하이퍼네트워크 모델의 점진적 학습에 의한 문법의 자동 생성: 문장 생성 및 분석

석호식, 작가멧, 장병탁
서울대학교 컴퓨터공학부 바이오지능연구실
hsseok@bi.snu.ac.kr, jakramate@bi.snu.ac.kr, btzhang@cse.snu.ac.kr

Automatic Grammar Induction by Incrementally Learning a Hypernetwork Model: Sentence Generation and Analysis

Ho-Sik Seok, Jakramate Bootkrajang, Byoung-Tak Zhang
Biointelligence Lab., School of Computer Sci. and Eng., Seoul National University

요 약

본 논문에서는 학습에 활용하는 코퍼스(Corpus)의 난이도와 양을 증가시키면서 하이퍼네트워크에 기반한 문장 생성 모델을 학습시킨 후, 모델 학습 단계 별로 생성된 문장의 문법적 특성을 분석한다. 하이퍼네트워크 모델은 가중치를 갖고 있는 에지와 노드의 집합으로 코퍼스의 확률 분포를 학습할 수 있는 한가지 방법을 제공한다. 본 논문에서는 생성되는 문장의 문법적 특성의 변화를 통해 언어 모델의 특성을 분석하며 원래 코퍼스와 생성된 문장의 문법 규칙 분포에 대한 KL divergence를 계산하여 학습 데이터와 언어 모델에 축적된 문법 규칙의 분포간 차이를 확인하였다. 학습 데이터와 언어 모델에 축적된 문법 규칙의 분포에 대하여 KL divergence를 계산한 결과 최대 0.06을 초과하지 않았다. 언어 모델에서 생성된 문장의 파싱을 분석해 본 결과, 학습이 진행됨에 따라 생성되는 문장의 문법적 타당성이 점차 증가함을 관찰하였으며, 코퍼스의 특성으로 인해 학습 초기 과정에 문법 규칙이 일정한 분포로 수렴함을 확인하였다.

1. 서론

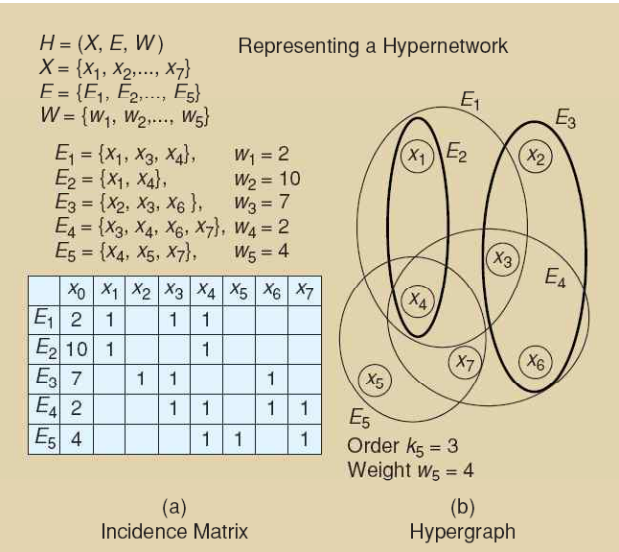
본 논문에서는 하이퍼네트워크 모델[1]에 기반하여 언어 모델을 학습한 후 학습된 모델에서 문장을 생성하는 실험을 소개한다. 기존의 자연언어 학습 모델[2]에서는 내재된 문법 규칙을 사용하거나 템플릿을 사용하여 자연 언어 문장을 생성하기 때문에 문장 생성의 유연성이 매우 낮다. 그러나 본 논문에서 소개하는 방법에서는 코퍼스 분석을 통해 단어간 관계성을 분석하고, 분석된 관계성에 기반하여 문장을 생성하기 때문에 매우 표현력이 큰 방법을 사용하여 문장을 생성할 수 있다.

본 논문에서 사용한 하이퍼네트워크 방법은 가중치가 있는 하이퍼그래프를 이용하여 단어 조합 패턴의 확률 분포를 찾는 방법이다. 하이퍼네트워크 방법론을 이용하면 단어 조합이 생성되기 때문에 일련의 단어가 큐로 주어졌을 때 주어진 큐와 관련성이 높은 단어 조합을 이용하여 순차적으로 단어를 추가하여 문장을 생성할 수 있다. 본 논문에서 사용한 문장생성 방식을 통해 훈련 코퍼스와 문법 규칙 분포가 근사한 문장을 생성할 수 있음을 확인하였다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 사용 방법론인 하이퍼네트워크를 소개하고, 3장에서는 문장 생성 결과를 제시한다. 그리고 4장에서는 결론 및 향후 추가할 연구 방향에 대하여 논의한다.

2. 방법론

하이퍼네트워크는 가중치가 있는 하이퍼그래프로 노드 간의 상호작용은 하이퍼에지로 표현된다. 하이퍼네트워크 $H=(X,E,W)$는 노드 집합 $X=\{x_1,x_2,...,x_l\}$, 에지 집합 $E=\{E_1,E_2,...,E_{|E|}\}$, 가중치 집합 $W=\{w_1,w_2,...,w_{|E|}\}$로 정의된다. 하이퍼네트워크는 데이터 $D=\{x^{(n)}\}_{n=1}^{N}$를 저장하는 확률적 연상 메모리로 사용될 수 있다.



	x0	x1	x2	x3	x4	x5	x6	x7
E1	2	1		1	1			
E2	10	1			1			
E3	7		1	1			1	
E4	2			1	1		1	1
E5	4				1	1		1

그림 1 하이퍼네트워크 모델 구축의 예. 훈련을 위하여 제공된 코퍼스 문장에서 무작위로 k개의 단어를 추출하여 오더 k의 하이퍼그래프를 생성한 후 훈련을 통해 생성된 하이퍼그래프의 가중치를 조정한다.

표 1 하이퍼네트워크의 학습 알고리즘

초기화

코퍼스에서 문장 획득

획득된 문장에서 무작위로 단어를 추출한 후 단어간 관계성을 분석하여 하이퍼에지 $E=\{E_1,E_2,...,E_{|H|}\}$ 생성

학습

1. 가중치 부여

(네트워크는 반복적으로 문장을 관찰하여 학습됨)

1.1 문장 $x^{(n)}$을 관찰한 후 질의 하이퍼에지 $E^{(q)}$를 반복 샘플링하여 생성된 하이퍼네트워크 H의 에지와 비교.

1.2 에지와 비교한 후 매칭되는 에지의 가중치를 증가.

학습 과정에서의 목표

학습에 사용된 문장 $x^{(n)}$을 선택한 후 선택 문장에서 질의 에지를 생성한 후 문장 생성 결과 $x^{(n)}$이 생성될 때까지 가중치 조정

제안된 방법은 샘플링 방식에 따라 기존에 널리 널리 사용되는 자연 언어 분석 방법인 *n*-그램[3] 방식을 모사할 수도 있다. 학습 과정에서 샘플링되는 단어를 단어수 *n*의 윈도우에서 추출하게 될 경우 *n*-그램 모델에서 사용한 것과 유사한 효과를 기대할 수 있다. 만약 크기 *n*의 윈도우 제한을 넘어서서 문장 전체에서 무작위로 단어 조합을 추출할 경우 *n*-그램 모델의 단점을 극복하는 단어 관계성 분석이 가능하다.

3. 실험 결과

3.1 사용 코퍼스

실험에 사용한 코퍼스는 유아용 비디오의 스크립트로 사용한 유아용 비디오는 Miffy [4], Looney tunes [5], Caillou [6], Dora Dora [7], Macdonald's farm, Thomas & Friends [8], Timothy, Pooh [9]의 8종이며 유아용 비디오보다 문장 수가 월등히 많은 시트콤 Friends의 스크립트 데이터가 추가되었다. 실험에 사용된 코퍼스는 총 $O(100K)$ 의 문장으로 구성되어 있다. 그림 2에서 알 수 있듯 코퍼스가 늘어나면서 문법 규칙의 다양성이 늘어나기 때문에 코퍼스 증가에 따라 난이도가 높아지는 문제에 해당한다.

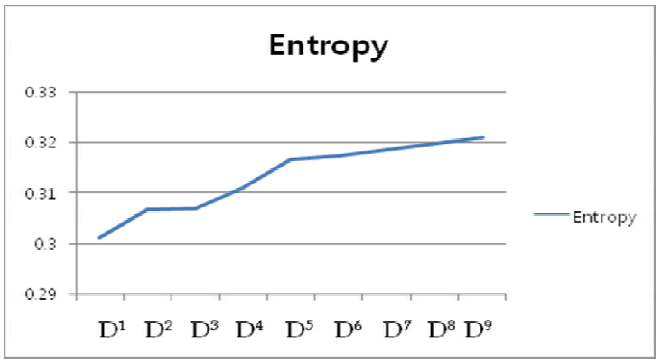


그림 2 엔트로피를 이용하여 표현한 코퍼스의 문법 규칙 다양성. 코퍼스가 늘어날수록 문법 규칙의 다양성이 높아짐을 알 수 있다(D^1= Miffy, $D^2=D^1$+ Looney, $D^3=D^2$+ Caillou, $D^4=D^3$+ DoraDora, $D^5=D^4$+ Macdonald, $D^6=D^5$+ Thomas, $D^7=D^6$+ Timothy, $D^8=D^7$+ Pooh, $D^9=D^8$+ Friends).

3.2 실험 및 결과

실험에 사용한 스크립트의 사용 나이에 따라 최초에는 Miffy 만 사용(D^1=Miffy), 두 번째 단계에서는 D^1에 Looney 추가 사용($D^2 = D^1$ + Looney)과 같이 실험에 사용하는 코퍼스를 점차 추가시켰다. 실험 순서는 다음과 같다. D^1= Miffy, $D^2 = D^1$ + Looney, $D^3 = D^2$ + Caillou, $D^4 = D^3$ + DoraDora, $D^5 = D^4$ + Macdonald, $D^6 = D^5$ + Thomas, $D^7 = D^6$ + Timothy, $D^8 = D^7$ + Pooh, $D^9 = D^8$ + Friends.

문장 생성은 "I love * * * *"(*: 공란 의미)와 같이 문장 일부에 단어가 존재하는 방식으로 질의 문장이 주어지면 하이퍼네트워크 H 에서 학습된 언어 모델을 활용하여 공란을 채우는 방식으로 진행된다.

하이퍼네트워크의 언어 모델에 "love + you…(1)", "love + sentimental…(2)"과 같은 단어가 존재하고 (1)의 가중치가 (2)보다 클 경우 "you"가 선택되어 "I love you * * *"를 만들게 된다. 이와 같은 방식으로 기존에 존재하는 단어를 갖고 있는 단어 조합 중 가중치가 높은 하이퍼에지를 선택하여 공란을 채우는 방식으로 문장 생성을 진행하게 된다.

표 2 생성된 문장: 양호한 문장 및 양호하지 못한 문장

(양호한 문장)
On my first day of school.
Yes timothy it is time to go to school.
Thomas and Percy enjoy working in the spotlight.
Well it is morning.
(양호하지 못한 문장)
He couldn't way to go outside and shoot.
The gas house gorillas are a lot of fun players.

표 2에서는 생성된 문장의 예를 보여주고 있다. 표 2에서 알 수 있듯 문법적으로 타당할 뿐 아니라

의미적으로도 성립하는 문장이 생성되기도 하며, 문법적·의미적으로 모두 성립하지 않는 문장이 생성되기도 한다.

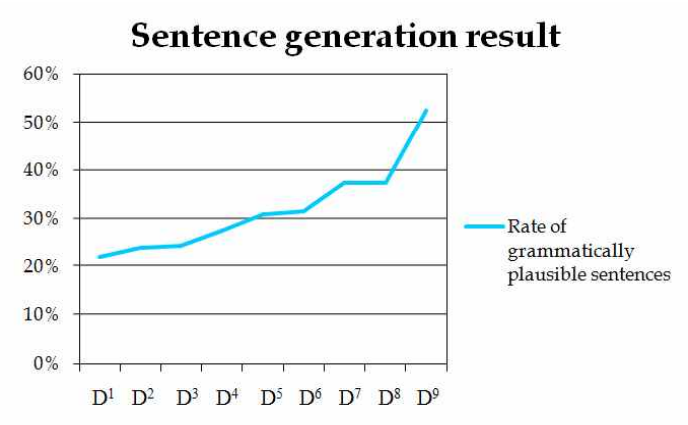


그림 3 문법적으로 유의미한 문장 비율
(D*의 의미는 그림2의 경우와 동일)

그림 3은 생성된 문장 중 문법적으로 가능한 문장의 비율을 보여주고 있다. 200개 테스트 문장이 주어지고 평가자 1명이 생성된 문장의 문법적 타당성을 직접 평가한다. 생성된 문장의 의미적 타당성은 고려되지 않았으며 문법적 타당성만을 평가하였다. 그림 3에서 알 수 있듯 학습이 진행됨에 따라 문법적으로 타당한 생성 문장 비율이 높아지며, 문장이 많이 존재하는 시트콤 Friends를 반영하는 경우 문법적으로 타당한 문장의 비율이 50%를 초과한다.

그림 4는 식 (1)에 따라 KL divergence[10]를 계산한 결과를 보여준다. KL divergence는 서로 다른 확률 분포간의 차이를 보여주는 식으로 이번 실험에서는 학습에 사용한 코퍼스의 문법 규칙 분포와 생성된 문장의 문법 규칙 분포간 어느 정도 차이가 존재하는지 확인하기 위하여 계산되었다. 기본적으로 본 논문의 실험에서는 훈련 코퍼스의 문법적 규칙이 갖는 확률 분포와 동일한 분포를 갖는 테스트 문장을 생성하고자 한다. 따라서 KL divergence가 낮을수록 양호한 수치이다.

KL divergence :

$$D_{KL}(P||Q) = \sum_i P(i)\log P(i)/Q(i) \quad \text{...식(1)}$$

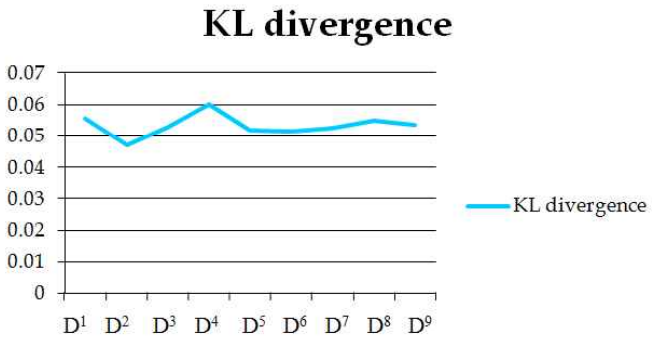


그림 4 KL divergence
(D*의 의미는 그림2의 경우와 동일)

그림 4에서 보여지듯 KL divergence 계산 결과 두 분포의 차이는 0.06을 초과하지 않는다. 문법 규칙의 분포는 스탠포드 파서[11]을 이용하여 문장을 파싱한 후 파싱된 결과에서 등장한 문법 규칙을 분석하는 방식으로 이루어졌다.

표 3 생성 문장 파싱 결과 분석: 파싱 트리에서 획득된 문법 규칙 중 빈도수가 가장 높은 문법 규칙

G1. S = NP + VP	G2. NP = PRP
G3. S = VP	G4. PP = IN + NPNP = NN
G5. NP = DT + NN	G6. ADVP = RB
G7. NP = NP + PP	G8. SBAR = S
G9. VP = VB + NP	G10. NP = FW
G11. NP = JJ + NN	G12. VP = VBZ + NP
G13. PRT = RP	G14. VP = VBP + NP
G15. NP = PRPS + NN	G16. NP = NNS
G17. VP = TO + VP	G18. ROOT = NP
G19. ROOT = FRAG	

표 3은 생성 문장의 파싱 결과를 분석하여 등장 빈도수가 가장 높은 문법 규칙을 내림 차순으로 19개 선정한 결과이다. 그림 5는 분석된 문법 규칙이 차지하는 비율을 보인 것이다. 생성 문장에서 획득된 문법 규칙 중 상위 19개 규칙이 전체 문법 규칙 빈도수의 80%를 차지하고 있음을 알 수 있다. 또한 각 문법 규칙이 차지하는 비율도 거의 일정하게 유지된다.

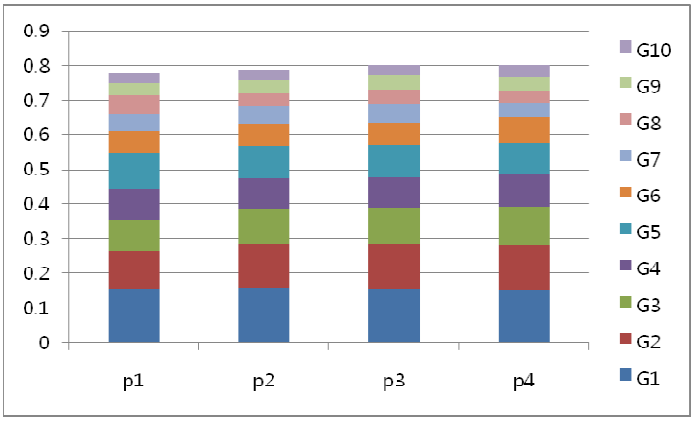


그림 5 생성된 문장을 파싱한 결과에서 상위 19개 문법 규칙이 차지하는 비율(그림 4의 경우 Miffy, Looney, Caillou, Dora Dora의 네 개 코퍼스를 분석한 결과임)

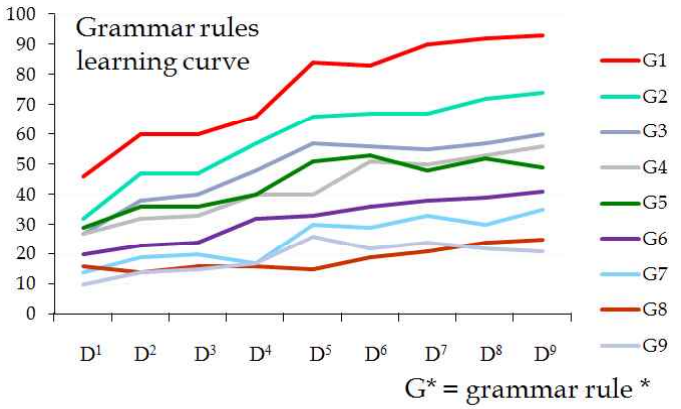


그림 6 학습 진행에 따른 파싱 결과에서의 상위 9개 문법 규칙의 관찰 빈도수 변화

그림 6은 학습이 진행됨에 따라 문법 규칙의 빈도수가 어떻게 변하는지 보이고 있다. 그림 5와 달리 그림 6은 관찰된 빈도수를 보인다. 각 코퍼스에서 문법 규칙이 등장하는 빈도수 비율이 거의 일정하기 때문에 코퍼스를 중첩하여 학습을 진행함에 따라 해당 문법 규칙의 등장 빈도수가 커지는 것을 관찰할 수 있다.

4. 결론

본 논문에서는 하이퍼네트워크에 기초하여 언어 모델을 생성한 후 생성된 언어 모델에 바탕하여 새로운 문장을 생성하는 실험을 통해 하이퍼네트워크 모델의 언어적 타당성을 확인하였다.

기존의 자연언어 생성 모델과 달리 본 논문에서 사용한 방법은 내재된 문법규칙과 템플릿의 존재를 가정하지 않고 있다. 그럼에도 불구하고 학습 코퍼스가 축적됨에 따라 문법적으로 타당한 문장이 생성되는 비율이 점차 높아짐을 확인할 수 있었다.

본 논문에서 소개한 하이퍼네트워크 기초 언어 모델은 훈련용 코퍼스에 존재하는 단어 조합이 갖는 패턴의 분포를 학습하는 것을 목표로 한다. 훈련 코퍼스와 생성된 문장의 단어 조합 패턴 분포를 확인하기 위하여 KL divergence를 계산한 결과 매우 낮은 차이를 보임을 확인하였다. 추후 연구에서는 가설 검증을 통해 두 분포가 동일한지 여부를 확인할 것이다.

본 논문에서 사용한 하이퍼네트워크 기반 언어 모델은 단어간 조합만을 고려할 뿐 한 하이퍼에지와 다른 하이퍼에지를 결합하는 방법을 고려하지 않고 있다. 일종의 은닉 변수를 이용하여 하이퍼에지와 하이퍼에지를 결합하거나, 문장 정체의 엔트로피를 고려하는 방법[12]을 활용한다면 생성 문장의 문법적 타당성을 더욱 높일 수 있을 것이라고 생각된다.

감사의 글

본 연구는 지식 경제부 한국산업기술평가관리원 산업원천기술개발사업(MARS, KEIT-2009-A1100-0901-1639, 교육인적자원부 학술진흥재단(KRF-2008-314-D00377), BK21사업의 지원을 받아 이루어졌습니다.

참고 문헌

[1] Zhang, B.-T. Hypernetwork: A molecular evolutionary architecture for cognitive learning and memory, *IEEE Computational Intelligence Magazine*, Vol. 3, No. 3, 49-53, 2008.

[2] Reiter, E. and Dale, R. Building applied natural language generation systems, Natural Language Engineering, Vol. 3, No. 1, 57-87, 1995.

[3] Bahl, L., Jelinek, F., and Mercer, R. A maximum likelihood approach to continuous speech recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 5, No.2, 179-190, 1983.

[4] http://www.miffy.com/

[5] http://looneytunes.kidswb.com/

[6] http://www.caillou.com/

[7] http://www.nickjr.co.uk/shows/dora/index.aspx

[8] http://www.thomasthetankengine.kr/

[9] http://www.winniethepoohbear.net/

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

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

[12] Rosenfeld, R., Chen, S. F., and Zhu, X. Whole-sentence exponential language models: a vehicle for linguistic-statistical integration, *Computer Speech and Language*, Vol. 15, No. 1, 55-73, 2001.