단어 간 관계 패턴 학습을 통한 하이퍼네트워크 기반 자연 언어 문장 생성

(Hypernetwork-based Natural Language Sentence Generation by Word Relation Pattern Learning)

석 호 식 † 작 가 멧 † 장 병 탁 ††
(Ho-Sik Seok) (Jakramate Bootkrajang) (Byoung-Tak Zhang)

요 약 본 논문에서는 단어간 관계 패턴을 학습한 후 이에 기반하여 자연 언어 문장을 생성하는 방법을 소개한다. 기존의 문장 생성 방법론에서는 내재된 문법 규칙의 존재를 가정하거나 템플릿을 사용하고 있으나, 본 논문에서 소개하는 방법론에서는 태깅 등의 부가 정보 없이 단어의 동시 등장 빈도만을 활용하여 단어간 관계 패턴을 학습한다. 단어간 관계 패턴은 하이퍼네트워크 방법론에 기반하여 학습되었다. 학습이 진행됨에 따라 하이퍼네트워크의 복잡도가 높아지며, 학습 모델에 축적되는 언어 관계 패턴의 수가 증가한다. 학습된 모델의 유효성은 학습 패턴에 기반한 자연 언어 문장 생성을 통해 확인하였다. 실험 결과 학습이 진행됨에 따라 문법적으로 성립하는 문장의 비율이 향상하였다. 파서를 이용하여 생성된 문장을 구성하는 문법 규칙을 분석한 후 문법 규칙의 분포를 학습에 사용한 코퍼스의 문법 규칙 분포와 비교한 결과 학습에 사용된 코퍼스의 문법적 특성을 학습할 수 있는 잠재력을 갖고 있음을 확인하였다.

키워드 : 하이퍼네트워크, 자연언어문장생성, 기계 학습

Abstract We introduce a natural language sentence generation (NLG) method based on learning of word-association patterns. Existing NLG methods assume the inherent grammar rules or use template based method. Contrary to the existing NLG methods, the presented method learns the words-association patterns using only the co-occurrence of words without additional information such as tagging. We employ the hypernetwork method to analyze and represent the words-association patterns. As training going on, the model complexity is increased. After completing each training phase, natural language sentences are generated using the learned hyperedges. The number of grammatically plausible sentences increases after each training phase. We confirm that the proposed method has a potential for learning grammatical properties of training corpuses by comparing the diversity of grammatical rules of training corpuses and the generated sentences.

Key words : Hypernetwork, natural language generation, machine learning

· 이 연구는 한국학술진흥재단(KRF-2008-314-D00377), 지식경제부 및 한국산업기술평가관리원의 IT산업원천기술개발사업(2009-F-051-01, 차세대 맞춤형 서비스를 위한 기계학습 기반 멀티모달 복합 정보 추출 및 추천기술 개발), BK21-IT 사업에 의하여 지원되었음

† 학생회원 : 서울대학교 컴퓨터공학부
hsseok@bi.snu.ac.kr
jakramate@bi.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
btzhang@bi.snu.ac.kr
(Corresponding author임)

논문접수 : 2009년 11월 23일
심사완료 : 2010년 1월 5일

정보과학회논문지: 소프트웨어 및 응용 제37권 제3호(2010.3)

1. 서 론

컴퓨터에 의한 자연언어 처리 방식에 아직 제약이 많기 때문에 기존의 컴퓨터 기반 자연 언어 생성(Natural Language Generation, NLG) 방식에서는 이미 저장되어 있는 문장을 변화 없이 출력하거나 저장된 템플릿을 약간 변형하여 문장을 출력하는 방식, 혹은 이미 저장되어 있는 문법 규칙을 이용하여 파싱 트리를 만든 후 이를 활용하여 문장을 생성하는 방식을 취해 왔다[1,2]. 그러나 이 같은 방식에서는 사전에 설정한 문장 생성 능력을 능가할 수 없으며, 인터넷 상에 끊임없이 축적되고 있는 데이터를 사용할 수 없다는 단점이 있다. 또한 기계 학습 기법에 기반한 지능적 처리 능력 부여도 어렵다는 문제가 있다. 본 논문에서는 부가 정보가 없는 텍

스트 데이터에 기반한 단어간 관계 패턴 학습을 통해 문장 생성 능력을 향상시키는 방법을 소개한다. 제안 문장 생성 방법론의 특징은 다음과 같다. 첫째, 사전에 주어진 문법 규칙의 존재를 가정하지 않으며, 둘째, 코퍼스(corpus)의 증가에 따라 문장 생성 능력이 향상된다. 그리고 셋째, 연상 기억 학습의 특징에 기반하고 있다.

본 연구에서는 하이퍼네트워크[3] 방법론을 이용하여 단어간 상관관계를 확보한다. 하이퍼네트워크는 병렬 연상 메모리 모델로 제안된 것으로, 많은 수의 랜덤 하이퍼에지로 구성되어 있다. 각 하이퍼에지는 노드들의 고차원 상호작용에 가중치가 부여된 것으로 하이퍼에지의 조합을 통해 정보를 처리하게 된다. 하이퍼네트워크 모델은 차원의 제한 없이 노드들의 고차원 상호작용을 표현할 수 있도록 허용하기 때문에 여러 개의 단어가 나타내는 단어 간 관계를 표현할 수 있는 잠재력을 갖고 있다.

학습 코퍼스로는 비디오의 스크립트를 선정하였다. 학습에 사용한 비디오 스크립트는 총 9개(유아용 비디오 8종, 시트콤 1종)이다. 사용 코퍼스의 문법적 규칙의 복잡도를 엔트로피 측면에서 분석한 결과 CHILDES 데이터베이스[4]의 복잡도와 유사한 코퍼스라는 것을 확인하였다. 하이퍼네트워크 방법론을 이용한 언어 모델 성능 확인을 위해, 문장의 일부만 큐로 주어진 문장을 입력으로 받아 큐를 이용하여 문장을 생성하는 실험을 실시하였다. 실험 결과 코퍼스 공급에 따라 문법적으로 성립하는 문장의 비율이 향상하는 것을 확인할 수 있었으며, KL 다이버전스(Kullback-Leibler divergence)를 이용하여 훈련용 코퍼스의 문법 규칙 분포와 생성된 문장의 문법 규칙 분포 차이가 크지 않음을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 제안 방법을 보다 자세히 설명한다. 3장에서는 실험 결과를 제시하고 실험 결과를 분석한다. 4장에서는 본 연구와 관련된 다른 기존 연구들을 살펴보며, 5장에서는 본 논문의 의의를 설명하고 향후 연구 방향을 제시한다.

2. 제안 방법

2.1 하이퍼네트워크 방법론

하이퍼네트워크는 가중치가 있는 하이퍼그래프로 노드간의 상호작용은 하이퍼에지로 표현된다. 하이퍼네트워크 $H=(X,E,W)$는 노드 집합 $X=\{x_1, x_2, ..., x_{|X|}\}$, 에지 집합 $E=\{E_1, E_2, ..., E_{|E|}\}$, 가중치 집합 $W=\{w_1, w_2, ..., w_{|E|}\}$로 정의된다. 하이퍼에지 E의 기수는 $k(k\geq 1)$로 하이퍼에지는 2개 이상의 노드를 연결할 수 있다.

하이퍼네트워크 모델은 데이터 집합 $D=\{x^{(n)}\}_{n=1}^{N}$을 저장하는 확률 연상 메모리로 사용될 수 있기 때문에 주소가 아닌 컨텐츠를 이용하여 메모리 내용을 추출한다. 하이퍼네트워크의 에너지는 다음과 같이 정의된다.

$$e(x^{(n)};W)=-\sum_{i=1}^{|E|} w_{i_1,i_2,...,i_{|E_i|}} x_{i_1}^{(n)} x_{i_2}^{(n)} ... x_{i_{|E_i|}}^{(n)} \quad (1)$$

여기서 W는 하이퍼네트워크 모델을 위한 인자(하이퍼에지 가중치)를 의미한다. $x_{i_1}^{(n)} x_{i_2}^{(n)} ... x_{i_{|E_i|}}^{(n)}$는 데이터 아이템의 조합을 의미하는 것으로 결국 식 (1)은 주어진 데이터 아이템을 조합한 후 이에 가중치를 부여한 항의 합으로 하이퍼네트워크를 표현한다는 것을 나타내고 있다.

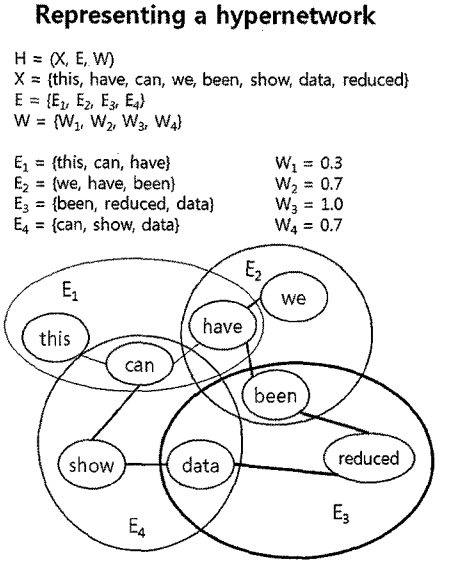

그림 1 하이퍼네트워크의 표현 예. 하이퍼네트워크 $H=(X, E, W)$는 주어진 데이터 노드의 조합에 가중치를 부여한 항의 합으로 데이터의 분포를 표현한다.

하이퍼네트워크에서 데이터가 생성될 확률은 깁스 분포(Gibbs distribution)[5]에서 주어진다.

$$P(x^{(n)}|W)=\frac{1}{Z(W)}\exp\{-e(x^{(n)};W)\}$$

여기서 $exp\{-e(x^{(n)};W)\}$를 볼츠만 팩터라고 하며 $Z(W)$는 다음과 같이 표현된다.

$$Z(W)=\sum_{x^{(n)}}\exp\{-e(x^{(n)};W)\}$$
$$=\sum_{x^{(n)}}\{\sum_{i=1}^{|E|} w_{i_1 i_2 ... i_{E_i}} x_{i_1}^{(m)} x_{i_2}^{(m)} ... x_{i_{ij}}^{(m)}\}$$

즉 하이퍼네트워크는 하이퍼에지와 가중치를 이용하여 데이터 집합의 확률적 모델을 표현하게 된다. 데이터 집합 $D=\{x^{(n)}\}_{n=1}^{N}$가 주어졌을 때 학습의 목적은 아래

의 우도함수(likelihood function)를 최대화하는 하이퍼네크워크를 발견하는 것이다.

$$P(D|W)=\prod_{n=1}^{N}P(x^{(n)}|W) \tag{2}$$

식 (2)에 대하여 로그를 취하면

$$\ln P(D|W)=\ln\prod_{n=1}^{N}P(x^{(n)}|W)$$

$$=\sum_{n=1}^{N}\{[\sum_{k=1}^{K}\frac{1}{C(K)}\sum_{i_1i_2\ldots i_k}w_{i_1i_2\ldots i_k}^{(k)}x_{i_1}^{(n)}x_{i_2}^{(n)}\ldots x_{i_k}^{(n)}]-\ln Z(W)\} \tag{3}$$

식 (3)을 가중치 W에 대하여 미분하면

$$\frac{\nabla}{\nabla W_{i_1i_2\ldots i_k}^{(K)}}\ln\prod_{n=1}^{N}P(x^{(n)}|W)$$

$$=\frac{\nabla}{\nabla W_{i_1i_2\ldots i_k}^{(K)}}\sum_{n=1}^{N}\{[\sum_{k=1}^{K}\frac{1}{C(K)}\sum_{i_1i_2\ldots i_k}w_{i_1i_2\ldots i_k}^{(k)}x_{i_1}^{(n)}x_{i_2}^{(n)}\ldots x_{i_k}^{(n)}]-\ln Z(W)\}$$

$$=\sum_{n=1}^{N}\{\frac{\nabla}{\nabla W_{i_1i_2\ldots i_k}^{(K)}}[\sum_{k=1}^{K}\frac{1}{C(K)}\sum_{i_1i_2\ldots i_k}w_{i_1i_2\ldots i_k}^{(k)}x_{i_1}^{(n)}x_{i_2}^{(n)}\ldots x_{i_k}^{(n)}]$$

$$-\frac{\nabla}{\nabla W_{i_1i_2\ldots i_k}^{(K)}}\ln Z(W)\}$$

$$=\sum_{n=1}^{N}\{x_{i_1}^{(n)}x_{i_2}^{(n)}\ldots x_{i_k}^{(n)}-<x_{i_1}x_{i_2}\ldots x_{i_k}>_{P(x|W)}\}$$

$$=N\{<x_{i_1}x_{i_2}\ldots x_{i_k}>_{Data}-<x_{i_1}x_{i_2}\ldots x_{i_k}>_{P(x|W)}\}$$

여기서 $<x_{i_1}x_{i_2}\ldots x_{i_k}>_{Data}=\frac{1}{N}\sum_{n=1}^{N}[x_{i_1}^{(n)}x_{i_2}^{(n)}\cdots x_{i_k}^{(n)}]$

$<x_{i_1}x_{i_2}\ldots x_{i_k}>_{P(x|W)}=\sum_{x}[x_{i_1}x_{i_2}\cdots x_{i_k}P(x|W)]$ 데이터 집합의 하이퍼에지 평균 빈도수와 하이퍼네트워크 모델 하이퍼에지 평균 빈도수의 차이를 최소화하여 우도(likelihood)를 최대화하면서 학습이 진행된다[3].

2.2 문장 학습 및 문장 생성

2.2.1 문장 학습

문장 학습을 위한 하이퍼네트워크 언어처리 알고리즘의 슈도 코드는 다음과 같다.

표 1 하이퍼네트워크 언어처리 알고리즘 슈도 코드

```
초기화
코퍼스에서 문장 획득
획득된 문장에서 무작위로 복수 개의 단어를 추출한 후 단어간의 관계성을
분석하여 하이퍼에지 E = {E1, E2, ..., E|E|} 생성

학습
시작: H = {X, E, W} = {0,0,0}
1. 하이퍼에지 생성
순차적 랜덤 샘플링을 이용하여 훈련 문장에서 하이퍼에지 ei 생성
 - 단계 1 에서
   E ← E ∪ {ei}
   X ← X ∪ {xi|xi ∈ ei}
   W ← W ∪ {wi|wi ∈ winit}
 - 단계 k + 1에서(k ≥ 1)
 k + 1 번째 코퍼스에서 생성한 하이퍼에지 ei에 대하여
   W의 모든 원소 wi에 대하여 wi = wi/3
   E ← E ∪ {ei}
   X ← X ∪ {xi|xi ∈ ei}
 k번째 코퍼스에서 생성한 하이퍼에지 ei 및 새로운
 하이퍼에지 ej'에 대하여
                    wj ← wj ∈ winit
 (1) ej ≠ ei 인 ei가 존재하지 않는 경우
     W ← W ∪ {wj}
 (2) ej = ei인 ei가 존재하는 경우
     W ← W ∪ {w|w ← max{wi, wj}}
2. 사전 정의된 종료 조건까지 반복
```

본 논문에서 제안하는 언어처리 방식에서는 샘플링을 통해 하이퍼에지의 가중치를 변경시킨다. 제안된 방식은 샘플링 방식에 따라 널리 사용되는 자연 언어 분석 방법인 n-그램 방식[6] 보다 큰 표현력을 갖는다. 예를 들어 학습 과정에서 샘플링되는 n개의 단어를 단어 수 n의 윈도우에서 추출할 경우 n-그램 모델과 동일한 효과를 기대할 수 있으며, 샘플링되는 윈도우 크기를 확장하면 n-그램 방법론의 표현력을 능가하게 된다.

본 논문에서는 순차적 랜덤 샘플링(sequential random sampling) 방식을 이용하여 n-그램의 효과와 하이퍼그래프의 특징을 모두 유지하고자 하였다. 순차적 랜덤 샘플링 방식에서는 문장을 구성하는 단어의 순서를 유지하면서 주어진 기수에 맞게 무작위로 단어를 선정한다.

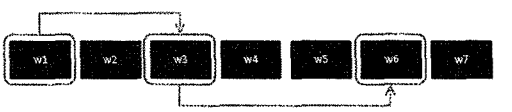

그림 2 기수 3의 하이퍼에지 생성. 7개의 단어로 구성된 문장에서 3개의 단어를 순차적으로 샘플링하여 기수 3의 하이퍼에지를 구성한다.

이를 통해 하이퍼네트워크가 문장 내에서 분리된 단어간의 관계를 지킬 수 있는 가능성이 높아진다. 만약 주어진 오더에 맞게 연속된 단어를 샘플링하게 된다면 n-그램 모델과 동일한 방식으로 동작하게 된다.

2.2.2 문장 생성

문장 생성을 위해서는 문장을 구성하는 단어의 일부만 존재하는 문장이 주어진다. 문장에 남아 있는 단어와

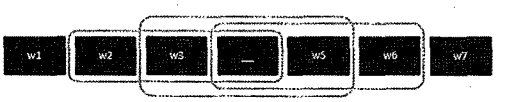

그림 3 문장 생성. 상기 그림과 같이 공백("_")이 있는 문장이 주어지면 해당 공백을 제외한 나머지 단어를 포함하고 있는 하이퍼에지를 하이퍼네트워크 모델에서 검색한 후 가중치가 가장 높은 하이퍼에지의 노드를 공백의 단어로 결정한다.

공백을 결합하여 하이퍼에지를 생성한 후, 하이퍼네트워크 모델에 존재하는 하이퍼에지들과 비교하여 매칭된 하이퍼에지를 선택한다. 선택된 하이퍼에지의 가중치를 관찰한 후 가장 큰 가중치를 갖는 하이퍼에지의 단어를 공백에 해당하는 단어라고 결정한다.

3. 실험 결과

3.1 실험용 코퍼스

실험에 사용한 코퍼스는 유아용 비디오의 스크립트 텍스트로 선정된 비디오는 Miffy, Looney tunes, Caillou, Dora Dora, Macdonald's farm, Thomas&Friends, Timothy, Pooh의 8종이며, 훈련에 사용되는 문장의 수가 늘어났을 경우의 효과를 확인하기 위하여 시트콤 Friends의 스크립트 데이터를 추가하였다. 실험에 사용된 코퍼스는 총 $O(100K)$의 문장으로 구성되어 있다. 훈련에 사용되는 코퍼스는 다음과 같은 순서로 공급된다.

D1 = Miffy, D2 = Looney, D3 = Caillou, D4 = Dora-Dora, D5 = Macdonald's farm. D6 = Thomas&Friends, D7 = Timothy, D8 = Pooh, D9 = Friends.

그림 5는 CHILDES 데이터베이스의 자료를 2세~성인까지 9개의 그룹으로 나누어 문법 규칙의 복잡도를 계산한 결과이다. 문법 규칙의 복잡도 측면에서 하이퍼네트워크 모델 생성에 사용한 코퍼스는 엔트로피 측면에서 아동이 사용하는 언어의 복잡도와 근사한 복잡도를 갖고 있다. 훈련과정에서 공급되는 코퍼스의 수는 단계별로 하나이지만, 이전 단계에서 공급되었던 코퍼스의 훈련 결과가 하이퍼네트워크의 하이퍼에지에 남아 있기 때문에, 단계별로 훈련 코퍼스를 축적하면서 훈련을 진행하는 것과 동일한 효과를 갖게 된다. 코퍼스의 복잡도는 스탠포드 파서[7]를 이용하여 문장을 구성하는 문법 규칙을 분리한 후, 분리된 문법 규칙의 엔트로피를 계산하여 측정되었다. 그림 4는 계산된 엔트로피 결과이다.

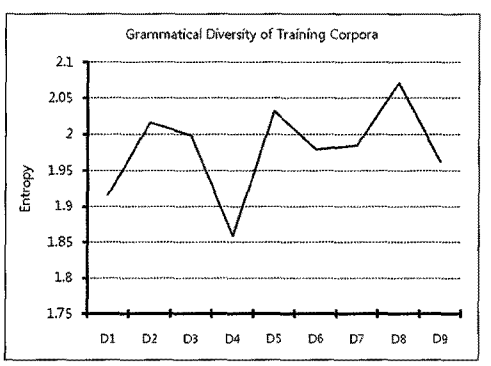

그림 4 훈련 코퍼스의 복잡도. 훈련용 코퍼스를 파싱한 후 파싱 트리의 문법 규칙 엔트로피로 복잡도를 계산

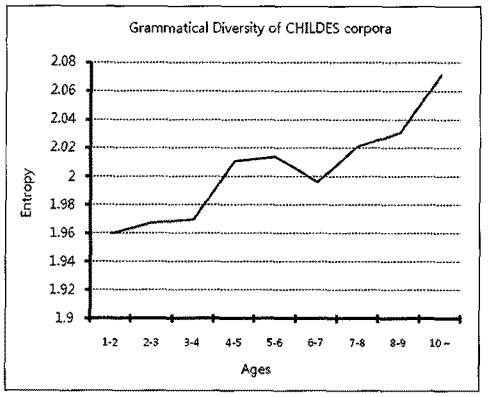

그림 5 CHILDES 데이터베이스의 자료를 연령별로(2세-성인) 9개의 그룹으로 나누어 엔트로피를 계산한 결과. 그림 4와 그림 5를 비교해 보았을 때, 훈련에 사용한 코퍼스의 복잡도는 엔트로피 측면에서 아동이 사용하는 언어 복잡도와 유사하다.

3.2 문장 생성 실험 결과 및 분석

훈련 결과 총 91,000개의 훈련 문장에 대하여 기수=3(하이퍼에지를 구성하는 노드의 수 = 3)의 하이퍼에지를 생성시킨 후 코퍼스를 공급하였다. 대략 3,000,000개의 하이퍼에지가 생성되었으며, 생성된 하이퍼에지에는 표 1의 알고리즘에 따라 가중치를 부여하였다.

문장 생성 실험을 위하여 D1~D9의 코퍼스에서 200개의 문장을 선정한 후, 선정된 문장은 하이퍼네트워크 모델 훈련에 사용하지 않고 테스트용 문장 집합을 구성하였다. 그림 6은 문장 생성을 위하여 주어진 문장의 예이다.

D1~D9까지 9개의 코퍼스 그룹에 대하여 새로운 코퍼스가 공급될 때마다 하이퍼네트워크 모델을 학습시킨 후 학습된 모델을 이용하여 200개의 테스트 문장에 대

thomas and _ _ working in the _
_ _ how do _ _ the eggs
one morning _ _ out of _ window
_ macdonald is _ _ _ _ _ _ _
today we _ going to _ _ garden
_ you think _ _ beat us all by _
the gas house _ are a _ of _ players
how does _ _ _ the _ look like
hey _ why did _ _ wake _ up
later the _ _ met gordon _ james

그림 6 문장 생성을 위한 입력 문장 예. 밑줄로 표시된 부분은 언어 모델로 채워야 할 부분이다. 원래 문장의 단어 수 중 최대 50%에 해당하는 단어를 무작위로 선정하여 선정된 부분을 삭제하는 방식으로 테스트 문장을 생성하였다.

문법적으로 성립하는 문장의 예
Look at all the snow
Gilbert what are you doing here
A place to sing and dance
Later the two engines met Gordon and James

문법적으로 성립하지 않는 문장의 예
I am the for everyone to wear
Let me in fact said garden
It has your come on it

그림 7 생성된 문장의 예. 문법적으로 성립하면서 의미적으로도 성립하는 문장, 문법적으로 성립하나 의미적으로는 성립하지 않는 문장, 문법적으로도 의미적으로도 모두 성립하지 않는 문장이 생성된다. 본 논문에서는 문법적으로 성립하는지 여부에 집중하였다.

하여 문장을 생성하는 실험을 9회 반복하였다. 2명의 평가자가 생성된 문장을 평가한 후 의미적으로 성립하지는 않더라도 문법적으로 성립하는 문장의 비율을 판정하는 방식으로 학습 모델의 성능을 평가하였다.

이 실험은 제안 방법의 문법 규칙 학습 능력을 확인하기 위한 것으로, 하이퍼에지의 가중치만으로 문법적으로 유의미한 문장을 생성할 수 있는 단어간 관계 패턴 생성 능력을 확보할 수 있음을 보이고자 하였다.

그림 7에 생성된 문장의 예를 보이고 있다. "A place to sing and dance"와 같이 문법적으로 유효하면서 의미적으로도 유의미한 문장이 생성되는 경우도 있으며, "Later the two engines met Gordon and James"라는 문장의 경우 컨텍스트에 따라 의미적으로 유의미한 것인지 변화할 수 있으나 문법적으로는 유효한 문장이 생성되기도 한다. 본 논문에서는 의미론적 유의미함의 여부는 고려하지 않고 문법적 유효성만을 고려하였다.

그림 8에서 문법적으로 성립하는 것으로 판정된 문장의 비율을 보이고 있다. D1(Miffy)만을 이용하여 하이퍼네트워크 모델을 구축한 경우 문법적으로 성립하는 문장이 20% 초반에 불과하였지만, 주어진 코퍼스를 모두 학습에 사용한 결과 생성된 문장의 60%에 가까운 문장들이 문법적으로 성립함을 확인할 수 있었다. 그림 8에서는 하이퍼네트워크 모델 학습 코퍼스가 늘어남에 따라 문장 생성 능력이 높아진다는 것을 보여준다. 학습이 진행됨에 따라 코퍼스에서의 등장 빈도가 높은 단어 패턴(하이퍼에지)에 높은 가중치가 부여되고, 문법적으로 유효한 단어 패턴이 살아남을 확률이 높아지며 따라서 유효한 문장 비율이 높아진다.

표 2에서 가중치가 높은 하이퍼에지의 예를 보이고 있다. 제시된 하이퍼에지에는 연달아 존재하는 것으로 유추되는 단어 조합도 있으며 연달아 존재하는 단어 패턴은 아닌 패턴들도 존재한다. 즉 순차적 랜덤 샘플링 방식을 사용한 결과 n-그램 모델의 샘플링 효과를 포함하면서 표현력이 더 큰 샘플링 방식을 사용할 수 있게 되었음을 알 수 있다.

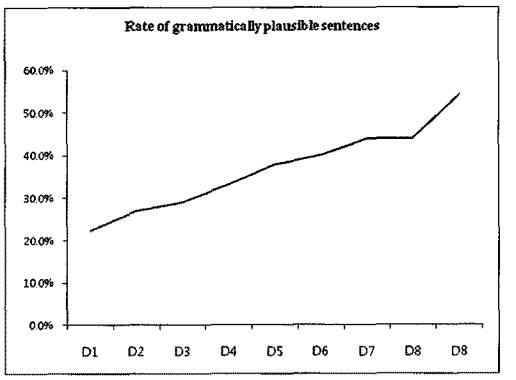

그림 8 문법적으로 성립하는 문장의 비율. 200개의 테스트 문장에 대하여 두 명의 평가자가 문법적 성립 여부를 평가한 결과. 두 명의 평가자가 문법적으로 성립 여부를 판정한 후 판정 결과의 평균치를 결과로 제시하였다.

표 2 가중치가 높은 하이퍼에지

Oh my god	I don't know
You know what	Hoo hoo hoo
Wait a minute	I love you
Are you doing	What is it
What are you	Here we go
What is that	There you go
I am not	What you doing
What do you	I got it

생성된 문장의 문법 규칙 분포를 관찰하기 위하여 스탠포드 파서를 이용하여 생성된 문장의 문법 규칙을 탐색한 후, 탐색된 문법 규칙의 엔트로피를 계산해 보았다(그림 9). 그림 8과 그림 9의 결과를 종합해 볼 때, 코퍼스가 공급됨에 따라 문법 규칙이 축적되고 축적된 문법 규칙의 복잡도가 높아짐에 따라 또한 문법적으로 성립하는 문장의 비율이 높아진다는 것을 확인할 수 있다. 즉 언어 모델의 복잡도가 높아지면서 다양한 문법 규칙을 표현할 수 있게 되는 것이다.

기계 학습의 학습 목표는 주어진 훈련 데이터의 분포를 학습하는 것이다. 코퍼스의 문법 규칙 분포에 대한 하이퍼네트워크 알고리즘의 학습 능력을 확인하기 위하여 학습 코퍼스의 문법 규칙 분포와 생성된 문장의 문법 규칙 차이를 KL 다이버전스(Kullback-Leibler divergence, $D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}dx$)로 확인하였

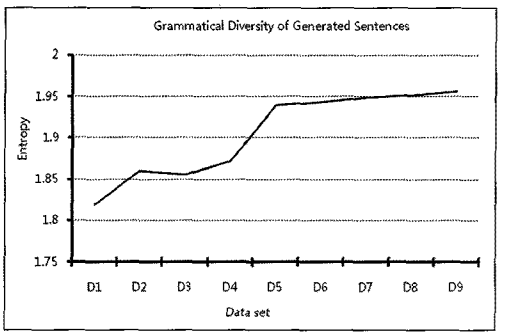

그림 9 생성된 문장의 엔트로피. 스탠포드 파서를 이용하여 생성된 문장의 문법 규칙을 찾은 후 문법 규칙의 복잡도를 엔트로피로 표현하였다. 생성된 문장의 엔트로피 분석을 통해 문법을 통해 문장 생성의 기반이 되는 하이퍼네트워크 모델의 복잡도가 높아졌음을 유추할 수 있다.

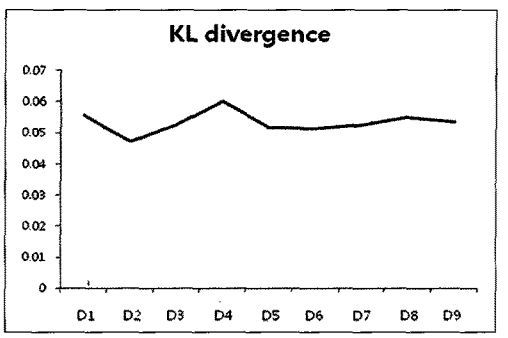

그림 10 학습 코퍼스의 문법 규칙 분포와 생성된 문장의 문법 규칙 분포 비교. 스탠포드 파서를 이용하여 획득한 문법 규칙의 분포에 대하여 KL 다이버전스를 계산한 결과

다. 그림 10에서 KL 다이버전스를 계산한 결과를 보이고 있다. 계산 결과 학습에 사용한 코퍼스의 문법적 규칙 분포와 생성된 문장의 파싱 결과 획득된 문법적 규칙 분포 차는 0.06이하의 차이를 갖고 있다.

스탠포드 파서를 이용하여 생성된 문장의 문법 규칙을 분석해 보았다. 즉 생성된 파싱 트리에서 각 노드를 루트로 하고 루트 노드의 바로 아래 노드를 구성원으로 하는 서브 트리에 해당하는 문법 규칙을 추출한 후 분석하였다. 표 3은 파싱 트리에서 발생 빈도가 높은 상위 10개의 문법 규칙을 나열한 것이다. 그림 11은 해당 10개 문법 규칙의 발생 빈도수 변화 그래프이다.

그림 11에서는 학습이 진행됨에 따라 구체적으로 어떤 변화가 발생하는지 등장 빈도가 높은 문법 규칙을 이용하여 보이고 있다. G1~G10까지 10개 문법 규칙은

표 3 생성 문장에서 관찰된 문법 규칙 상위 10개

G1. S = NP + VP	G2. NP = PRP
G3. S = VP	G4. PP = IN + NPNP
G5. NP = DT + NN	G6. ADVP = RB
G7. NP = NP + PP	G8. SBAR = S
G9. VP = VB + NP	G10. NP = FW

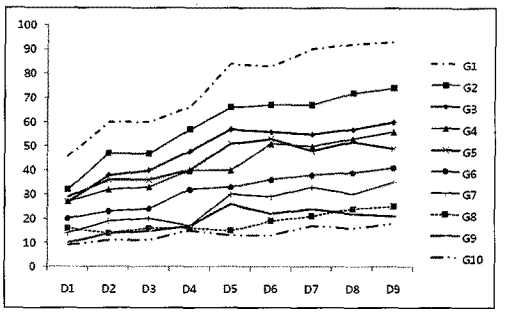

그림 11 상위 10개 문법 규칙의 빈도수 변화. 이 그림에서는 스탠포드 파서를 이용하여 분석된 문법 규칙 중 상위 10개 문법 규칙의 발생 빈도수 변화를 보이고 있다.

분석된 문법 규칙 중 43%의 등장 빈도를 차지하는 문법 규칙들로 대체적으로 문법적으로 성립하는 문장의 비율이 늘어남에 따라 문법 규칙의 발생 빈도도 높아진다. 그러나 각 문법 규칙의 변화 양상이 일정한 것은 아니다. 예를 들어 G3, G5, G7, G9의 문법 규칙의 경우 발생 빈도 수가 단조 증가하지 않는다. 발생 빈도가 대체적으로 증가하기는 하지만 단조 증가하지 않는 현상은 다음과 같은 두 가지 의미를 갖는다. 첫째, 이와 같은 변화 양상은 제안 알고리즘의 동작 특성을 나타낸 것이다. 제안 알고리즘의 경우 기억 감소(memory decay) 효과를 모사하기 위하여 가장 최근에 주어진 훈련 코퍼스에 존재하는 하이퍼에지의 가중치가 존재하지 않는 하이퍼에지의 가중치보다 높도록 설계되었다. 따라서 코퍼스의 규칙 분포에 따라 문법 규칙을 구성하는 하이퍼에지의 가중치가 변화하면서 그림 11과 같이 빈도수가 낮아지는 현상이 나타나게 된다. 둘째, 훈련 코퍼스를 구성하는 문법 규칙의 분포가 동일하지 않다는 사실을 나타낸다. 훈련 코퍼스의 복잡도(그림 4)에서도 알 수 있지만, 훈련 코퍼스를 구성하는 문법 규칙의 분포는 매우 다양하다. 따라서 그림 11과 같은 변화 양상이 발생한다.

3.3 실험 결과에 대한 토의

본 논문에서는 순차적 랜덤 샘플링을 통해 구축한 하이퍼네트워크 모델을 이용하여 언어의 문법적 특성을 학습하는 연구를 수행하였다. 문법적 지식을 사전에 부

여하거나, 템플릿을 사용하는 등 언어 처리 능력을 사전에 부여하지 않은 상태에서 하이퍼네트워크 샘플링을 통해 습득한 단어 관계 패턴만을 이용하여 문장을 생성하는 것으로 제안된 모델의 언어 처리 능력을 확인하였다. 본 논문의 의의는 다음과 같다.

첫째, 학습을 위하여 사용되는 코퍼스를 누적 증가시키는 것과 같은 효과를 갖도록 학습을 진행한 결과, 단어 관계 패턴이 증가함에 따라서 문장 생성 능력이 향상됨을 확인할 수 있었다. 이는 코퍼스 규모의 중요성을 다시 한 번 확인시켜 주는 결과로 보다 큰 규모의 코퍼스를 공급할 경우 문장 생성 능력이 어느 정도 향상될 것인지 추가 확인이 필요하다.

둘째, 본 논문에서는 순차적 랜덤 샘플링이라는 단순한 방법으로도 문장 생성 능력을 갖춘 언어 모델을 생성할 수 있음을 확인하였다. 본 논문에서 사용한 방법은 단어 조합을 만든 후, 학습 코퍼스에서의 관찰 빈도를 통해 가중치를 조정하는 단순한 방법에 불과하며, 태깅을 이용한 부가 정보를 활용하지 않았다. 그럼에도 불구하고 학습된 언어 모델에 문장 생성 능력이 존재함을 확인하였으므로, 학습을 통해 유사한 성질을 갖춘 단어를 판별하고 이를 통해 문법 규칙을 만들어 갈 수 있는 능력을 갖춘다면 언어 모델을 성능을 한층 높일 수 있을 것이라고 기대할 수 있다.

4. 관련 연구

언어 습득 및 활용에 대해서는 다양한 이론이 존재한다. 어떤 연구자들은 인간에게 내재된 언어 특화 능력이 언어 학습을 조정한다고 주장하기도 한다[8]. 다른 학자들은 언어 학습 능력 역시 일반적인 학습 능력의 한 형태에 불과하며 학습자와 주변 환경과의 상호작용을 통해 언어를 습득하게 된다고 주장한다[9,10].

본 장에서는 이와 같은 논란에 실마리를 제시하기 보다는 전산학 측면에서의 언어 처리에 초점을 맞추어서 언어 습득 및 처리를 살펴보고자 한다. 이를 위해 특히 통계적 측면에서 언어 습득을 설명하고자 한다.

언어에는 다양한 형태의 규칙성이 존재하며, 일종의 척도를 이용하여 발견된 규칙성을 표현하는 것이 가능하다. 인간 역시 언어의 통계적 규칙성을 이용하여 언어를 학습한다는 연구가 제시되고 있다[11]. 즉 성인 뿐 아니라 유아 역시 강력한 통계적 학습 능력을 보유하고 있고 이를 이용하여 언어를 습득한다는 것이다[12].

이와 같은 실세계 연구와 궤를 같이하여 통계적 언어 모델에서는 자연언어의 규칙성을 발견하고자 한다. 통계적 언어 모델은 가능한 문장 s에 대한 확률 분포 $P(s)$를 찾는 것을 목표로 한다[13]. 통계적 언어 학습은 주로 베이지안 법칙(Bayesian Law)의 문맥에서 사용된다. 예를 들어 음성 신호 A가 주어졌을 때 음성 신호 A에 대한 언어 모델 $P(L)$은 아래 식에서 사전 확률(prior)의 역할을 수행하게 되는데

$$L^* = \text{argmax}_L P(L|A) = \text{argmax}_L P(L|A) \cdot P(L)$$

여기서 $L = w_1^n \equiv w_1 w_2 \ldots w_n$ 이라고 하면 $P(L)$을 표현하는 한 가지 방법은 체인룰을 이용하는 것이다.

$$P(L) = \Pi_{i=1}^{n} P(w_i \mid w_1^{i-1}) \quad (4)$$

대부분의 통계적 언어 모델에서는 식 (4)를 추정하고자 한다. 식 (4)를 추정하는 방법으로 가장 흔히 사용되는 것이 n-그램[6] 모델이다. n-그램 모델에서는 현재 관찰하는 단어의 이전 $n-1$ 개의 단어에 집중한다. n-그램 모델은 현재 관찰하고 있는 단어를 분석하기 위하여 문장의 시작에서부터 관찰 중인 단어 이전 단어까지 모두 관찰하는 것이 아니라 그 중 일부 단어만을 이용하여 이력을 근사하는 것을 목표로 한다. 즉 $P(w_i|w_1, \ldots, w_{i-1}) \approx P(w_i|w_{i-n+1}, \ldots, w_{i-1})$을 이용하여 추정 문제의 차원을 줄이는 것이 n-그램 모델의 목표이다. 다양한 분야에서 n-그램 모델이 널리 사용되고 있기는 하지만[14], n-그램 모델에는 몇 가지 단점이 존재한다[15].

첫째, n-그램 모델은 n이라는 제한된 범위 바깥에서 일어나는 현상이나 제약 조건에 무지하다.

둘째, n-그램 모델에서 단어 예측은 언어적 역할이 아닌 문장에서의 위치에 의해 결정된다.

이와 같은 n-그램 모델의 문제를 해결하기 위하여 의존 관계 문법(dependency grammar)[16]이 제안되었다. 이 모델에서는 문장 S 와 연결관계(Linkage) K의 결합확률인 $P(S,K)$를 이용하여 문장을 표현하는데, 연결관계란 단어 및 해당 단어 및 다른 단어와의 문법적 관계성을 의미하는 것으로 하나 하나의 관계성을 d^i라고 표현할 때 결국 결합확률은

$$P(S,K) = \prod_{i=0}^{n} P(w^i d^i | h^i)$$

(여기서 $h^i = w^0 d^0 \ldots w^{i-1} d^{i-10}$)로 표현된다.

의존 관계 그래프와 하이퍼네트워크 모델은 n이라는 제약을 깨고 보다 넓은 범위의 단어간 관계성을 고려한다는 점에서 유사하다[17]. 그러나 제안 방법론은 태깅 등을 이용하여 품사 정보를 추가하거나 별도의 문법 규칙을 학습에 활용하는 것이 아니라 주어진 코퍼스의 텍스트 데이터만을 활용하여 단어 관계 패턴을 구성하고 이에 바탕하여 문장 생성을 시도했다는 점에서, 기존의 연구와 다른 독창성이 있다.

컴퓨터에 기반하여 자연 언어 생성을 하는 경우 어떤 이론에 기반하고 있느냐에 따라 모듈의 복잡도가 틀려진다[18]. 유연성에 따라 분류하면 캔드 텍스트 시스템(canned text systems), 템플릿 시스템(template sys-

tems), 프레이즈 기반 시스템(phrase-based systems)로 구분할 수 있다. 캔드 텍스트 시스템은 사전에 입력된 문장을 그대로 재생성하는 것이며, 템플릿 시스템은 템플릿에 기반하여 아주 사소한 변화만 부여하여 문장을 생성하고, 프레이즈 기반 시스템에서는 입력을 패턴으로 구분한 후 보다 자세한 패턴으로 분리하여 문장을 생성해 나간다.

하이퍼네트워크를 이용한 문장 생성의 경우 훈련 코퍼스에서 획득한 단어 관계 패턴을 활용하기 때문에 기존의 문장 생성 방식과 같은 제약에서 자유롭다. 기존의 하이퍼네트워크 활용 문장 처리 연구에서는 훈련 코퍼스의 문장을 리콜(recall)하거나[17], 하이퍼네트워크 방법론을 이용하여 유아의 단계적 언어 능력 발달 모습을 모사해 보았고[19] 코퍼스를 변화시키면서 문장을 생성하여 보았다[20]. 본 논문에서 제안하는 방법의 경우 [17]과는 달리 훈련 코퍼스와는 다른 별도의 테스트 문장을 생성하며, 생성 문장의 난이도 측면에서 [19]와 차이가 있고, 생성 방법 및 분석에 있어 [20]의 연장선에 있다.

로봇 축구 경기 중계[21]나 기상 예보[22] 등에서 자연 언어 생성 프로그램이 사용되고 있지만 아직 충분한 자유도를 갖고 있는 자연 언어 생성 프로그램의 등장은 요원한 상태이다. 하이퍼네트워크 모델을 이용한 문장 생성의 경우 학습에 사용한 코퍼스의 어휘 수준 및 단어 관계의 패턴을 문장 생성에 사용할 수 있는 잠재력을 지니고 있으므로 자연 언어 생성에 있어 보다 높은 자유도를 구사할 수 있는 가능성을 갖고 있다.

5. 결론 및 향후 연구

본 논문에서는 하이퍼네트워크 모델을 이용하여 언어 모델을 학습한 후 학습된 모델을 이용하여 자연 언어를 생성하는 실험을 통해, 하이퍼네트워크 모델이 갖고 있는 언어 처리 능력을 확인해 보았다. 학습 모델의 복잡도가 높아짐에 따라 다양한 단어간 관계를 학습할 수 있음을, 문법적으로 성립하는 문장 비율을 통해 관찰하였으며, KL 다이버전스 계산을 통해 생성된 문장의 문법 규칙 분포가 훈련 코퍼스의 문법 규칙 분포와 유사함을 확인하였다. 본 연구를 통해 문법 규칙 학습의 잠재력은 확인하였지만, 어떤 문법 규칙들이 학습되었는지 탐색할 수는 없었다. 향후 연구에서는 학습된 하이퍼에지를 결합하여 학습된 문법 규칙을 생성할 수 있는 방법을 고안하여, 단어 관계 패턴에서 추상화된 문법 규칙을 도출할 수 있는 방법을 찾아보고자 한다.

참 고 문 헌

[1] E. Reiter and R. Dale, "Building Applied Natural Language Generation Systems," *Natural Language Engineering*, vol.3, no.1, pp.57-87, 1995.

[2] D. Traum, M. Fleischman, and E. Hovy, "NL Generation for Virtual Humans in a Complex Social Environment," *Proceedings of the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*, pp.151-158, 2003.

[3] B.-T. Zhang, "Hypernetworks: A Molecular Evolutionary Architecture for Cognitive Learning and Memory," *IEEE Computational Intelligence Magazine*, vol.3, no.3, pp.49-63, 2008.

[4] http://childes.psy.cmu.edu/data/

[5] D. J. C. MacKay, "Information Theory, Inference, and Learning Algorithms," Cambridge University Press, 2005.

[6] L. Bahl, F. Jelinek, and R. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.5, no.2, pp. 179-190, 1983.

[7] http://nlp.stanford.edu/software/lex-parser.shtml

[8] N. Chomsky, "Aspects of the Theory of Syntax," MIT Press, 1965.

[9] R. L. Gmez and L. Gerken, "Infant Artificial Language Learning and Language Acquisition," *Trends in Cognitive Sciences*, vol.4, no.5, pp.178-186, 2000.

[10] D. McAllester and R. E. Schapire, "Learning Theory and Language Modeling," Exploring Artificial Intelligence in the New Millennium, Morgan Kaufmann Publishers, pp.271-287, 2003.

[11] J. R. Saffran, "Statistical Language Learning: Mechanisms and Constraints," Current Directions in Psychological Science, vol.12, no.4, pp. 110-114, 2003.

[12] C. Yu and D. H. Ballard, "A Unified Model of Early Word Learning: Integrating Statistical and Social Cues," *Neurocomputing*, vol.70, pp.2149-2165. 2007.

[13] R. Rosenfeld, "Two Decades of Statistical Language Modeling: Where Do We Go From Here?," *Proceedings of IEEE*, vol.88, pp.1270-1278, 2000.

[14] A. L. Buchsabau and R. Giancarlo, "Algorithmic Aspects in Speech Recognition: An Introduction," *Journal of Experimental Algorithmics*, vol.2, no.1, pp.1-44, 1997.

[15] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer Speech and Language*, vol.10, pp.187-228, 1996.

[16] C. Chelba, et al, "Structure and Performance of a Dependency Language Model," *Proceedings of Eurospeech 97*, pp.2775-2778, 1997.

[17] B.-T. Zhang and C.-H. Park, "Self-Assemblying Hypernetworks for Cognitive Learning of Linguistic Memory," *Proceedings of International Con-*

ference on Computer, Electrical, and Systems Science and Engineering, vol.27, pp.134-138, 2008.
[18] C. Mellish and R. Evans, "Implementing Architecture for Natural Language Generation," *Natural Language Engineering*, vol.10, no.3/4, pp. 261-282, 2004.
[19] J.-H. Lee, E. S. Lee, and B.-T. Zhang, "A Hypernetwork Memory-based Model of Sentence Learning and Generation in Children: How a Child Learns to Produce Language from a Video Corpus," *KCC2009*, pp.134-138, 2009.
[20] H.-S. Seok, J. Bootkrajang, and B.-T. Zhang, "Automatic Grammar Induction by Incrementally Learning a Hypernetwork Model: Sentence Generation and Analysis," *The 36th KIISE Fall Conference*, pp.221-224, 2009.
[21] D. L. Chen and R. J. Mooney, "Learning to Sportcast: A Test of Grounded Language Acquisition," *Proceedings of the 25th International Conference on Machine Learning*, pp.128-135, 2008.
[22] E. Reither, S. Sripada, J. Hunter, J. Yu, and I. Davy, "Choosing Words in Computer-generated Weather Forecasts," *Artificial Intelligence*, vol.167, no.1/2, pp.137-169, 2005.

석 호 식
1999년 서울대학교 컴퓨터공학 학사. 2001년 서울대학교 컴퓨터공학 석사. 2001년~2004년 육군사관학교 전산학과. 2004년~현재 서울대학교 컴퓨터공학부 박사과정. 관심분야는 기계학습, 자연언어 획득, 문장 생성

작 가 멧
2008년 서울대학교 컴퓨터공학 학사. 2010년 서울대학교 컴퓨터공학 석사. 2010년~현재 University of Birmingham, Department of Computer Science, Postgraduate research. 관심분야는 기계학습, Text Mining, Information extraction

장 병 탁
정보과학회논문지 : 소프트웨어 및 응용
제 37 권 제 2 호 참조