인지신경기반의 한국어 어휘습득 계산주의적 모델

유원희*, 박기남**, 류기곤*, 임희석*, 남기춘***
*한신대학교 컴퓨터정보소프트웨어학부
**고려대학교 컴퓨터교육학과
***고려대학교 심리학과
e-mail:galadous@naver.com

Cognitive-Neuro Computational Model of Lexical Acquisition in Korean

Wonhee Yu, Kinam Park, Kigon Lyu, Heuiseok Lim, Kichun Nam
Division of Computer, Information and Software, Hanshin University
Department of Computer Education, Korea University
Department of Psychology, Korea University

요 약

본 논문은 인간의 어휘획득(Lexical Aquisition)과정을 하이브리드(hybrid)한 형태의 계산주의적(Computational) 모델을 설계,반복 실험을 통해 인지신경기반의 어휘습득 모델을 구현하고 실험하였다. 이 연구를 통해 인간의 어휘획득 과정을 모사(simulate)할수 있었고, 이로인해 인지신경기반 어휘 정보처리 시스템 개발을 위한 자동어휘 획득, 심성 어휘집 표상, 어휘 인식(word recognition)의 계산주의적 모델 개발에 기여할 수 있을 것이다.

1. 서론

최근 인간의 인지 기능 원리 규명에 대한 지적 관심 증대, 지능형 로봇 개발을 위한 뇌과학의 필요성 증대 등으로 인지신경과학 분야의 연구와 중요성이 날로 증가하고 있는 실정이다. 이에 대뇌의 정보처리 규명과 뇌에 대한 이해와 모델링, 그리고 이를 지능형 공학 시스템에 접목하기 위해서는 인지과학과 신경 과학의 접목 뿐만 아니라 신경 심리학(neuropsychology), 컴퓨터과학(computer science), 전자과학(electronic engineering), 통계학, 물리학 등의 접목을 통한 다학제적 연구(multidisciplinary)가 필요하다.

본 연구는 인지신경과학과 컴퓨터과학이 접목된 인지신경계산학(cognitive neuro-computational study)의 한 분야라 할 수 있다.본 연구는 인지신경계산학은 대뇌에서 일어나는 인지 기능의 정보처리적 원리, 구조적 원리 등을 추상화를 통하여 계산주의적 모델을 개발하고 구현하며, 어휘획득모델의 제시와 대뇌의 어휘 표상 연구(mental lexicon representation) 모델을 제시하여 인간의 언어정보처리 과정의 이해를위한 시뮬레이션과 심도 있는 연구를 가능하게 할 수 있다.

본 연구는 인지신경기반 어휘 정보처리 시스템 개발을 위한 자동어휘 획득, 심성 어휘집 표상, 어휘 인식(word recognition)의 계산주의적 모델 개발에 기여할 수 있을 것이다.

1.1. 한국어 심성어휘 표상

심성 어휘집에 형태소, 단어, 어절이 어떻게 표상되어 있을 것인가에 대한논의는 다양하게 전개되어 왔다. 현재 심성어휘집 표상에 대한 연구는 결합 모델(full-list model), 분해모델(decomposition model), 하이브리드 모델(hybrid model)로 구분할 수 있다

[1][2][3][4][5][6].

결합 모델은 활용(inflection)과 파생(derivation) 등으로 여러 개의 형태소로 결합된 단어나 어절이 언어 생활에서 사용되는 형태 그대로 심성어휘집에 저장되고, 단어 인식(word recognition) 또는 어절 인식 시 심성어휘집에 등록된 형태대로 탐색이 이루어진다는 모델이다. 분해 모델은 결합 모델과는 달리 형태소와 같은 단어나 어절을 이루는 하위 단위의 어휘 정보가 심성어휘집에 저장되어 있으며, 단어 인식 시 입력된 단어를 심성어휘집에 저장된 단위로 분해하여 분해된 각 단위를 사전에서 탐색하고 탐색해서 얻은 정보를 통합하여 단어를 인식한다는 모델이다. 하이브리드 모델은 결합 모델과 분해 모델 모두가 심성어휘집의 표상과 탐색에 적용된다는 모델이다.

본 논문은 하이브리드 모델을 바탕으로 하는 계산주의적 어휘획득 모델을 제안한다.

2. 계산주의적 어휘획득 모델

본 논문이 제안하는 모델을 도식화한 그림은 <그림 1>과 같다.

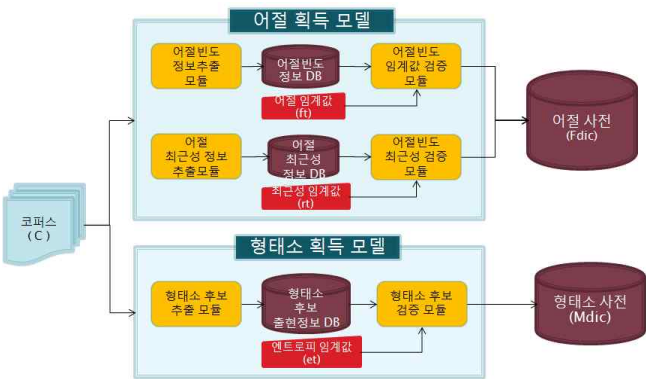

그림 1. 인지신경기반 어휘 획득 모델

제안하는 모델을 간단히 설명하면, 어절 획득 모델에서는 입력된 어절의 빈도가 어절 임계값(frequency)이상일때 그 어절은 충분히 학습이 가능한 빈도 이상의 어절로 간주하여 어절사전에 등록한다. 또한 특정 어절 크기 내에서 반복해서 출현하였는지를 어절 최근성 검증 모듈을 통하여 최근성 빈도가 높은 어절도 어절 사전에 등록한다. 형태소 획득 모델은 업력된 어절의 가능한 모든 음절 경계에서 '머리문자열'과 '꼬리문자열'의 형태로 나누고 머리문자열의 바로 다음 문자의 발생 엔트로피가 머리문자열의 발생 엔트로피(entropy)값와 꼬리문자열의 바로 전 문자열의 엔트로피값을 계산하여 엔트로피값의 변화가 형태소 가능성을 보일 경우 해당 머리문자열과 꼬리문자열을 형태소 사전에 등록 시키는 형태이다.

3. 실험 및 결과

3.1 학습 데이터

입력값인 한국어는 21세기세종계획의 문어 말뭉치(corpus)로 1천만 어절과 구어 말뭉치 85만 어절을 사용하였다.

3.2 어절사전 학습

본 논문이 제안하는 어휘 획득 모델의 결합 모델 실험을 하기 위해서는 빈도 임계값, 최근성 임계값이 결정 되어야 한다. 반복 실험을 통해서 어절 사전의 학습 형태를 분석하는데 <표 1>에서 보는 바와 같은 결과를 빈도임계값 [100]과 최근성임계값 [10]에서의 학습된 결과의 초기데이터다. 총 11,761 어절이 학습 되었고, 학습어절의 빈도의 합이 학습 코퍼스의 58%정도를 차지한다.

표 1. 어절사전 학습 결과1

번호	어절	학습시기
1	수혜는	5397
2	있었다	13219
3	있었다	18340
4	수혜가	18678
5	그	21394
6	있는	24865
7	한	28509
8	너의	32748

아래 <표 2>는 빈도임계값 [500]과 최근성임계값 [20]에서의 학습된 결과의 초기 데이터다. 총 2,097개의 어절이 학습되었으며 학습된 어절의 빈도의 합이 학습 코퍼스 1천만의 38.63에 해당하였다.

표 2. 어절사전 학습 결과2

번호	어절	학습시기
1	나는	53207
2	그	74580
3	있는	89058
4	수	108006
5	이	124014
6	있다	125613
7	것이다	154870
8	한	158764

<표 1>과 <표 2>의 결과를 비교해 보았을 때 임계값들이 500과 20으로 정해져 있을 때 초기 어절의 획득 결과가 더 좋은 결과를 보여 주는 것을 볼 수 있다.

3.3 형태소사전 학습

본논문이 제안하는 어휘 획득 모델의 분해 모델의 실험 결과중 일부가 <표 3>이다.

표 3. 형태소사전 학습결과

머리형태소	꼬리형태소	추출어절
가꾸어	야겠다	가꾸어야겠다
가능하	니까	가능하니까
가다듬	는데	가다듬는데
가두어	버렸다	가두어버렸다
가라앉	도록	가라앉도록
가레니나	조차도	가레니나조차도
가르시아	에게	가르시아에게
가만있	겠습니까	가만있겠습니까
가버렸	습니다	가버렸습니다
가버리	잖아요	가버리잖아요
가부장	적인	가부장적인
가상현실	시스템	가상현실시스템
가석방	시켜야	가석방시켜야
가속화	하는	가속화하는
가운데	에서	가운데에서
가이드	로서	가이드로서
가정주부	에서	가정주부에서
가톨릭	적인	가톨릭적인

총 3,488개의 머리형태소와 꼬리형태소가 학습 되었으며, 올바른 형태소가 획득 되었는지는 <식 1>의 정확도 계산에 의해서 이루어졌으며 99.87%의 정확도를 보였다.

<식 1>형태소획득정확도 계산식

$$\text{형태소획득정확도} = \frac{\text{정확하게획득된형태소수}}{\text{획득된형태소수}}$$

4. 결론

본 논문은 인간의 언어 학습 원리를 반영한 어휘획득의 계산주의적 모델을 제안하였으며, 이를 한국어에 적용하여 실험 하였다. 제안하는 모델은 한국어에만 적용되는 모델이 아니라 다국어 어휘획득 모델이며, 다양한 다른 학습데이터만 있으면 해당 학습데이터를 사용하여 어휘획득의 과정을 알아볼수 있다.

여러 가지 반복 실험을 통해 가장 적절히 심성어휘집과 같은 전자 사전 구축과, 품사정보, 구문정보, 의미정보등 그 어휘가 가지는 언어적 지식을 추가하는 연구를 수행중이다.

향후 다른 언어의 적용으로 모델의 검증과, 구축된 사전을 통해 인지신경기반 어절인식 모델을 개발하고, 어휘획득 모델과 어절인식 모델을 이용하여 인지신경기반 어휘 정보 처리 모델을 개발하는 것이 목표이다.

5. 참고문헌

[1] Butterworth, B., "Lexical representation of derivational relation", In M. Aronoff & M..L. Kean(Eds.), Juncture, 37-55. Saratoga, CA: Anma Libri., 1983.

[2] Anshen, F., Aronoff, M. “Producing morphologically complex words”, *Linguistics*, 26, 1988.

[3] Cole, P., Segui, J., Taft, M., “Words and Morphemes as Units for Lexical Access”, *Journal of Memory and Language*, 28, 1997.

[4] Jung, J., Lim, H., Nam, K., Morphological Representations of Korean compound Nouns. Journal of Speech and Hearing Disorders, Vol. 12, pp. 77-95, 2003.

[5] Lim, H., Nam, K., Hwang, Y., A Computational Model of Korean Mental Lexicon, Lecture Notes in Computer Science, Vol. 3480, pp. 1129-1136, 2005.

[6] Nam, K., Seo, K., Choi, K., The word length effect on Hangul word recognition. Korean Journal of Experimental and Cognitive Psychology, 9, 1-18, 1997.