cdots\cdots\cdots\cdots\cdots\cdots (2)$$

예를 들어 실제 단어인 “ABC”를 시스템이 “ABL”로 오인식 했다면 오인식 된 각 문자에 대응하는 HMM 모델들은 λ_A, λ_B, λ_L 이다. 이때 λ_A, λ_B, λ_L 모델들을 통해 사전에 있는 단어들 중에서 “ABC”에 대응하는 관측열의 발생 확률 $P(O_A|\lambda_A) \times P(O_B|\lambda_B) \times P(O_C|\lambda_L)$가 가장 높다면 시스템은 단어 “ABC”로 보정을 할 수가 있게 된다.

4. HMM 구조 및 훈련

초기 본 논문에서 사용된 HMM 모델의 구조는 left-to-right 모델이며, 상태 전이의 수는 자기 전이와 다음 상태까지만 전이 할 수 있는 구조를 갖는다. 각 문자에 대응하는 HMM 의 상태 수는 10 개로 하였고, 각 상태에서의 각 심볼의 발생

가능성은 같게 하였으며 초기 전이 확률 분포는 균일 확률 분포로 하였다.

각 문자는 인터페이스의 박스 당 한 문자씩 필기하게 하여 외부 분할 방식으로 분리되게 하였으며, 테블릿으로부터 입력되는 데이터는 먼저 거친 점 제거와 평활화 등의 전처리 과정을 거친 후 코드화 하기 위해 일정 시간 간격으로 샘플링 된 일련의 점들을 8 방향 체인코드로 만들어 사용하였다. 본 논문에서는 훈련을 위해 Baum-Welch 알고리즘을 사용하였다[2].

5. 실험 및 결과 분석

제안한 방식들의 성능을 분석하기 위해 TMC300XCi 테블릿 피시를 사용하였으며, C#과 Microsoft Tablet PC Platform SDK API 로 구현하였다. 본 논문에서 문자 훈련은 5 명의 데이터를 사용하였고, 각 사람은 각 문자를 3 번 필기하도록 해, 각 모델은 평균적으로 15 개의 해당 문자를 가지고 훈련되었다. 인식에 사용 할 단어들은 외부 분할 된 3 개의 문자로 구성된다. 또한 사전에는 100 개의 단어가 미리 정의되어 있으며, 각 단어는 3 개의 문자로 구성되고, 각 문자는 0~9, A~Z 의 36 개 문자 중 랜덤하게 선정하였다.

5.1 오인식 모델을 통한 입력 문자의 발생 확률

우선 입력된 문자가 유사한 다른 문자에 대응하는 모델로 오인식이 되었을 때, 해당 모델에서 입력된 문자에 대응하는 관측열의 발생 확률도 여전히 높게 나오는지를 조사했다. 조사 결과 36 개의 문자 중에서 3 개 문자는 오인식을 보였다. <표 1>은 오인식 된 문자들과 해당 모델에서 높은 순위로 인식된 문자들을 보여준다.

입력문자	오인식문자	인식문자 우선순위
Ø	6	6,Ø,5,8,Q
B	P	P,D,6,Ø,Q
O	6	6,O,Ø,8,Q

<표 1> 문자에 대한 오인식 결과

표 1 에서 알 수 있듯이 비록 문자가 오인식이 되었다 할지라도 해당 오인식 모델에서 원래 입력 문자가 발생할 확률은 여전히 높게 나옴을 알 수가 있다. 그러나 B 의 경우는 특수한 경우로, B 와 비슷한 D 의 발생 확률은 높게 나왔으나 B 의 발생 확률은 나오지 않았다. 이것은 아마도 훈련데이터의 부족, 모델의 간단한 환경 설정, 그리고 8 방향 체인코드만을 사용해 전체적인 인식률이 떨어져 생긴 현상으로 보인다.

5.2 사전을 통한 오인식 감소율

본 논문에서 제안하는 방법을 사용하였을 때 사전을 통한 오인식 감소율을 조사했다. 사전을 이용하지 않은 경우에 100 개의 단어 중에서 34 개의 오인식을 보인 반면, 사전을 이용한 경우에는 15 개의 오인식을 보였다. 따라서 사전을 사용하였을 경우 인식률이 29%로 증가하였다. <표 2>는 전체의 데이터 중에서 오인식이 생기는 부분에 대한 일부 데이터를 보여준다.

입력데이터	사전 미사용	사전 사용
LTØ	LT6	LTØ
MØ7	M67	MØ7
OIJ	6IJ	OIJ
OKF	6KF	OKF
MBO	MPO	결과 없음

<표 2> 사전을 이용하여 수정된 오인식 결과

<표 2>를 통해 제안한 방법을 통한 사전을 적용하였을 경우 오인식을 많이 보정 함을 알 수 있다. 다만 단어 MBO 의 경우처럼 단어에 B 문자가 포함된 경우들은 B 의 발생 확률이 없기 때문에, 사전에 있는 해당 단어 확률이 Ø 이 되어 어떠한 단어로도 찾지 못하거나 아예 틀린 단어로 보정함을 알 수가 있었다.

6. 결 론

문자 인식에서 오인식은 유사한 문자들 사이에서 일어날 경우가 많다고 볼 수 있기에,

입력된 문자가 유사한 다른 문자에 대응하는 모델을 통해 오인식이 되었다고 할지라도, 해당 모델에서는 입력된 문자의 발생 확률도 여전히 높다고 볼 수 있다. 따라서 사전을 통한 후처리 시, 오인식 된 단어에 대응하는 모델들을 통해 사전에 있는 각 단어의 발생 확률을 계산해서 오인식을 줄일 수 있음을 보였다. 향후 연구로는 각 모델에 맞는 환경 설정과 많은 훈련 데이터 등을 사용해 좀 더 정확한 결과를 도출해 낼 예정이다.

참고문헌

[1] 이성환, 문자인식- 이론과 실제, 홍릉과학출판사, 1993.

[2] L. R. Labiner, “A Tutorial on HMM and Selected Applications in Speech Recognition,” Proc. IEEE, 1989.

[3] G. T. Toussaint, “The Use of Context in Pattern Recognition,” Proc. IEEE Computer Society Conf. on Pattern Recognition and Image Processing, 1977.

[4] S. N. Srihari, J. J. Hull and R. Choudari, “Integrating Diverse Knowledge SOurces in Text Recognition,” ACM Trans. on Office Information Systems. 1983.

[5] R. Shinghal and G. T. Toussaint, “Experiments in Text Recognition with the Modified Vitgerbi Algirithm,” IEEE Trans. on Pattern Analysis and Machine Intelligence, 1979.

[6] R. Shinghal, “A Bottom-up and Top-down Approach to Using Context in Text Recognition,” Int. Journal of Man-Machine Studies, 1979.