한국어 시각단어재인 과정에서의 음운정보 역할 규명을 위한 계산주의적 모델[1)]

박기남*, 임희석**, 한군희***
*고려대학교 컴퓨터교육과
**한신대학교 컴퓨터정보소프트웨어학부
***백석대학교 정보통신학부
spknn@korea.ac.kr

Computational Model for Proving Phonological Information a Role in Visual Korean Word Recognition

Kinam Park*, Heuiseok Lim**, KunHee Han***
Department of computer Education, Korea University*
Information and Software, Hanshin University**
Division of Information and Communication, Baekseok University***

요 약

본 논문은 인간의 언어정보처리 과정 중 시각단어재인(visual word recognition) 과정에서 음운정보와 철자정보의 역할 및 심성어휘집의 표상 형태를 알아보기 위해, 계산주의적 모델을 제안하고, 제안된 모델을 이용하여 실험하였다. 실험결과 계산주의적 모델은 한국어에 대한 시각 단어재인 시 보이는 언어현상 중 음운, 철자 이웃 크기효과(phonological and orthographic neighborhood effect)를 나타냈으며, 이를 통해 한국어 시각단어재인 과정에서 심성어휘집이 음운정보로 표상되어 있다는 것을 시사하는 증거를 보였다.

1. 서론

인간의 어휘정보처리 과정 중 시각단어재인(visual word recognition) 과정이란, 시각적 자극으로 이루어진 문자를 보고 그 의미를 파악하는 과정이다. 하지만, 인간의 언어처리는 빠르고(rapid), 무의식적(unconscious)이며, 자동적(automatic)으로 이루어지지기 때문에 그 과정을 파악하기란 쉽지 않다. 이러한 이유로 단어의 의미를 파악하는 시각단어재인 과정은 인간의 언어정보처리를 연구하는 학자들 사이에서 특히 많이 연구되고 있으며, 매우 중요한 분야 중 하나이다. 하지만, 최근 까지도 시각단어재인을 설명하는 여러 이론이 대립하고 있다. 또한 계산주의적(computational model) 모델을 이용한 시각단어재인 연구는 대부분 영어나 외래어를 대상으로 한 연구가 대부분이다. 이러한 이유들로부터 본 연구의 의의를 찾을 수 있다.

시각단어재인 과정 연구에 있어서 여러 가지 과제 범주판단과제, 어휘판단과제 등을 사용할 수 있는데, 본 본문에서는 한국어의 특성을 고려하여 인간의 어휘판단과제(lexical decision task) 시 나타나는 이웃 크기 효과를 이용하였다. 어휘판단과제를 이용하여 어휘접속과정을 연구한 많은 연구자들은 시각정보가 의미로 대응되는 언어처리과정과 언어처리과정에 관여하는 여러 요소에 대해 밝혀냈다. 그중에서도 이웃크기효과(neighborhood effect)는 한 단어의 이웃들이 그 단어를 재인할 때 영향을 주는 것을 말한다. 여기서 이웃이란 한국어의 경우 철자체계의 언어가 아니기 때문에 이웃의 단위를 단정지울 수는 없지만, [1]에 따르면 단어의 음절이 이웃의 단위일

1) 이 논문은 2006년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (No. M10644000033 - 06N4400 - 03310)

가능성이 가장 높다고 할 수 있다. 실제 그의 실험에서 이웃크기효과는 2음절 단어의 첫음절만 조작했을 때 나타났으며, 다른 음절을 조절했을 때에는 효과가 나타나지 않았다. 따라서 본 논문에서는 한국어 어휘정보처리 과정 중 시각단어재인과정의 계산주의적 모델을 제안하고, 제안된 모델을 이용하여 음운정보와 철자정보의 역할과 어휘의 의미에 접속하기 위한 어휘목록의 표상 형태를 알아보고자 하는데, 이를 위해서 본 본문에서는 한국어의 이웃의 단위를 음절로 보고 철자와 음운의 이웃크기를 변수로 설정하고 실험하였다. 본 논문에 사용한 단어는 2음절 단어로 음운변화(자음동화)를 겪는 단어만을 사용하였다. 왜냐하면 단순히 철자 이웃만이 이웃크기 효과를 일으키는 요인이 아닐 수 있기 때문이었다.

2. 계산주의적 모델 구조

본 논문에서 제안된 모델은 [그림 1]과 같이 4개의 층을 가진 전 방향 신경회로망(feed forward network)구조로 설계하였다. [그림 1]은 제안한 모델의 구조를 보이고 있으며, 2개의 입력층과 1개의 은닉층, 그리고 2개의 출력층으로 구성하였다. 입력층은 한국어 2음절 단어의 철자정보와 음운정보를 벡터 형태로 표현한 각각 32단위(units)로 설계 하였으며, 100개의 단위로 구성된 은닉층과 완전하게 연결(fully connect) 하였다. 그리고 은닉층 단위 역시 32개의 단위로 어휘목록(lexical entry)과 단어의 의미를 표현한 60개의 단위와 완전하게 연결되도록 하였다. 그리고 입력층의 철자단위와 출력층의 어휘목록 단위는 전문에서 언급한 촉진효과, 억제효과를 위해 부분적 회귀연결을 갖도록 설계하였다.

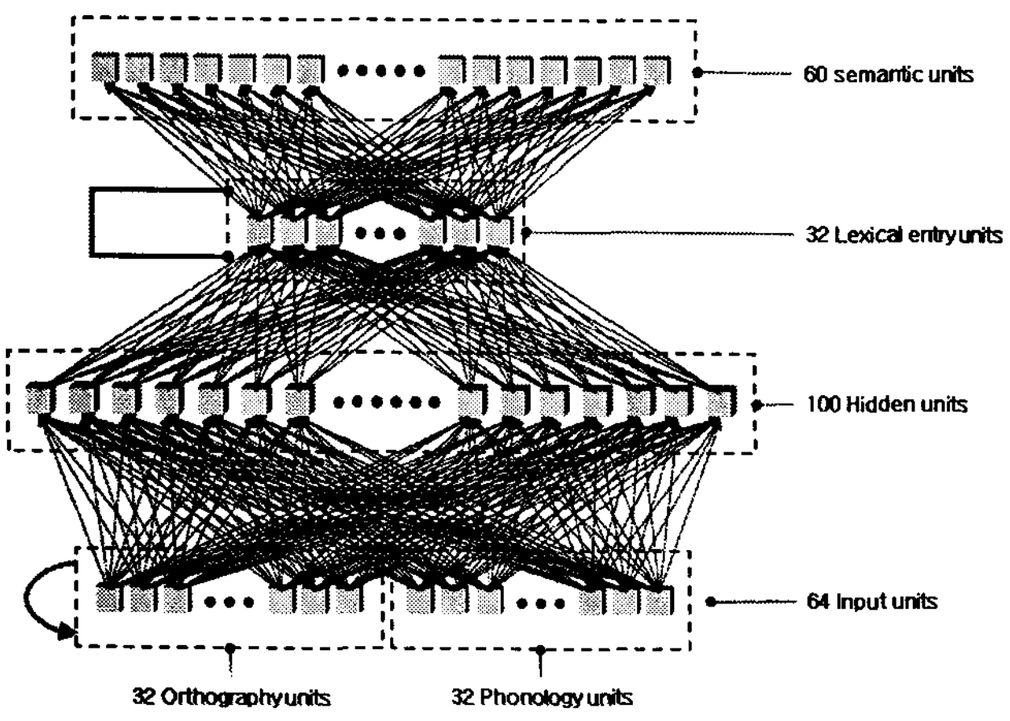

그림 1. 모델 구조

2.1 입력 설계

본 모델의 입력설계는 컴퓨터에서 가장 일반적인 언어표현 방법인 한글의 한 음절을 조합형 글자체계에서와 같이 초성, 중성, 종성에 각각 5비트(bit)씩을 할당하고, 최상위 비트를 영어 또는 한글을 표시하는 비트로 사용하여 2바이트(byte)로 표현하는 방법을 생각할 수 있다. 하지만, 본 모델에서는 추후 연구를 위해 한글뿐만 아닌 외래어 입력 설계를 고려하여 조합형 글자체계가 아니라, 16 비트의 유니코드(uni-code)를 이용하여 입력을 설계하였다. [표 1]은 한글의 모든 음절과 이에 대응하는 입력값을 나타낸 것이다.

표 1. 한국어 음절의 벡터표현 설계

유니코드로 표현 가능한 한국어 음절 (11172 자)		입력 설계 (vector)
1	가	1010110000000000
2	각	1010110000000001
3	갂	1010110000000010
:	:	:
11172	힣	1101011110100011

2.2 출력 설계

어휘판단과제를 사용한 시각단어 재인의 계산주의적 모델의 출력은 입력된 단어의 의미를 표현한다. 단어의 의미는 총 60개의 의미 자질로 구성된 벡터로 표현하였다. 출력 설계에 해당하는 의미의 표상(semantic representation)은 실제 단어의 의미를 정확하게 고려하여 설계할 수 있지만, 단순화를 위해 실제 의미를 표상하여 설계하지 않고, 입력문자와 의미간의 자의성과 단어의 의미들 간의 범주화를 고려하여 만들었다.

표 2. 단어의 의미를 표현한 출력 설계

		111111111122222222223333333333444444444455555555556 12345678901234567890123456789012345678901234567890123456789
문항	1	00000000000000000000100000000001000000000000000000110111111
물기	2	00001000000001000000000000000000000000000001011101100000000011
병오	3	00000000001011111111000000000000000000000001000000000000000000
:	:	:
가분	5941	01111111100000000000000001001000000000000000000000000000000000

2.3 계산주의적 모델의 학습 및 실험데이터

본 논문에서 제안한 모델을 학습시키기 위해 학습데이터의 입력값은 세종말뭉치 550만 어절에서 추출한 2음절단어 총 5,941개를 이용하여 만들었다. 추출된 단어는 저빈도 단어 집단에서 이웃크기가 큰 단어가 이웃크기가 작은 단어보다 재인 시간이 빨라진다는 점을 반영하기 위해서 음절의 자체 빈도를 고려하여 출현빈도 5회 이하의 저빈도 단어를 [표 1]

에서 제안한 입력패턴설계 방법을 이용하여 학습데이터를 구축하였다. 출력값은 입력된 단어의 의미표현이기 때문에 5,941개의 입력데이터와 같은 수를 [표 2]에서 제안한 의미패턴 설계구조를 바탕으로 만들었고, 입력데이터와 출력데이터는 랜덤하게 짝지어졌다. 이는 단어들 간의 철자조합 유사성과 의미간의 관계가 존재하지 않음을 나타내 주기 위한 것이다. 본 논문에서 제안하는 모델의 평가(evaluation)를 위하여 사용한 데이터는 철자와 음운 각각에 대해 자음동화를 일으키는 단어, 예를 들어 '인력'인 경우 철자를 공유하는 이웃이 많고 적음에 따라 단어를 선정하였으며, 동시에 발음은 '/일력/'으로 나기 때문에 '/일/'으로 발음되는 음운 이웃이 많고 적음으로 평가 데이터를 설계하였다. 평가를 위해 총 4종류의 데이터를 사용하였으며, 각 데이터마다 55개의 단어를 갖도록 하였다.

2.4 실험결과

본 논문에서 제안한 모델의 출력값은 의미강조값(semantic stress)값을 이용하여 분석하였다.2]. 의미강조값은 엔트로피값의 일종으로 오류값을 계산하는 방법 중 하나이다. 오류값이 크다는 것은 의미강조값이 낮다는 것이고, 어휘판단과제 시 어휘판단 반응시간이 길다는 것이며, 오류값이 작다는 것은 의미강조값이 크다는 것이고, 어휘판단 반응시간이 짧다는 의미이다. 의미강조값의 계산식은 [식 1]과 같고, 의미층에서의 출력값이 0에서 1사이의 값 중 0.5에 가까울 록 0이되고, 0이나 1에 가까울수록 1이 된다. 이는 본 모델에서 입력설계구조와 출력설계구조가 0혹은 1로 설계되었다는 점과 실험데이터가 목표값 없이 출력값과 목표값의 차이를 평균 제곱한 값을 이용할 수 없다는 점에서 모델평가 함수로써 적절하다.

$$S_j = s_j \log_2(s_j) + (1-s_j)\log_2(1-s_j) - \log_2(0.5)$$

S_j 의미강조값 s_j 의미층에서의 출력값

[식 1] 의미강조값

본 연구는 제안한 모델에 의해 한국어의 이웃의 단위를 음절로 가정하고, 철자이웃과 음운이웃의 많고 적음에 따라 이웃크기효과가 어떻게 나타나는지 알아보고자 하였다. 실험결과 음운이웃의 많고, 음운이웃의 적음의 크기조건에 상관없이 철자이웃이 큰 경우에 평가 데이터의 평균 어휘판단 반응시간이 빠르게 나타났다. 그리고 철자이웃의 많고, 적음의 차원에서 살펴보면, 크기조건에 상관없이 음운이웃이 큰 경우에 평가 데이터의 평균어휘판단 반응 시간이 느리게 나타났음을 알 수 있다 (t[20441]= 20.057, $p<0.05$), (t[20454]= 13.098, $p<0.05$). 이는 철자이웃이 큰 경우 철자이웃이 작은 경우에 비해 어휘판단 반응 시간에 대해 촉진적(facilitatory) 효과를, 음운이웃의 경우 이웃이 큰 경우가 작은 경우에 비해 어휘판단 반응 시간이 증가하는 억제적 효과(inhibitory effect)가 나타났음을 알 수 있었다.

3. 결론

본 논문은 인간의 언어정보처리 과정 중 시각단어재인 과정에서 음운정보와 철자정보의 역할 및 심성어휘집의 표상 형태를 알아보기 위해, 본 논문에서 제안한 계산주의적 모델을 통해 실험하였다.

실험결과 한국어 시각단어재인 과정에서 심성어휘집의 표상 형태가 음운정보로 이루어져 있음을 알 수 있었다. 실험 결과에 따르면 음운이웃이 크기가 크면 억제적 효과를 나타냈는데, 이는 의미에 접속하는 형태가 음운정보로 구성되었기 때문에 활성화 수준에 비례해서 다른 음운정보를 억제시키게 되는 것이다. 즉 이웃들과 연결된 노드가 목표단어보다 먼저 역치수준에 도달한다면 심성어휘집에서의 어휘경쟁원리(lexical competition principle)로 인해 활성화된 이웃들은 목표단어의 활성화를 억제한 것으로 해석된다. 그리고 철자이웃의 정보는 음운이웃 정보에 의해 활성화 되어있는 단어들을 구별하여 주는 보조적인 역할을 수행한 것으로 해석되어진다. 그러나 본 연구를 통해서는 음운정보와 철자정보 사이에 어느 정보가 더 우위에 있는지는 검증할 수 없었다. 왜냐하면 본 연구에서 제안한 모델이 음운정보와 철자정보를 동시에 사용하는 이중경로론(dual route theory)을 바탕으로 설계되었기 때문일 것으로 생각된다. 하지만 추후 연구를 통해 두 정보를 직접적으로 조작한다면 이 또한 검증할 수 있을 것이다.

참고문헌

[1] 조혜숙, "한국어 단어재인에서 나타나는 이웃효과", 고려대학교 대학원, 석사학위청구논문, pp.14-15, 2003.

[2] Plaut, D. C., "Structure and function in the lexical system: Insights from distributed model of word reading and lexical decion", Language and Cognitive Processes, 12, pp.765-805, 1997.