

비디오 스크립트를 이용한 문법적 패턴 습득 모델링

석호식⁰, 장병탁

서울대학교 컴퓨터공학부 바이오지능연구실

hsseok@bi.snu.ac.kr, btzhang@bi.snu.ac.kr

Modelling Grammatical Pattern Acquisition using Video Scripts

Ho-Sik Seok⁰, Byoung-Tak Zhang

Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University

요약

본 논문에서는 다양한 코퍼스를 통해 언어를 학습하는 과정을 모델링하여 무감독학습(Unsupervised learning)으로 문법적 패턴을 습득하는 방법론을 소개한다. 제안 방법에서는 적은 수의 특성 조합으로 잠재적 패턴의 부분만을 표현한 후 표현된 규칙을 조합하여 유의미한 문법적 패턴을 탐색한다. 본 논문에서 제안한 방법은 베이지안 추론(Bayesian Inference)과 MCMC (Markov Chain Monte Carlo) 샘플링에 기반하여 특성 조합을 유의미한 문법적 패턴으로 정제하는 방법으로, 랜덤하이퍼그래프(Random Hypergraph) 모델을 이용하여 많은 수의 하이퍼에지를 생성한 후 생성된 하이퍼에지의 가중치를 조정하여 유의미한 문법적 패턴을 탐색하는 방법론이다. 우리는 본 논문에서 유아용 비디오의 스크립트를 이용하여 다양한 유아용 비디오 스크립트에서 문법적 패턴을 습득하는 방법론을 소개한다.

주제어: 이용 기반 언어학(Usage based linguistics), 통계적 언어 학습, 문법적 패턴 습득, 랜덤하이퍼그래프 모델(Random Hypergraph Model), 기계학습 (Machine Learning)

1. 서론

인간의 언어 획득 과정을 모사할 때 해결해야 할 중요한 질문은 언어 습득 과정에서 어떤 능력이 사용되며, 이를 능력들은 어떻게 서로 상호 작용하는가의 문제이다. 기계 학습 관점에서 이 질문에 답변하기 위하여 우리는 자연 언어 코퍼스에 깔려 있는 문법적 패턴을 무감독학습(Unsupervised learning)을 이용하여 습득하는 방법을 소개한다. 본 논문의 제안 방법은 태그 등이 추가되지 않은 비가공 자연 언어 데이터를 이용하여 자연 언어의 구조적 특성(예. 문법 규칙)을 학습하는 방법으로, 본 연구와 같이 미가공 상태의 언어 데이터를 이용하여 자연 언어의 구조를 추론하려는 연구는 최근 많은 주목을 받고 있다 [1, 2]. 본 논문에서는 베이지안 추론(Bayesian inference)과 MCMC (Markov Chain Monte Carlo) 샘플링을 이용하여 자연 언어의 문법적 패턴을 탐색하고자 한다. 베이지안 추론 관점에서 언어 습득 현상을 설명하려는 시도는 최근 활발히 연구되고 있으며 다양한 연구 결과[3]를 통해 그 유용성을 확인 받고 있다. MCMC 샘플링은 자연 언어 처리와 같이 기저의 확률 분포 함수를 추정하기 어려운 문제를 처리할 때 큰 도움이 될 수 있는 방법으로 우리는 두 접근법을 결합하여 무감독학습으로 유의미한 패턴을 탐색한다.

본 제안 방법은 랜덤하이퍼그래프 모델[4] 방법론에 기반한 것으로 훈련용 코퍼스에서 무작위로 단어 패턴을 샘플링 한 후 샘플링한 단어 조합의 가중치를 조정하여 유의미한 패턴을 찾고 탐색된 결과를 문법적 패턴으로 재해석하는 일련의 과정을 거쳐 문법적 패턴을 도출한다. 우리는 유아용 비디오 9종의 스크립트를 훈련 및 테스트에 사용하였으며 학습된 문법적 패턴의 타당성은 문장의 일부만이 실마리로 주어진 문장이 주어졌을 때 주

어진 실마리를 바탕으로 문장을 새롭게 생성하고, 생성된 문장의 문법적 타당성을 실험자가 평가하는 방식으로 확인하고자 한다.

2. 제안 방법

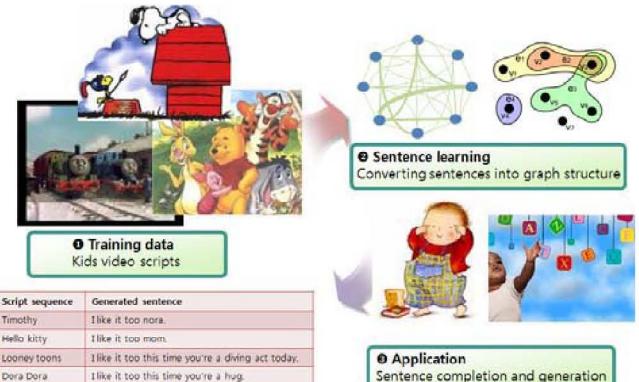


그림 1 제안 방법의 전체적인 설명. 유아용 비디오 스크립트를 분석하여(①) 문장을 구성하는 단어 조합을 무작위로 선정한 후 단어 조합의 가중치를 조정한다. 가중치에 기반하여 유의미한 단어 패턴을 문법적 패턴으로 재해석한 후(②), 재해석된 문법적 패턴을 새로운 문장 생성에 적용하여 그 타당성을 확인한다(③).

그림 1에서 제안 방법의 전체적인 동작을 보이고 있다. 주어진 훈련용 데이터에서 무작위로 단어 조합을 샘플링한 후 샘플링된 단어들을 하이퍼에지(그림 1의 ②

및 그림 2의 E_1, E_2, E_3 에 해당)로 조합한다. 하이퍼에지는 무작위로 샘플링된 n 개의 단어(노드 X)와 이들 단어를 연결하는 에지(E)그리고 해당 하이퍼에지에 부여된 가중치(W)의 쌍 (X, E, W) 으로 구성된다. 이를 하이퍼에지의 쌍이 모여 하이퍼네트워크(H)를 구성한다. 초기 하이퍼에지 집합을 생성한 후, 학습 과정에서 주어진 훈련 데이터에서 개별 하이퍼에지(h_i)의 등장 빈도를 관찰하여 해당 하이퍼에지의 가중치를 조정하는 것으로 유의미한 패턴의 후보를 생성한다. 생성된 유의미한 패턴(등장 빈도가 높은 패턴)은 문법적 패턴으로 재해석되어 별도의 자료 구조에 보관된다. 별도로 저장된 패턴은 새로운 문장의 생성에 적용되어 그 타당성을 평가받는다.

Representing a hypernetwork

$$\begin{aligned} H &= (X, E, W) \\ X &= \{\text{this, have, can, we, been, show, data, reduced}\} \\ E &= \{E_1, E_2, E_3, E_4\} \\ W &= \{W_1, W_2, W_3, W_4\} \end{aligned}$$

$$\begin{array}{ll} E_1 = \{\text{this, can, have}\} & W_1 = 0.3 \\ E_2 = \{\text{we, have, been}\} & W_2 = 0.7 \\ E_3 = \{\text{been, reduced, data}\} & W_3 = 1.0 \\ E_4 = \{\text{can, show, data}\} & W_4 = 0.7 \end{array}$$

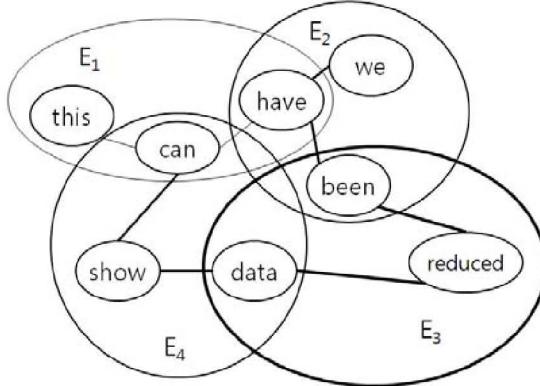


그림 2 하이퍼네트워크의 표현. 무작위로 선정된 단어를 모아서 하나의 하이퍼에지를 구성하며, 구성된 하이퍼에지에 학습을 통해 가중치를 부여한다.

2.1 학습 데이터

우리는 유아용 비디오 9종의 스크립트를 모아서 해당 스크립트에 포함된 문장을 훈련에 사용하였다. 훈련에 사용한 유아용 비디오는 Miffy, Looney tunes, Caillou, Dora Dora, Macdonald's farm, Thomas & Friends, Timothy, Winnie-the-Pooh, Kitty의 9종이며 각 스크립트 당 문장 규모는 최소 600단어에서 최대 8,600문장으로 총 문장 수는 29,000개에 달한다.

2.2 제안 알고리즘

주어진 훈련 데이터를 D ($D = (d_1, d_2, \dots, d_n)$), 이 논문

에서 d_i 는 개별 훈련용 문장을 의미), 찾고자 하는 문법 규칙의 집합을 GS 라고 나타낼 경우 학습을 통해 달성하고자 하는 목표는

$$GS' = \operatorname{argmax}_{GS} P(GS|D) \dots (1)$$

(1)을 만족하는 문법적 패턴 GS' 을 찾는 것이다. 여기서 $P(GS|D) \propto P(D|GS)P(GS)$ 이고 $P(D|GS) =$

$$\prod_{i=1}^n P(d_i|GS)$$

이다. 하이퍼에지의 집합(H)가 주어졌을 때 조건부 확률 아래의 식(2)

$$P(d_i|GS) \propto \operatorname{argmax}_s P(s|GS, d_i) \dots (2)$$

와 같이 표현할 수 있는데 여기서 s 는 주어진 문장 d_i 를 재생성할 수 있는 하이퍼에지(h_i)의 집합으로 각 하이퍼에지의 가중치를 이용하여, (2)를 표현할 수 있다.

제안 알고리즘은 주어진 훈련 문장을 설명할 수 있는 s 를 찾고 이를 바탕으로 GS 를 생성하며, 생성된 GS 를 바탕으로 다시 s 를 찾는 과정을 반복하는 EM (Expectation-Maximization) 프레임 워크[5]로 설명할 수 있다.



그림 3 GS 와 s 의 반복 추정

이 과정은 Expectation step ($P(s|GS, D)$)에 기반하여 s 생성과 Maximization step ($P(GS|s, D)$)에 기반하여 GS 생성으로 표현할 수 있다.

종료 조건을 만족할 때까지 이와 같은 과정을 반복한다고 하면 t 번째 학습이 진행될 때의 문법 규칙의 집합을 GS_t 그리고 가중치를 감안한 하이퍼에지 분포를 확률 변수 X_t 로 표현할 수 있다.

이 경우 $P(X_{t+1} = GS_j | X_0 = GS_0, \dots, X_t = GS_t)$ $= P(X_{t+1} = GS_j | X_t = GS_t)$ 라고 표현할 수 있는데 이는 우리가 MCMC 방법론과 샘플링 방법을 이용하여 GS 를 근사할 수 있음을 의미한다. 제안 방법을 통한 문법적 규칙의 학습은 학습된 문법적 패턴을 새로운 문장 생성에 적용한 후 생성된 문장의 문법적 타당성을 평가하는 방식으로 진행할 것이다.

제안 방법론은 [6]의 방법론과 (i)오더(에지를 구성하는 단어 수)의 제한이 없고 (ii) 주어진 훈련 데이터에 의존하며 (iii) 컨텍스트 의존적이라는 점에서 유사하다. 또한 제안 방법과 [6]의 방법 모두 통계적 정보에 기반

하여 미가공 데이터를 학습에 사용한다. 그러나 제안 방법은 무작위 샘플링된 단어 조합의 생성을 통해 문법 규칙을 보다 간명하게 표현할 수 있으며 무감독 학습 과정에서 주어진 학습 문장의 재생성을 학습 목표로 설정하여 학습을 진행한다는 차이가 있다.

3. 결론 및 추후 연구

본 논문에서 우리는 무감독 학습을 이용하여 자연언어 데이터를 분석하여 주어진 훈련 데이터로부터 문법적 패턴을 탐색하는 방법을 제안하였다. 본 논문에서 우리는 자연언어 습득 과정에 대한 단초를 제안하려고 한 것이 아니라, 기계 학습 관점에서 문법 규칙 습득 과정을 모사할 수 있는 방법론을 소개하고자 하였다.

본 논문에서 소개한 방법은 베이지안 추론과 MCMC 샘플링을 EM 프레임워크로 통합한 것으로, 미가공된 훈련 데이터가 주어진 상태에서 언어 학습을 설명할 수 있는 한 가지 가능성을 제공한 방법론이다. 그러나 언어 습득 현상은 순차적인 현상이며, 부분적인 단어 패턴의 결합은 본 논문에서 제안한 방법과는 비교도 할 수 없을 정도로 역동적으로 발생한다. 우리는 추후 연구를 통해 부분적인 패턴을 순차적으로 결합할 수 있는 방법론을 소개하여 제안 방법론을 한층 발전시키고자 한다.

감사의 글

본 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(No. 2010-0017734)이며 지식경제부 산업원천기술개발사업(10035348, 모바일 플랫폼 기반 계획 및 학습 인지 모델 프레임워크 기술 개발), BK21-IT 사업에 의해 일부 지원되었음. 본 연구를 위해 연구 장비를 지원하고 공간을 제공한 서울대학교 컴퓨터연구소에 감사드립니다.

참고문헌

- [1] J. G. Wolff, Learning syntax and meanings through optimization and distributional analysis, in: Y. Levy, I. M. Schlesinger, M. D. S. Braine (Eds.), *Categories and Processes in Language Acquisition*, Lawrence Erlbaum, Hillsdale, NJ pp. 179-215, 1988.
- [2] C. Bannard, E. Lieven, and M. Tomasello, Modeling children's early grammatical knowledge, *Proceedings of the National Academy of Sciences*, Vol. 106, No. 41, pp. 17284-17289, 2009.
- [3] J. R. Saffran, R. L. Aslin, and E. L. Newport, Statistical learning by 8-month-old infants, *Science*, Vol. 274, pp. 1926-1928, 1996.
- [4] 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.
- [5] T. M. Mitchell, *Machine Learning*, McGraw-Hill Companies, 1997.
- [6] Z. Solan, D. Horn, E. Ruppin, and S. Edelman, Unsupervised learning of natural languages, *Proceedings of the National Academy of Sciences*, Vol. 102, No. 33, pp. 11629-11634, 2005.