필기한글 단어인식에서 사전정보의 효과

김 호 연(金 湖 然), 임 길 택(林 吉 澤), 남 윤 석(南 潤 奭)
한국전자통신연구원 우정기술연구부
전화 : (042) 860-1160 / 팩스 : (042) 860-6508

An effect of dictionary information in the handwritten Hangul word recognition

Ho Yon Kim, Kil Taek Lim, Yun Seok Nam
Postal Technology Development Department, ETRI.
E-mail: hoyon@etri.re.kr

Abstract

In this paper, we analysis the effect of a dictionary in a handwritten Hangul word recognition problem in terms of its size and the length of the words in it. With our experimental results, we can account for the word recognition rate depending not only on character recognition performance, but also much on the amount of the information that the dictionary contains, as well as the reduction rate of a dictionary.

I. 서론

필기 문자인식에 관한 연구는 40여년 전부터 꾸준히 수행되어져 오면서 그 인식 대상이 낱자에서 단어로 확대되고있다. 그 중 필기 한글에 관한 연구는 필기 숫자나 영어 인식 연구에 비해 상대적으로 부족하며, 주로 낱자 인식에 관한 연구에 그치고 있다. 이는 아직까지 한글낱자인식에서도 만족할 만한 결과를 얻지 못하고 있기 때문으로 보인다[1].

낱자인식기를 기반으로 동작하는 단어인식기에서, 단어인식기의 성능은 낱자인식기의 성능에 크게 의존하기 때문에 낱자인식률이 낮으면 단어인식에서 높은 인식률을 얻기는 어렵다. 수치상으로 단어인식률은 문자인식률의 단어 길이만큼의 곱이 되므로 단어의 길이가 길수록 인식률은 현저하게 떨어진다. 예를 들면, 낱자 인식률이 90%라 하더라도 3 문자로 구성된 단어의 인식률은, 문자분할오류가 없다고 가정해도 0.9*0.9*0.9 = 0.729, 즉 약 73%밖에 되지 않는다. 물론, 이 수치는 인식대상에 대한 아무런 정보가 없을 경우에 해당한다. 인식대상 단어가 제한되어 있는 경우에는 인식결과를 검증할 수 있으므로 사전을 이용하여 위의 수치보다 인식률을 높일 수 있다고 알려져 있다[2]. 그러나, 실제로 단어사전이 주는 정보량이나 단어사전의 크기와 인식률과의 관계 등에 관해서는 깊이 있는 연구가 수행되지 않았다.

본 논문에서는 한글 단어인식에서 사전의 크기, 즉 인식대상 단어의 수와 인식성능과의 관계를 분석하고 제한된 단어사전의 효과를 평가하고자 한다. 특히, 사전의 크기, 단어의 길이, 사전이 갖는 정보량, 인식률 등이 갖는 의미와 이들의 상관관계를 분석하고 이를 바탕으로 한글단어인식의 가능성을 평가하고자 한다.

논문의 구성은 다음과 같다. 2 절에서는 필기한글단어인식 문제를 정의하고, 3 절에서는 본 논문에서 이용한 한글낱자인식기를 소개한다. 4 절에서는 개별낱자인식결과와 단어사전을 이용하여 한글단어를 인식하는 단어사전검증방법을 설명한다. 5 절에서는 실험내용과 분석결과를 소개하고, 6 절에서 결론과 향후 연구방향으로 끝을 맺는다.

II. 필기한글단어인식

필기한글 단어인식에 관한 연구는 개별 문자를 중심으로 단어를 인식하는 낱자기반 인식방법과, 단어 전체로부터 특징을 추출하고 이를 모델링하여 인식하는 단어기반 인식방법으로 구분할 수 있다. 낱자기반인식방법은 단어로부터 낱자를 분할할 때 발생하는 오류를 극복해야 한다는 난점이 있지만 모델링 단위가 상대적으로 작고, 문자의 조합을 통해서 다양한 단어를 생성할 수 있기 때문에 다양한 응용분야에 적용이 쉬운 장점이 있다. 이에 반해서 단어기반인식방법은 개별문자 분할오류는 피할 수 있으나 단어 자체를 모델링해야 하므로 이를 위한 특징추출이 쉽지 않다는 것과 응용분야에 따라서 매번 새롭게 모델링해야 한다는 단점이 있다.

인식대상에 따라서 단어기반인식방법이 더 효과적으로 적용되는 분야도 있겠지만, 다양한 분할후보를 고려할 경우 문자분할의 오류를 줄일 수 있고, 이미 개발된 낱자인식기를 활용하여 단어인식기를 만들 수도 있다는 장점이 있기 때문에 본 논문에서는 낱자기반인식방법으로 필기한글단어인식문제에 접근하고자 한다. 이러한 관점에서는, 필기한글단어인식의 과정을 먼저 단어에서 문자를 분할하여 개별문자를 인식하고, 그 결과를 사전정보를 활용하여 검증한 후에, 최종적으로

주어진 사전에 있는 단어 중에서 가장 확률이 높은 단어를 선택하는 것으로 요약할 수 있다.

실제 필기한글 단어인식 실험을 위한 한글단어영상은 낱자영상을 결합하여 이용하였다. 이는 사전효과를 분석하는데 불필요한 문자분할문제와 다양한 단어사전에 해당하는 필기단어영상을 수집해야 하는 어려움을 피하기 위해서이다.

III. 한글 낱자 인식기

단어인식에 사용된 낱자인식기는 우편 주소의 행정구역 주소계층에서 사용되는 한글 467자를 인식대상으로 하여 개발되었다. 인식기는 MLP 신경망으로 만들어졌으며, 학습에는 역전파 오류알고리즘을 이용하였다. 입력 특징은 그물망 특징과 기울기 특징을 각각 추출한 후 이들을 하나의 입력벡터로 결합하여 인식기의 입력 특징으로 하였다. 각 특징의 차원은 144차원이고, 결합된 특징은 288차원이다. MLP신경망 인식기의 구조는 하나의 은닉층을 가지는데 288개의 입력노드와 100개의 은닉노드 그리고 출력노드는 467개로 되어있다. 각 낱자의 특징벡터에 대한 목표벡터는 입력 낱자의 클래스에 할당된 출력 노드는 1이고 나머지는 0으로 하였다. 본 논문의 연구에서 낱자인식기는 중요한 논의대상이 아니므로 높은 성능의 인식기를 구현하기보다는 실험의 편의를 위해서 기존에 알려진 MLP 학습방법들을 그대로 적용하여 간단한 구조로 구현하였다.

IV. 단어사전 검증기

낱자인식을 기반으로 한 필기한글 단어인식에서 단어를 검증하기 위한 한가지 방법은 인식된 단어의 후보를 생성하고 후보 중에서 사전에 없는 것을 제거하는 것이다. 이러한 방식에서는 인식된 후보단어가 모두 사전에 없는 단어인 경우 인식 결과를 얻을 수 없다. 특히 사전의 크기가 작고 문자의 인식률이 낮은 경우에는 후보 문자만으로 사전에 있는 단어를 생성하지 못할 수 있기 때문에 단어사전 검증의 효과가 떨어지게 된다. 이러한 문제를 최소화 하기 위해서는 단어 후보 수를 늘려서 가능한 후보 중에 사전에 있는 단어가 포함될 확률을 높여야 한다. 물론, 후보 수를 늘리게 되면 단어검증시간도 같이 증가하게 된다.

단어 검증의 또 다른 방법은 인식된 결과를 단어사전에 있는 단어와 매칭하되 미리 정의된 문자간의 매칭거리, 혹은 유사도를 이용하여 각 단어의 적합도를 평가하는 것이다. 이 방법을 이용하기 위해서는 문자간의 유사도가 정의되어 있어야 한다. 이 방법은 사전에 있는 모든 단어의 유사도를 계산해야 하므로 단어사전의 크기가 작을 경우에는 효율적이나, 단어사전이 커질 경우 시간이 많이 걸린다는 단점이 있다.

전자를 사전검증법이라 하고 후자를 유사도측정법이라고 부른다면, 본 논문에서는 사전검증법을 변형하여 사용하였다. 매칭을 효율적으로 수행하기 위해서 인식된 문자를 단어로 확장해나가는 과정에서 사전의 부분적 매칭을 도입하였다. 이를 위해서 트리 구조의 단어사전을 이용하였으며 각 매칭 단계에서 사전에서 제거된 단어를 보충할 수 있도록 중간단계에서도 부분단어의 여러 후보를 고려하였다.

본 논문에서 이용한 방법을 점진적사전검증법이라고 하면, 점진적사전검증법은 각 단계에서의 후보 부분단어의 수에 따라 성능이 좌우될 수 있다. 후보 부분단어의 수가 크면 잘못 인식된 결과를 사전검증을 이용하여 교정할 수 있는 가능성이 높아지는 반면 검증시간이 많이 필요하게 되고 후보 부분단어의 수가 작으면 검증속도는 빨라지지만 사전검증의 효과는 줄어들게 된다. 후보 부분단어의 수는 인식기의 성능에 따라서, 혹은 필기된 문자의 난이도에 따라서 결정하는 것이 효과적이다. 후보 부분단어의 수에 아무런 제한을 두지 않을 경우에는 문자인식결과를 최대한 활용할 수 있으므로 주어진 정보에 대한 최적의 매칭단어를 얻을 수 있게 된다.

사전검증법이나 점진적사전검증법에서 필수적인 것은 문자인식기가 인식된 결과를 신뢰도 혹은 매칭확률로 표현된 여러 개의 결과후보로 출력해야 한다는 것이다. 사전검증의 결과는 인식기의 성능에 좌우되지만 단지 1순위의 인식결과 뿐 아니라 모든 후보 결과에 의존하므로 인식기의 성능 평가시 이러한 사항이 반영되어야 한다.

V. 실험결과 및 분석

필기한글 단어인식에서 단어사전이 인식에 미치는 효과를 분석하기 위해서는 다양한 단어사전과 이에 해당하는 단어영상이 필요하다. 아직까지 데이터베이스로 잘 구축된 단어영상을 구하기가 어렵기 때문에 본 논문의 실험에서는 실험에 필요한 단어영상을 낱자영상으로부터 생성하여 이용하였다. 결과적으로 문자분할오류는 없다고 가정하고 실험한 것이다.

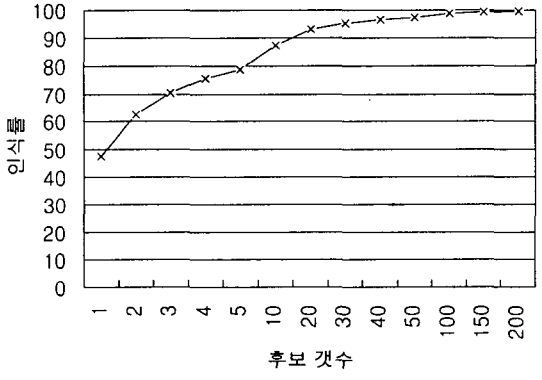

그림1. PE92데이터베이스에 대한 문자인식률
Fig. 1 Character recognition rate for PE92 DB

실험에 사용된 개별문자 영상은 한글 2350자의 100세트로 구성된 PE92 데이터베이스[3]의 이진영상을 이

용하였다. 그 중 처음 50세트는 신경망의 학습에 이용하였으며 나중 50세트는 테스트에 이용하였다. 그림1을 보면, 테스트 데이터에 대한 문자인식률은 인식대상문자를 467자로 하였을 때, 1순위가 47.6%, 2순위까지가 62.6%, 10순위까지가 87.3%로 낮은 값이다. 단어사전 검증에 이용될 인식기의 인식률은 1순위뿐 아니라 각 순위의 인식결과가 모두 중요하다. 따라서, 그래프가 전체적으로 위쪽에 존재하는 것이 좋은 성능을 보이게 된다. 사전의 축소율이 커질수록 아래쪽 순위의 인식결과가 단어인식 성능에 많은 영향을 끼치게 된다.

인식실험을 위한 한글단어사전은 랜덤으로 생성하였다. 예를 들면, 세 개의 문자로 구성된 500단어 랜덤사전을 만들기 위해서는 랜덤으로 단어를 생성하여 중복되지 않은 500단어를 선택한다. 참고적으로, 세 문자로 생성 가능한 모든 단어의 수는 467*467*467개이다. 실제 실험에서 단어사전의 크기는 인식대상문자 467자를 기준으로 2배수씩 증가시켰으며, 사전에서 단어의 길이는 2자부터 7자까지 다양하게 하였다. 실험을 위한 한글단어영상의 생성은, 일반적으로 한 단어는 한 사람에 의해 쓰여지므로, PE92데이터베이스의 50개의 파일 각각으로부터 단어사전의 모든 단어에 대해서 50세트를 반복 생성하여 실험하였다.

그림2는 사전의 크기별로 단어의 길이에 따른 인식률의 변화를 보이는 실험결과이며, 그림3은 동일한 결과를 단어길이별로 사전의 크기 변화에 따라 표현한 것이다. 예측한 바와 같이 사전의 크기가 작아질수록, 그리고 같은 사전의 크기에서는 단어의 길이가 길수록 인식률이 높아짐을 알 수 있다.

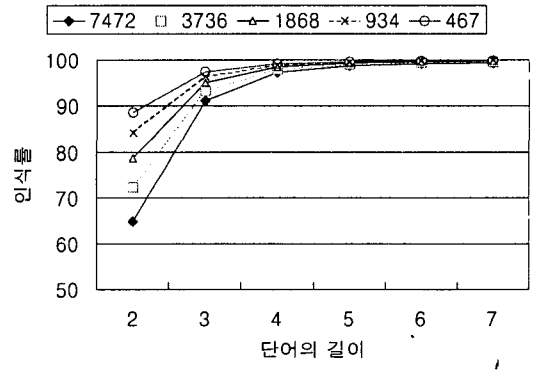

그림2. 사전 크기별 단어의 길이에 따른 인식률 변화
Fig. 2 Recognition rate as a function of the length of a word

사전의 크기가 작아질수록 인식률이 높아지는 이유는 오인식 됐을 때 결과가 사전에 없을 경우 이를 만회할 기회가 생기기 때문이고, 사전의 크기가 동일할 때 단어의 길이가 길수록 인식률이 높아지는 것은 사전의 축소율이 커지기 때문이라고 할 수 있다. 사전의 축소에 따른 오류감소율을 이용하면 이를 좀 더 명확하게 설명할 수 있다.

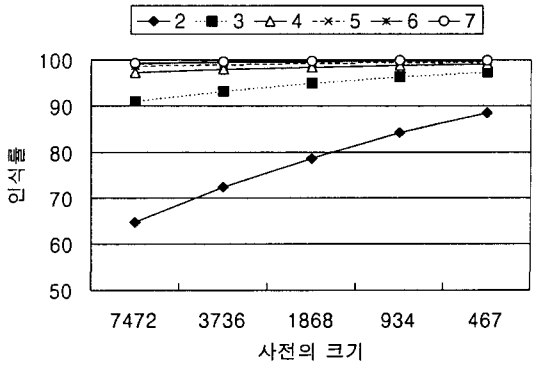

그림3. 단어길이별 사전 크기에 따른 인식률 변화
Fig. 3 Recognition rate as a function of a dictionary size

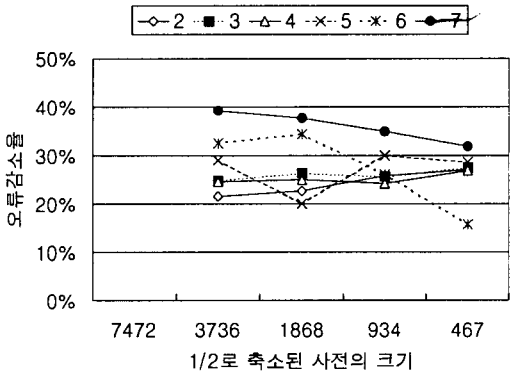

그림4. 사전 축소에 따른 오류감소율
Fig. 4 Error reduction rate when dictionary size is set to be 1/2 of its original size

그림4에서 사전의 크기나 단어의 길이에 따라 약간의 차이가 있으나 사전의 크기가 반으로 줄어듦에 따라 대체로 20 ~ 40%의 오류가 감소함을 알 수 있다. 길이가 2인 7472개의 단어를 3736개의 단어로 줄일 경우에는 오류감소율이 약 21.6%이다. 이를 바탕으로 두 글자로 표현 가능한 모든 단어 467*467개를 갖는 단어사전으로부터 사전의 크기를 반으로 줄여가며 426개의 단어가 될 때까지, 각 단계의 오류감소율이 21%라고 가정했을 때의 인식률을 이론적으로 추정한 결과가 그림5이다. 물론, 실제 오류감소율은 단어인식결과에서 각 후보인식률에 따라 각 단계에서 다소 차이가 있을 것이다. 그림4의 초기 인식률은 문자인식률 47.6%를 두 번 곱한 값인 22.7%이며, 사전을 반으로 줄일 때 21%의 오류가 줄어든다고 가정하고 계산하면, 426단어 사전에서는 약 90.7%의 인식률이 된다. 이 값은 그림1의 47.6% 낱자인식결과로부터 그림2의 467단어에 대해 실제로 얻어진 인식률 88.47%와 비슷하다. 이를 통해서 우리는 비교적 낮은 인식률을 보이는 낱자인식기를

이용하여 높은 단어인식률을 얻는 과정을 설명할 수 있다.

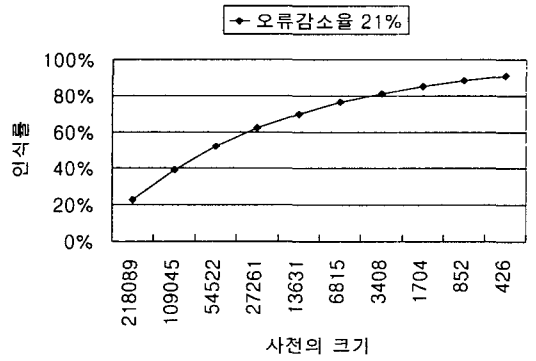

그림5. 오류감소율이 21%일 때, 사전을 1/2크기로 축소해감에 따른 인식률 추정치

Fig. 5 Expected recognition rate as a function of dictionary size when the error reduction rate is 21%

사전에서 단어의 길이가 길어지면 실제 표현 가능한 사전의 크기는 문자의 수를 곱한 것만큼 커지게 된다. 이 상태에서 사전의 크기를 동일하게 유지하면, 결과적으로 사전의 축소율은 높아지므로 그에 해당하는 만큼의 오류를 줄일 수 있게 된다. 따라서, 동일한 사전의 크기에서는 단어의 길이가 길수록 인식률이 높아진다.

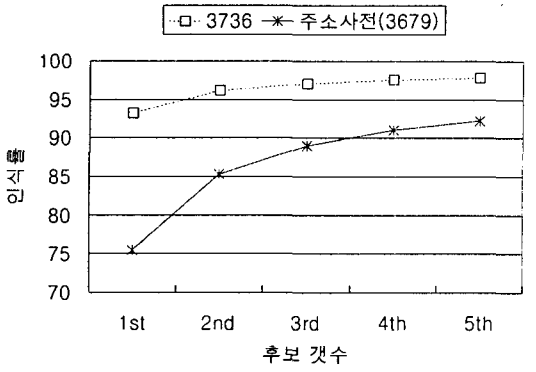

그림6. 랜덤 단어사전을 이용했을 때와 주소 단어사전을 이용했을 때의 인식률 비교

Fig. 6 Comparison of the recognition rates with a randomly generated dictionary and a dictionary obtained from the words in real addresses

그림6은 랜덤 단어사전을 이용한 결과와 우편주소에 나타나는 단어 중 길이가 3인 것만을 모은 사전을 이용한 결과를 비교한 것이다. 결과를 보면, 우편주소사전의 단어의 수가 약간 적음에도 불구하고 인식률은 오히려 현저히 떨어짐을 알 수 있다. 이는 인식과정에서 랜덤 단어사전으로부터 얻을 수 있는 정보량이 우편주소사전으로부터 얻을 수 있는 정보량 보다 많기 때문이다.

예를 들어 주소 단어사전의 화도읍 ~ 화현면 까지의 40개 단어 중에서 '화?읍'으로 된 단어가 5개, '화?면'이 10개 '화?동'이 19개, '화?군'이 3개, '화?리'가 3개 등으로 나타난다. 또한, 화암동, 화양동, 화양면, 화양읍, 화영동와 같이 두 자가 동일한 유사한 단어가 많이 존재한다. 이와 같이 주소에서 사용된 단어에는 중복성이 많기 때문에 사전으로부터 얻을 수 있는 정보가 랜덤 사전으로부터 얻는 것보다 훨씬 작다. 따라서, 동일한 사전 크기와 단어길이를 이용함에도 불구하고 인식률이 떨어지는 것이다.

VI. 결론 및 추후 연구

본 논문에서는 한글 단어인식에서 단어사전의 효과를 평가하기 위해서 단어사전의 크기와 인식성능과의 관계를 분석하였다. 단어인식실험에서 비교적 낮은 인식률의 낱자인식기를 이용했음에도 불구하고 사전의 크기나 단어의 길이에 따라서 매우 높은 단어인식률을 얻었다. 사전의 크기가 작을수록 높은 인식률을 얻었으며, 사전의 크기가 같은 경우에는 단어의 길이가 길수록 높은 인식률을 얻었다. 이를 요약하면 단어사전의 축소비율이 높을수록 높은 인식률을 얻었다고 할 수 있다. 단어의 인식률은 낱자인식률의 단어길이 지수승 만큼 떨어지고, 사전의 축약비율에따라 증가한다. 대부분의 문제에서는 문맥정보나 제한된 문제영역 등으로 인해 사전의 크기가 제한되기 때문에 단어의 길이가 길수록 높은 단어인식률을 얻을 수 있다.

단어사전으로부터 인식과정에 얻을 수 있는 정보량은 사전의 축소된 정도 뿐 아니라 사전 단어의 엔트로피와도 연관이 있다. 사전의 엔트로피가 높을수록 인식과정에서 많은 도움을 얻을 수 있다. 주소사전을 이용한 실험에서 사전크기가 비슷함에도 인식률이 떨어진 것은 주소사전이 주는 인식과정에서 주는 정보가 랜덤사전보다 적기 때문이다.

향후 연구방향으로는 단어실험을 실제 문제에 적용하여 깊이 있게 실험하는 것과, 엔트로피를 기반으로 사전이 주는 정보량을 평가하여 인식기의 관점에서 사전을 평가하는 방안을 마련하는 것이다.

참고문헌

[1] 김호연, "계층적 랜덤그래프 표현과 학습 및 이를 응용한 필기한글인식 시스템 개발", KAIST CSD. 학위논문 1999.

[2] 이관용, 권진욱, 이일병, "단어 수준의 음절 공기확률을 이용한 한글 주소 인식", 한국정보과학회논문지(B) 제25권 12호, 1998, pp.1758- 1768

[3] Dae-Hwan Kim, Young-Sup Hwang, Sang-Tae Park, Eun-Jung Kim, Sang-Hoon Paek and Sung-Yang Bang, "Handwritten Korean character image database PE92," IEICE Transaction on Information and Systems, Vol.E79-D, No.7, 1996, pp.943-950