2002년도 한국음향학회 학술발표대회 논문집 제21권 제1(s)호

한국어 어절 재인의 시뮬레이션 모델*

임희석°, 남기춘
천안대학교 정보통신학부°, 고려대학교 심리학과

A Simulation Model for Korean Eojeol Retrieval

Heuiseok Lim°, Kichun Nam
Dept. of Information & Communications, Cheonan Univ°.
Dept. of Psychology, Korea Univ.

limhs@infocom.chonan.ac.kr

요 약

본 논문은 한국인 피험자를 대상으로 이루어진 어절 재인 실험 시 관찰된 언어 현상인 길이 효과, 빈도 효과, 그리고 이웃 효과를 설명할 수 있는 한국어 어절 재인 시뮬레이션 모델을 제안한다. 제안한 모델은 코퍼스에서 나타난 어절의 빈도를 이용하여 정렬한 트라이(trie) 구조를 기반으로 하고 있다. 본 모델은 피험자들의 어절 재인 현상을 모두 설명할 수 있으며 피험자들을 대상으로 한 실험에서 사용한 동일 자료를 이용하여 시뮬레이션한 결과 유의미한 상관 관계를 보였다. 현재 시뮬레이션 중 발견된 언어 현상이 한국인 피험자에서도 나타나는지를 규명하기 위한 실험과 영어 단어 재인시의 언어 현상에 대해서도 적용할 수 있는 확장 방안에 대하여 연구를 수행하고 있다.

1. 서론

인간의 지식 표상 규명에 대한 연구는 인간을 대상으로 연구하는 심리학에서뿐만 아니라 인간의 지능을 컴퓨터를 이용하여 구현하고자 하는 인공지능 학문에서도 오래 전부터 매우 중요한 화두가 되고 있다. 특히 인간의 지식 중 언어 지식에 대한 연구는 인간의 언어처리 과정 및 현상을 규명하고 이해하고자 하는 심리언어학에서뿐만 아니라 인간의 언어를 컴퓨터를 이용하여 처리하고자 하는 전산언어학 연구에 있어서도 매우 중요하다. 언어 지식 표상 구조에 대한 이해는 인간의 언어 이해 및 생성 과정, 언어를 통한 사고 과정에 대한 연구와 인간의 언어 처리 능력과 유사한 언어처리 시스템의 개발에 결정적인 단서를 제공할 것이다.

인간의 언어 정보 처리 과정 및 심성어휘집의 표상규명을 위하여 본 연구팀이 인간 피험자를 대상으로 실험한 결과 심성어휘집내에는 형태소 단위의 사전과 어절 단위의 사전이 모두 공존하는 것으로 나타난다. 자연어처리 시스템을 구현하기 위한 전자 사전도 형태소 단위의 사전과 어절 단위의 사전이 사용될 수 있는데, 인간의 심성어휘집내의 사전도 이처럼 두 가지의 형태의 사전이 모두 사용되는 것으로 추정된다.

형태소 단위의 사전을 사용할 경우 형태소 단위로 정보를 가지고 있으므로 기억 공간을 적게 사용하나 어절을 분석하기 위하여 형태소 단위의 분리 및 결합 여부를 판단하기 위한 처리 과정이 필요하다. 어절 단위의 사전을 사용할 경우 어절의 가능한 분석 결과를 사전에 저장하고 있으므로 분석하고자 하는 어절의 탐색 과정으로만 어절 분석이 이루어질 수 있어 빠른 정보처리가 가능하나 많은 기억 공간을 사용하여야 하는 특징을 가지고 있다. 이러한 특징은 심성어휘집내의 사전의 종류에 따른 정보처리 과정의 차이로 인한 것이며 형태소 사전이나 어절 사전내에서의 표상 구조에 따른 차이는 알 수가 없다. 사전내의 언어 지식의 표상 방법 및 구조는 언어처리의 시간 및 정보 처리 과정을 결정짓는데, 중요한 역할을 한다. 하지만 외래어에 대한 연구와는 달리 한국어에 대한 심성어휘집의 표상 구조에 대한 연구는 미흡한 편이다.

본 연구는 심성어휘집내의 어절 사전의 표상 구조에 대해

* 본 연구는 한국과학재단 목적기초연구(R01-2000-00407)지원으로 수행되었음.

서 논의하며 심성어휘집의 어절 사전의 시뮬레이션 모델에 대한 것이다. 본 연구팀이 인간을 대상으로 실험한 한국어 어절 재인 실험 결과, 어절의 재인 시 길이 효과와 빈도 효과가 나타남을 확인할 수 있었으며 본 연구는 이러한 결과를 반영하는 심성어휘집의 어절 표상의 시뮬레이션 모델의 제안과 실험 결과 및 시뮬레이션 결과를 통해 추정되는 인간의 어절 재인 발생하는 언어 현상에 대해서 토의하고자 한다.

2. 어절 재인시의 언어 현상

단어 빈도와 단어 길이는 단어 재인에 관한 연구에 있어서 중요시 다루어지는 변수이다. 단어 길이는 어절에 포함된 낱자, 음절 혹은 글자로 정의될 수 있으며 단어 빈도는 학습자들에게 노출된 빈도를 의미한다.

한국어 단어 재인에 관한 연구 결과에 의하면 어절 재인 시 길이 효과(length effect)와 빈도 효과(frequency effect)가 나타나는 것으로 제시되었다[1, 5, 6, 7]. 어절 재인시의 길이 효과와 빈도 효과를 정의하면 다음과 같다. 정의 1과 정의 2에서 LDT는 어휘 판단 과제(lexical decision test) 수행시의 시간을 의미하고, freq(x)와 len(x)는 각각 어절 x의 빈도와 길이를 계산하는 함수이다.

정의 1 : 빈도 효과

$$LDT(x) < LDT(y)$$

$$where\ len(x) \geq len(y)\ \text{and}\ freq(x) > freq(y)\ \blacksquare$$

정의 2 : 길이 효과

$$LDT(x) < LDT(y)$$

$$where\ len(x) < len(y)\ \text{and}\ freq(x) \leq freq(y)\ \blacksquare$$

정의 1의 빈도 효과는 어절 x의 길이가 y의 길이보다 크거나 같은 경우라도 빈도가 높은 어절 x의 어휘 판단 시간이 더 빠름을 의미하고, 정의 2의 길이 효과는 빈도가 높거나 같은 어절 x일지라도 어절의 길이가 짧은 x의 어휘 판단 시간이 더 빨라지는 것을 의미한다.

3. 언어처리를 위한 전자 사전

본 논문은 인간 피험자를 대상으로 한 어절 재인 실험 결과 관찰된 언어 현상을 시뮬레이션할 수 있는 전자 사전 구조를 개발하고, 이를 이용한 시뮬레이션 결과를 토의하는 것이다. 이를 위하여 기존의 전자 사전 구조에 대한 이해가 필요하며 다음은 현재 언어 처리 프로그램에서 많이 사용되고 있는 전자 사전에 대하여 설명한다.

전자 사전(machine dictionary)이란 언어처리 프로그램이 형태소 또는 어절의 정보 획득을 위해서 사용하는 사전으로 컴퓨터 입장에서의 인간의 심성어휘집과 같은 역할을 하며 기계 가독형 사전(machine readable dictionary)라고도 한다. 전자사전 구축시의 최대 관건은 빠른 시간에 원하는 정보를 접근할 수 있도록 하는 접근성과 저장 공간의 효율성이다. 이와 같은 이유로 등재되는 어휘의 수준이나 사전의 구성이 언어학적 관점에서의 사전과 다소 차이가 있을 수 있다.

현재 전자 사전 구현을 위하여 많이 사용되고 있는 구조는 해싱(hashing), B-tree, FST(finite state tranducer), 그리고 Trie 등이다[2, 3].

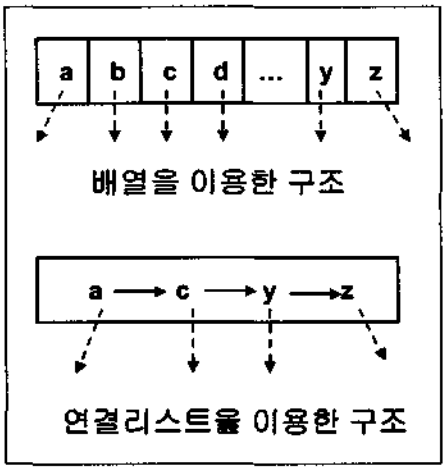
해싱은 저장할 어휘 문자열을 숫자 형식으로 변환하여 그 문자열이 저장될 주소 공간을 생성하는 해쉬 함수를 사용하여 생성된 주소 공간에 어휘를 저장한다. 검색 시에는 사전 구성 시 사용했던 동일한 해쉬 함수를 사용하여 탐색할 주소 공간을 생성하여 해당 위치를 탐색한다. 해싱 방법은 완전 해쉬 함수(perfect hash function)을 사용하는 경우 사전에 저장된 어휘의 개수에 영향을 받지 않고 시간의 복잡도 O(1)만으로 사전을 탐색할 수 있다는 장점을 갖는 반면 어휘의 삽입 삭제 시 해쉬 테이블을 다시 구성해야 하는 부담이 따르게 된다.

B-tree 방식은 B-tree의 노드에 어휘를 색인하는 구조로 저장 공간을 효율적으로 사용하기 위하여 가변형 B-tree 구조가 많이 사용되며 어휘의 삽입 삭제가 용이한 반면 사전 탐색의 시간 복잡도가 어휘의 개수에 따라 증가하게 된다.

FST 방식은 FST를 이용하여 완전 해쉬 함수를 만들고, 이를 이용하여 어휘를 저장하거나 탐색하는 방식으로 대량의 사전 표제어에 대해서도 사전의 크기가 많이 커지지 않고, 탐색 속도도 표제어의 수에 영향을 받지 않는다는 장점이 있지만 해싱과 같이 표제어의 삽입, 삭제 시 사전을 재구성해야 한다는 부담이 따른다.

Trie는 검색을 의미하는 'reTRIEval'에서 이름을 따온 구조로 표제어를 구성하는 기본 문자를 포함하는 정점으로 구성된 트리 구조라고 설명할 수 있다. 예를 들어 영어 단어는 키가 26개의 알파벳으로 구성되어 있다고 가정할 수 있으며 영어 어휘를 구성하기 위한 trie의 정점의 구조는 [그림 1]과 같이 배열을 이용한 방법과 연결 리스트를 이용한 방법을 사용할 수 있다. [그림 1]에서 점선은 해당 문자로 시작되는 표제어를 포함하는 영역의 포인터이고, 실선은 한 노드내에 사

용된 알파벳을 연결하는 포인터를 의미한다. 배열과 연결리스트를 이용한 두 구조 중 배열을 이용한 구조가 표제어의 탐색 시간이 연결리스트를 사용한 구조보다 빠른 반면 많은 저장 공간을 요구한다는 단점이 있다. 일반적으로 Trie를 이용한 방법은 사전 탐색 시간이 표제어 양에 영향을 받지 않고, 문자열의 길이를 k라 할 때 시간의 복잡도 O(k)의 계산량으로 표제어의 검색 및 삽입이 가능하다.



[그림 1] Trie 정점의 구조

4. 빈도 기반 Trie

앞 장에서 살펴본 전자 사전의 구조는 사람들의 어절 재인 시 보이는 길이 효과, 빈도 효과, 이웃 효과를 제대로 설명할 수 없으며, 그나마 문자열의 길이에 따라 표제어 탐색 시간이 좌우되는 trie 구조는 길이 효과를 반영한다고 할 수 있다.

본 논문은 trie 구조를 이용하여 나머지 효과를 설명할 수 있는 빈도 기반 trie(frequency based trie) 구조를 제안한다. 빈도 기반 trie는 아래의 조건을 만족하는 trie 구조이다.

1) 정점의 구조는 연결형 리스트를 사용한다.
2) 정점의 알파벳은 자소와 음소의 집합이다.
3) 정점내의 알파벳은 코퍼스내의 빈도에 따라 오름차순으로 정렬되어 있다.

일반적으로 trie 구조는 정점내에 알파벳이 알파벳의 순서에 따라 정렬되어 있으나 빈도 기반 trie는 각 정점에서 사용되는 알파벳이 빈도에 따라 정렬되어야 한다. 알파벳의 빈도는 코퍼스 상에 나타난 어절의 빈도에 따라 계산되며 이는 사람들이 언어를 학습하면서 자주 접하게 되는 어절의 빈도를 모델링하기 위함이다. 이는 학습자가 한국어 학습을 시작한 초창기에는 대부분의 어절의 빈도가 유사하여 어떤 어절이든지 어절 재인 시간이 비슷하나 학습 과정이 진행되면서 자주 보고 사용하게 된 어절, 즉 고빈도 어절인 경우에는 어절의 재인 시간이 빨라지는 현상을 모델링하기 위해서이다.

빈도 기반 trie의 정점을 구성하기 위한 자료 구조를 C언어를 이용하여 나타내면 [표 1]과 같고, 코퍼스를 이용하여 빈도 기반 trie를 구성하는 과정은 [표 2]와 같다.

[표 1] 빈도기반 trie의 정점 구조

```
typedef struct node {
        char ch;
        int freq;
        struct node *next;
        struct node *child;
        short int isfinal;
} nodetype;
```

[표 1]에서 'ch'는 알파벳을 나타내고 'freq'는 'ch'의 빈도를 나타내는 변수이다. 'next'와 'child'는 각각 한 정점 내의 다음 알파벳과 자식 정점을 나타내는 변수를 의미하며, 'isfinal'은 현재까지의 문자열이 어절을 의미하는지 또는 어절의 중간 부분인지를 나타낸다.

[표 2] 빈도기반 trie 구축 과정

1) 코퍼스를 분석하여 어절의 빈도를 계산
2) trie에 등록할 어절의 각 음절을 3바이트 코드로 변환
3) 어절과 어절의 빈도를 trie에 저장
4) 3)에서 구축된 trie의 각 정점내의 알파벳을 오름차순으로 정렬

5. 실험 결과

본 논문은 빈도 기반 트라이 구축을 위하여 천만 어절의 원시 코퍼스를 사용하였다. 빈도 기반 트라이 사전은 학습자의 학습 시간의 따른 어절 재인의 양상을 시뮬레이션하기 위하여 천만 어절 코퍼스를 1백만 어절, 500백만 어절, 1000만 어절 크기 나눈 3가지의 코퍼스를 이용하여 빈도 기반 트라이 사전을 구축하였다.

어휘 판단 과제의 시간 측정은 개별 단어 재인에 필요한 시간이 너무 짧아 초(second)단위로 나타내기 어려웠으며, 한 어절을 재인하기 위해서 방문하게되는 트라이 상에서의 정점의 개수를 시간 측정을 위한 단위로 사용하였다. 실험 자료는 [5]에서 사용한 실험 자료를 사용하였으며 실험 결과 천만어절 크기의 코퍼스를 이용한 경우, 피험자의 어휘 판단 시간과 어절의 길이, 어휘 판단 시간과 빈도간의 상관 관계가 가장 큰 것을 확인할 수 있었다.

시뮬레이션 실험에서 흥미로운 새로운 현상이 발견되었는데, 이는 동일한 빈도와 동일한 길이의 어절 x와 y의 재인 시간의 차이가 나타나는 것이다. 실험 결과를 분석해 본 결과 어절에 고빈도의 형태소 또는 음절을 포함한 어절일수록 재인 시간이 더 짧아짐을 확인할 수 있었다. 이러한 결과는 어절 재인에 영향을 미치는 변인이 어절의 길이와 빈도뿐만 아니라 어절을 구성하는 형태소나 특정 음절의 빈도도 어절 재인에 영향을 주는 변인으로 작용할 수 있음을 시사한다. 현재 이러한 이웃 효과(neighboring effect)가 피험자들에게도 보이는지 실험을 통하여 확인하고자 하며, 실험을 통하여 이웃 효과가 있음이 밝혀지면 본 논문에서 제안하는 어절 표상 구조가 인간의 어절 표상을 올바르게 모델링한 것임을 지지하는 결과가 될 것으로 사료된다.

6. 결론 및 토의

본 논문은 어절 재인 시 피험자가 보인 빈도 효과, 길이 효과를 설명할 수 있는 전자 사전 구조를 제안하고 그 결과를 제시하였다. 제안된 전자 사전 구조는 기존의 트라이 구조의 정점들이 코퍼스내의 어절 빈도를 이용하여 오름차순으로 정렬되어 있는 구조로 어절의 길이 효과와 빈도 효과를 자연스럽게 설명할 수 있으며 시뮬레이션 결과 인간 피험자들의 실험 결과와 동일한 결과를 보였다.

실험 결과, 어절 재인 시의 언어 현상뿐만 아니라 인간의 형태소 학습 원리를 일부 설명할 수 있는 결과를 얻을 수 있었다. 자주 사용되는 형태소를 포함하는 어절은 여러 형태로 코퍼스에 나타나지만 빈도 기반 트라이 사전에서 해당 형태소까지의 알파벳의 빈도가 매우 높고 기능어가 시작되는 부분의 알파벳에서 빈도가 현저히 저하되는 것을 확인할 수 있었다. 이는 인간이 어절 단위로 학습하다가 특정 빈도 이상이 되는 부분 문자열을 형태소 또는 신조어로 학습할 수 있다는 것을 의미하며, 본 연구팀은 이를 입증하기 위한 실험을 수행하고자 한다. 또한 시뮬레이션을 통하여 확인하게된 이웃 효과에 대한 실험을 준비 중에 있다.

현재 언어 심리학을 중심으로 인간의 심성 어휘집의 표상 구조 규명을 위한 노력이 진행 중이며 본 논문은 심성 어휘집 내의 어절 표상 구조만을 다루었다. 아직 한국어 언어 처리를 위한 심성어휘집의 구성 단위가 형태소인지 어절인지 또한 형태소와 어절 단위가 모두 사용되는지 확실히 밝혀지지 않고 있으나 본 연구팀의 연구 결과 형태소와 어절 두가지 단위가 사용되고 있음이 일부 확인되고 있다. 향후 형태소의 표상 구조를 포함한 심성 어휘집 전체의 표상 구조 규명에 대한 연구를 계속 수행할 것이며, 이를 이용한 인지 신경기반 전자 사전을 개발하고자 한다.

참고 문헌

[1] Cosky, M. J., The role of letter recognition, Memory and Cognition, Vol. 4, 204-214, 1976.

[2] E. Horowitz, S. Sahni, S. A. Freed, Fundamentals of Data Structures in C, Computer Science Press, 1998.

[3] R. Dale, H. Moisl, H. Somers, Handbook of Natural Language Processing(edited), Marcel Dekker, Inc., 2000.

[4] Foster, K. I., Chamber, S. M., Lexical acess and naming time, Journal of Verbal Learning and Verbal Behavior, Vol. 12, pp. 627-635, 1973.

[5] Whaley, C. P., Word-nonword classification time, Journal of Verbal Learning and Verbal Behavior, Vol. 17, pp. 143-154, 1978.

[6] 남기춘, 서광준, 최기선, 한글 단어 재인에서의 단어 길이 효과, 한국인지과학회지:실험 및 인지, Vol. 9, No. 2, 1-18, pp. 1-18, 1997.

[7] 남기춘, 최기선, 시각 단어 재인에서의 단어 빈도 효과: 단어 빈도 효과의 본질과 단어 빈도가 영향을 주는 정보처리 과정, 한국과학기술원 연구센터 Technical Report, CAIR-TR-97-66.