유아 언어학습에 대한 하이퍼망 메모리 기반 모델

(Hypernetwork Memory-Based Model for Infant's Language Learning)

이 지 훈 † 이 은 석 ††
(Ji-Hoon Lee) (Eun Seok Lee)

장 병 탁 †††
(Byoung-Tak Zhang)

요 약 유아들의 언어습득에 있어서 중요한 점 하나는 학습자에 대한 언어환경의 노출이다. 유아가 접하는 언어환경은 부모와 같은 인간뿐만 아니라 각종 미디어와 같은 인공적 환경도 포함되며, 유아는 이러한 방대한 언어환경을 탐색하면서 언어를 학습한다. 본 연구는 대용량의 언어 데이터 노출이 영향을 미치는 유아언어학습을 유연하고 적절하게 모사하는 인지적 기제에 따른 기계학습 방식을 제안한다. 유아의 초기 언어학습은 문장수준의 학습과 생성 같은 행동들이 수반되는데, 이는 언어 코퍼스에 대한 노출만으로 모사가 가능하다. 모사의 핵심은 언어 하이퍼망 구조를 가진 기억기반 학습모델이다. 언어 하이퍼망은 언어구성요소들 간의 상위차원 관계 표상을 가능케 함으로써 새로운 데이터 스트림에 대해 유사구조의 적용과 이용을 도모하여 발달적이고 점진적인 학습을 모사한다. 본 연구에서는 11 개의 유아용 비디오로부터 추출한 문장 32744개를 언어 하이퍼망을 통한 점진적 학습을 수행하여 문장을 생성해 유아의 점진적, 발달적 학습을 모사하였다.

키워드 : 언어학습, 언어생성, 문장생성, 하이퍼망 학습, 하이퍼네트워크

Abstract One of the critical themes in the language acquisition is its exposure to linguistic environments. Linguistic environments, which interact with infants, include not only human beings such as its parents but also artificially crafted linguistic media as their functioning elements. An infant learns a language by exploring these extensive language environments around it. Based on such large linguistic data exposure, we propose a machine learning based method on the cognitive mechanism that simulate flexibly and appropriately infant's language learning. The infant's initial stage of language learning comes with sentence learning and creation, which can be simulated by exposing it to a language corpus. The core of the simulation is a memory-based learning model which has language hypernetwork structure. The language hypernetwork simulates developmental and progressive language learning using the structure of new data stream through making it representing of high level connection between language components possible. In this paper, we simulates an infant's gradual and developmental learning progress by training language hypernetwork gradually using 32,744 sentences extracted from video scripts of commercial animation movies for children.

Key words : Language learning, Language generation, Sentence generation, Hypernetwork learning

1. 서 론

1.1 언어습득 이론

현대언어학을 지배하고 있는 생성문법은 언어습득의 핵심은 구문syntax의 학습에 있다고 주장한다. 구문론은 잘 형성된 문장형태와 잘못 형성된 형태를 구분하는 기준이 되며 문장 S를 구문적으로 옳거나/그른 코드로 맵핑하는 것을 배우는 과정이 언어습득과정이라고 본다. 또 이러한 과정을 가능하게 하는 기제는 태어날 때부터 가능하다고 본다[1].

하지만 소리/단어/문장 스트림으로부터 언어적 패턴을 사실상 추출해내는 확률기반 기계학습적인 접근을 통한 연구로 인해 생성문법이론에 대해 많은 비판이 가해졌다. 구문론적 규칙을 가지고 해 공간을 탐색하는 것보다는, 언어의 확률적이고 통계적인 측면에 대한 정보의 증가를 통해 언어학습이 가능하다는 입장이 확률기반 언어습득론이다[2]. 기계학습이 좀더 풍부한 데이터를 가용할

· 이 연구는 한국학술진흥재단(KRF-2008-314-D00377), 과학재단 미래유망 파이오니어 사업, 정보통신연구진흥원 IT산업원천기술개발사업(IITA-2009-A1100-0901-1639), BK21-IT에 의하여 지원되었음
· 이 논문은 2009 한국컴퓨터종합학술대회에서 '하이퍼망 메모리 기반 유아 언어학습 및 생성 모델'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 서울대학교 생물정보학협동과정
jhlee@bi.snu.ac.kr
†† 학생회원 : 서울대학교 인지과학협동과정
eslee@bi.snu.ac.kr
††† 종신회원 : 서울대학교 컴퓨터공학부 교수
btzhang@bi.snu.ac.kr

논문접수 : 2009년 8월 14일
심사완료 : 2009년 10월 8일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제15권 제12호(2009.12)

수 있게 되면서, 이 같은 논의는 더욱 활발해지고 있다. 즉 언어적 "자극Stimulus"은 그간 언어학자들이 생각했던 것보다도 훨씬 더 풍부하고 방대하다. 두 데이터 간의 맵핑-언어형태와 그 형태가 지시하는 의미-이 cross-situational 추론을 통해 이루어질 수도 있고[3], 언어데이터에 대해 방대한 양의 확률정보를 암묵적으로 저장하고 이를 확률적 제약조건으로 삼아 언어가 학습되기도 한다[2]. 이 이론적 입장은 언어학습은 인간의 보편적인 학습기제로 인해 일어난다는 가설을 지지하고 있다.

언어습득에 있어서 최근 나타나는 또다른 계산론적 방법론으로 발달 모델을 들 수 있다. 앞서의 두 입장은 언어의 특정 현상을 시간의 흐름을 무시한 채 그 자체만으로 고립시켜놓은 상태에서 다룬다. 이러한 접근은 언어습득의 과정을 중요하게 뒷받침하는 발달적 측면을 도외시할 가능성이 높다. 발달 모델은 이러한 점을 보충한다. 즉 incremental, open-ended 방식으로 학습이 가능하도록 한다[4-6]. 이 방법론에서 유아는 자신을 둘러싼 언어환경을 탐색하면서 관찰된 자극들과 상호영향을 주고받는다. 이렇게 함으로써 학습 단계가 늘어가면서 시간의 흐름에 따라 주어진 언어자극을 저장/변형/생성하는 경험을 늘려간다. 즉 한꺼번에(앞의 두 접근처럼) 언어의 복잡성에 노출되는 것이 아니라, 단계적으로 복잡성을 증가시켜나가는 것이다.

1.2 연구 목적

본 연구에서는 이와 같은 언어의 발달적 측면을 고려한 확률기반의 언어학습 양상을 보인다. 구체적으로 시뮬레이션하는 내용은 언어환경으로부터의 자극이 증가함에 따라 유아(기계학습자)가 점점 그 자극에 대해 연합된 개념이 풍부해지고 추상적 구조(구문론적 구조)의 출현이 점점 강화되는 양상이다. 이 기제의 핵심은 언어 하이퍼망 구조를 가진 기억기반 학습모델이다. 언어 하이퍼망은 언어구성요소들 간의 상위차원 관계 표상을 가능케 함으로써 새로운 데이터 스트림에 대해 유사구조의 적용과 이용을 도모하여 발달적이고 증가적인 학습을 시뮬레이션한다[7-9].

언어학습에 대해 가장 활발한 논의가 진행되는 영역은 구문론이다. 유아가 어떤 방식으로 처음에는 무분별하게 나열된 데이터의 스트림으로부터 소리, 단어, 구, 절을 구분해내는가? 본 연구에 적용된 언어 하이퍼망은 어떠한 명시적인 구문론적 규칙이 없이도 계속되는 학습 단계별로 주어진 자극으로부터 연합기억을 만들어내고, 이를 바탕으로 구문론적으로 정합적인 문장을 생성하고, 또 더 나아가서 의미론적으로도 정합적인 문장을 생성하는 데 적합한 모델이다.

1.3 유아에게 노출되는 언어환경의 특징

유아가 언어환경에 노출되면서 드러나는 언어환경의 특징을 약술하면 다음과 같다. 첫째 언어자극들이 random하게 나열되어 있고, 자극들 간에 연합되어 제시되는데 그 방식 또한 유아 입장에서는 알 수가 없다. 따라서 언어형태-의미간 맵핑 함수는 유아에게 전혀 가용하지 않다. 물론 유아의 부모와 같은 언어학습을 매개하는 기능이 있으나 이것이 유아의 언어발달에 전면적인 영향을 끼치지는 않는다[10].

둘째 언어학습 환경은 생각보다 훨씬 방대하며 복잡하다. 따라서 환경 전체를 정확하고 면밀하게 모사한 환경자극을 제시해야만 시뮬레이션의 학습자도 현실적이고 유연하게 언어학습의 양상을 보일 수 있을 것이다.

셋째 언어환경은 시간의 흐름에 따라 연속적으로 유아에게 입력된다. 이 환경은 똑 같은 양상으로 자주 나타날 수도 있고, 한번 나타났던 것이 다시는 나타나지 않을 수도 있다. 즉 그 환경의 시간에 따른 변화가 심하다. 이 환경을 얼마나 현실적으로 입력할 수 있는가 역시 발달적 학습의 요소를 도입하는 데 핵심적인 문제일 것이다.

2. 언어 하이퍼망 모델

하이퍼망 H는 두 개 이상의 여러 개의 정점의 집합 X, 두 개 이상의 정점의 집합인 하이퍼에지 E, 하이퍼에지의 가중치(weight) 집합 W로 구성된다[8]. 본 논문의 하이퍼망 언어 모델에서는 문장의 단어가 정점이 되고, 순서정보가 있는 연결된 단어 묶음이 하이퍼에지가 되며, 각 하이퍼에지의 출현 빈도가 가중치가 된다. 이 때 연결되는 단어의 개수를 Order라 한다. 예를 들어, 그림 1에서 언어 하이퍼망은 7 단어 good, friend, best, my, a, have, your 를 정점으로 가지고, Order 3의 5개 문장, my good friend, my best friend, have a friend, your best friend, a good friend 를 하이퍼에지로 가진다. 각각의 하이퍼에지들은 순서대로 weights(3,2,1,4,1)를 가진다. 각각 하이퍼에지들을 연결하면 그림 1에서와 같이 하이퍼망을 만든다.

기존의 하이퍼망은 하이퍼에지 내 정점들 간의 방향성을 고려하지 않지만 본 논문의 언어 하이퍼망에서는 정점의 순서를 고려한다. 이것은 문장 생성 문제에 있어서 연결된 단어간의 선후 관계에 의미를 부여함으로써 언어 특성에 맞는 하이퍼망을 만들기 위해서이다.

언어 하이퍼망은 유아의 언어 자극 정도에 따라 다르게 생성되며, 생성된 하이퍼망을 통해 문장을 생성함으로써 유아의 언어 학습 양상을 모사할 수 있다.

3. 문장 생성 실험

3.1 실험 재료 및 단계

언어환경은 총 11단계로 구성하여 단계별로 총합적으로 incrementally 학습시키고, 각 단계별로 주어진 키워

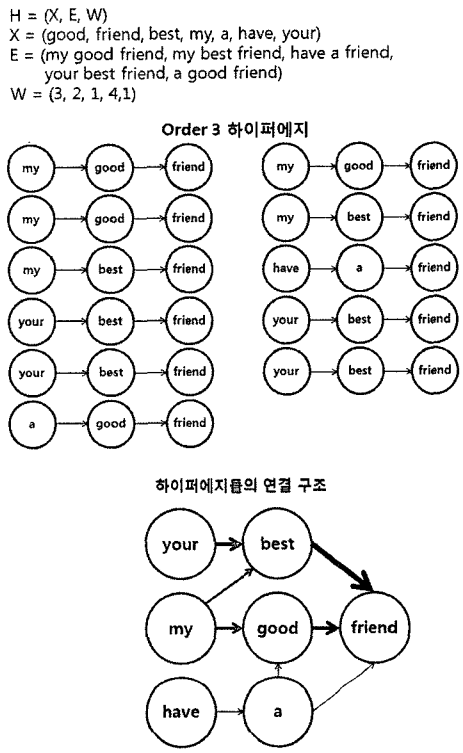

그림 1 하이퍼망 H는 단어 집합 X, 하이퍼에지 집합 E, Weight 집합 W로 구성된다. Order 3의 하이퍼에지들이 모여 하이퍼망 구조를 만든다.

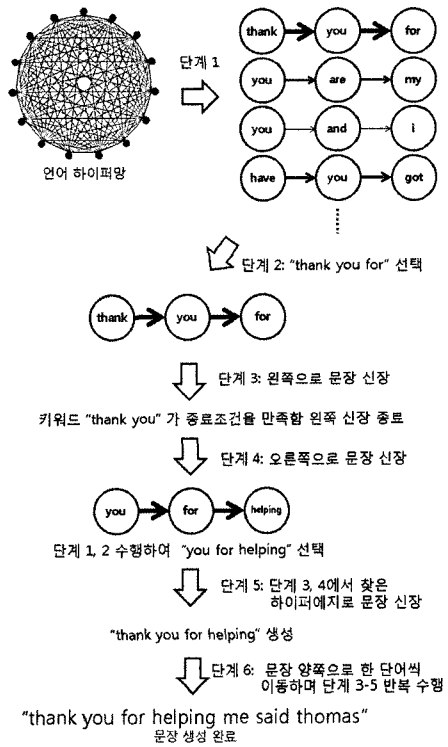

그림 2 언어 하이퍼망을 통한 문장 생성 예시. 키워드는 'you', 생성된 문장은 "thank you for helping me said thomas"

드에 따라 생성되는 문장 100개를 추출한 후, 이들 중 구문론적으로 옳은 문장과 의미론적으로 옳은 문장들의 분포와 비율을 분석하였다.

각 단계별로 학습시키는 문장은 유아용 비디오의 자막으로 구성되어 있다. 실험에 쓰인 문장데이터의 비디오는 <늑대와 7마리 아기양> <미피와 친구들> <루니튠스> <까이유> <도라도라> <싱어롱 맥도널드 농장> <꼬마기관차 토마스> <티모시네 유치원> <곰돌이 푸> <굿모닝 헬로 키티> <찰리브라운과 스누피>이다. (학습순서에 따라 나열함)

언어 하이퍼망을 구성하는 문장데이터의 용량은 각 단계별 평균 100kb이다. 본 실험에서 사용된 키워드는 명사/형용사 형태로, Kucera and Francis 빈도에서 상위에 위치하는 단어 'you'를 사용하였다[11].

키워드에 따라 생성된 문장들 중 100개를 무작위로 추출하고, 이를 다시 구문론적으로 옳은 문장과 의미론적으로 옳은 문장, 그리고 옳지 않은 문장으로 분류하여, 학습단계 발전에 따른 생성 양상을 살펴보았다.

3.2 하이퍼망 학습 및 생성

유아의 언어 학습 양상을 모사하기 위한 방법으로 언어 하이퍼망을 통한 문장 생성 실험을 수행하였다. 하이퍼망을 Order 3으로 생성하였을 경우 문장은 그림 2에서 처럼 키워드를 중심으로 양방향으로 생성된다.

문장생성 알고리즘은 다음과 같다:

단계 1. 주어진 키워드 $L_q = (x_q)$를 하이퍼망 H에서 검색하여 동일한 키워드를 포함하는 하이퍼에지를 찾아 $M = \{L_1, L_2, \ldots, L_m\}$에 저장.

단계 2. M 에서 Roulette wheel selection을 통해 하이퍼에지 $L_h = (x_{q-1}, x_q, x_{q+1})$ 선택

단계 3. 키워드 $L_q = (x_{q-1})$ 을 업데이트 한 후 단계 1, 2 를 다시 수행 하여 $L_{left} = (x_{q-2}, x_{q-1}, x_q)$ 를 정한다.

단계 4. 키워드 $L_q = (x_{q+1})$ 을 업데이트 한 후 단계 1, 2 를 다시 수행하여 $L_{right} = (x_q, x_{q+1}, x_{q+2})$ 를 정한다.

단계 5. L_{left} 를 L_{right} 연결하여 부분 문장 $L_h = (x_{q-2}, x_{q-1}, x_q, x_{q+1}, x_{q-2})$ 을 생성한다.

단계 6. 단계 3-5 를 $L_q = (x_{q-2})$ 과 $L_q = (x_{q+2})$ 에 수행한다. 이 과정을 문장이 종료조건을 맞이할 때까지 반복 수행한다.

문장 생성의 종료는 생성되고 있는 문장의 양쪽 말단에 위치한 단어들이 하이퍼망 상에서 종료 단어이거나 시작 단어일 확률이 높을 경우 종료하도록 하였다.

4. 결과 및 분석

표 1은 하이퍼망의 언어 데이터 학습 후 'you'라는 키워드를 중심으로 생성된 문장들의 예문이다. 제시된 문장들은 생성된 문장 100개 중에서 구문론적/의미론적으로 공히 정합적인 문장들이다. 이 문장들은 단어 4개부터 10개로 구성되어 있고, 학습 단계가 진행될수록 정합적이면서 길이가 긴 문장들이 더 많이 생성되었다.

학습 데이터를 구성하는 형태별 어휘들의 개수(type number)는 총 6124개이고 사용된 모든 어휘 개수(token number)는 총 252936개로서 small world를 형성한다. 하이퍼망이 형성하는 전체 가설공간, 즉 총 문장 개수의 크기는 32744개이다. 학습단계는 총 11회에 걸쳐 진행했다. 주어진 키워드에 대해 생성된 문장의 정합성에 대한 분석 기준은 다음과 같다:

* **문장의 구문론적 정합성**: 생성된 문장이 문법적으로 옳은지의 여부를 살펴보았다.
* **문장의 의미론적 정합성**: 의미적으로 옳은 문장인지 여부. 문장이 문법적으로 옳다고 하더라도 의미적으로 옳지 않은 경우가 많다. 가장 생성되기 어려운 기준이므로 이 기준에 부합하는 문장 수도 가장 적다.

그림 3에서 보는 것과 같이, 실험 결과 학습 단계가 진행될수록 생성된 문장에서 구문론적, 의미론적 정합적인 문장의 개수가 늘어나는 것을 볼 수 있었다.

표 1 문장 생성 결과

Round 1	Are you all right, You will know him at once
Round 2	Don't you read a book, You can do Where are you carrying Thank you, I think you have made me happy Don't you take your crayons, You must be hungry
Round 3	So are you trying to paint, You must be very hungry Just wait till you got too much talent Thought I told you I'd clean up I will come with you anyway You must sit and be patient Shoot me now or wait till you got your pencil
Round 4	Let me help you, Oh no you can't go out together Well why don't you read a book, What are you doing Why don't you come back here, Why don't you ask
Round 5	Do you see a bridge, Can you find the right one Let me help you keep the track, You find the right one It's all thanks to you, ill you help us get some stamps
Round 6	Do you see the witch, You can do card tricks You can be the best at starting off The fruit will be good for you, Hearts will stay with you Here's another little song for you You don't have to say hello again What are you making for your radio company Do you want mommy to help me do my thing
Round 7	So what do you mean I can find me You are my best friend, You can do it All you have to do is dip the sponge Percy told Gordon all about you, Here's what you do
Round 8	Do you see a lion Of course you know what that means Can you find the key, I just thought I'd help you You can be safe in this great country of ours When you are here to see the dinosaurs
Round 9	We'll get you there, To all you have to say goodbye How would you like that, Here's what you do So what do you do, Do you see another egg And what do you mean I can make you feel better
Round 10	What are you making, Do you think so too I demand you tell him who you are, Did you get it Do you see a farmer standing knee deep in snow But they were surprised to see you Do you want to go, I wish I could give you
Round 11	Did you get the basket of treats to grandma's house Do you have the flu, Can you help me shovel What do you think of that grizzly bear What are you doing down there Thank you for helping me, Caillou was happy to see you You know what I say What do you think of that grizzly bear Do you think that's what we can do better tomorrow? Hey what are you doing under there

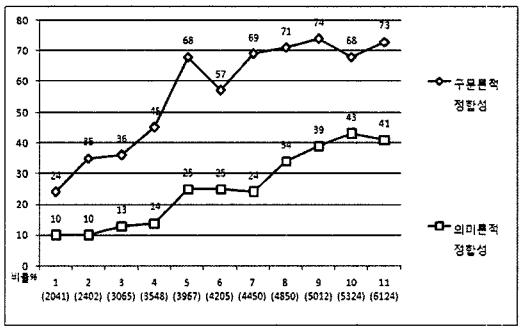

그림 3 각 학습 단계별 정합적 문장생성 비율(%). 총 11단계까지 진행했다. 괄호 안은 학습 어휘 type의 개수이다.

5. 결론 및 논의

본 연구는 언어학습에 대한 선험적인 규칙 없이 하이퍼망 기반의 기억모델을 통해 언어환경의 확률적 분포와 시간에 따른 언어환경 및 기억 변화로부터 유아의 언어획득 및 생성이 가능한지에 대해 시뮬레이션 실험으로 고찰해보았다. 언어환경 구축은 유아용 비디오의 자막 데이터를 사용하였고, 하이퍼망을 사용하여 총 11단계의 점진적, 발달적 학습을 모사하였다. 이후 하이퍼망이 주어진 키워드에 따라 생성하는 문장들의 구문론적/의미론적 정합성을 확인하였다.

하이퍼망은 학습 데이터에 대한 명시적인 규칙 없이 데이터 항목들 사이의 연결 양상으로부터 구조를 파악하고 주어진 구조와 비슷한 데이터를 생성하는 데 유용하다. 또 데이터가 증가적으로 늘어날수록 학습 효율과 유연성이 뛰어나 언어학습처럼 가설공간이 사실상 무한에 가깝고 보편적인 인지기제들이 동원되는 언어적 작업수행에 좋은 능력을 보인다.

실험을 통해, 데이터가 증가하면서 정합적인 문장의 생성 능력이 늘어감을 알 수 있었다. 또 정합적인 문장

의 길이 또한 증가하였기 때문에, 유아언어학습의 보편적인 양상을 반영하는 부분이라 할 수 있다. 종합적으로, 본 실험이 언어학습의 발달적 측면을 일부나마 잘 모사한 것으로 볼 수 있다. 차후 좀더 많은 양의 데이터와 학습단계, 언어학습에 영향을 끼치는 다양한 심리학적 변인들로 연계 실험을 진행할 필요가 있다.

참 고 문 헌

[1] Marcus, G., *Poverty of the stimulus arguments*, Cambridge MA, MIT Press, pp.660-661, 1999.

[2] Seidenberg, M. and MacDonald, M., A probabilistic constraints approach to language acquisition and processing, *Cognitive Science*, vol.23, pp.569-588, 1999.

[3] Siskind, J., A computational study of cross-situational techniques for learning word-to-meaning mappings, *Ccognition*, vol.61(1-2), pp.39-91, 1996.

[4] Weng, J., A theory for mentally developing robots, In *Second International Conference on Development and Learning*, IEEE Computer Society Press, 2002.

[5] Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G., Developmental robotics: A Survey, *Connection Science*, vol.15(4) pp.151-190, 2003.

[6] Oudeyer, P.-Y., Kaplan, F., and Hafner, V., Intrinsic motivation systems for autonomous mental development, *IEEE Transactions on Evolutionary Computation*, vol.11(2) pp.265-286, 2007.

[7] Zhang, B.-T. and Kim, J.-K., DNA hypernetworks for information storage and retrieval, Lecture Notes in Computer Science, *DNA12*, vol.4287, pp. 298-307, 2006.

[8] Kim, S., Heo, M.-O., and Zhang, B.-T., Text classifiers evolved on a simulated DNA computer, *IEEE Congress on Evolutionary Computation*, pp. 9196-9202, 2006.

[9] Ha, J.-W., Eom, J.-H., Kim S.-C. and Zhang, B.-T., Evolutionary hypernetwork models for aptamer-based cardiovascular disease diagnosis, *The Genetic and Evolutionary Computation Conference*, vol.4, pp.2709-2716, 2007.

[10] Chomsky, N., *Rules and Representations*, Oxford: Basil Blackwell, 1980.

[11] Kucera, H. & Francis, W., *Computational analysis of present-day American English*, Providence, RI: Brown University Press, 1967.