자동 스토리텔링을 위한 하이퍼그래프 언어모델 기반의 문장의 전후 관계 분석

장하영 ○ 장병탁

서울대학교 전기컴퓨터공학부

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

Context Analysis based on Hypergraph Language Model for Automatic Storytelling

Ha-Young Jang○ Byoung-Tak Zhang

School of Computer Science and Engineering, Seoul National University

요 약

자동 스토리텔링을 위한 대부분의 기법들은 미리 정의된 규칙이나, 문장간의 인과 관계가 미리 정의된 데이터를 사용하게 된다. 이와 달리 데이터 기반의 자동 스토리텔링을 위해서는 주어진 데이터에 존재하는 문장간의 전후 관계를 자동으로 분석하는 기법이 필요하다. 이를 위해서 본 논문에서는 하이퍼그래프 언어모델(Hypergraph Language Model)을 이용하여 문장간의 전후 관계를 자동으로 분석하는 새로운 기법을 제안한다. 인자 언어모델(Factored Language Model)에서 단어를 형태학적 인자들의 벡터로 간주하는 것처럼, 하이퍼그래프 언어모델에서는 문장을 단어들의 벡터로 간주하여 언어모델을 구축함으로써 문장간의 상관관계를 분석하는 것이 가능하고 이를 기반으로 하여 데이터 기반의 자동 이야기 생성이 가능하다. 제안된 방법론은 단순히 문장간의 선후관계만을 분석하는 것이 아니라 이야기의 동적인 변화를 분석함으로써 사용자와의 상호작용을 통해 동적으로 작용하는 인터랙티브 스토리텔링(Interactive Storytelling) 등의 분야에도 적용이 가능하다.

1. 서 론

스토리텔링은 컴퓨터와는 다른 인간의 창의성과 관련된 문제라고 인식되어 왔다. 그러나 최근 몇 년간 창의적인 스토리텔링을 위한 연구가 활발히 진행되고 있고, 이러한 연구들은 사용자에게 이야기를 전달하는 전통적인 의미의 스토리텔링에서부터 사용자와 매체의 상호작용 속에 이루어지는 인터랙티브 스토리텔링까지 다양한 분야에서 진행되고 있다. 스토리텔링을 위한 이야기를 만드는 방식은 크게 이야기의 등장인물들을 정의하는 규칙들을 기반으로 이야기를 만드는 방식과 이야기를 구성하는 대사들을 구성하기 위한 특수한 형태의 문법을 기반으로 이야기를 만드는 방식으로 구분할 수 있다[4, 7, 8, 10].

기존의 방법론에서는 이야기를 만들어 내기 위해서는 사람이 직접 입력한 규칙이나 문법 등이 필요한 반면에 본 논문에서는 하이퍼그래프 언어모델을 이용한 데이터 기반의 스토리모델 구축을 위한 새로운 방법론을 제안하였다. 제안한 방법론은 규칙기반의 방법론들과는 달리 태깅(tagging)되어 있지 않은 말뭉치를 이용하여 문장간의 전후 관계를 학습하여 모델을 만들 수 있는 방법을 제공한다. 이러한 특성으로 인해서 과거와는 달리 쉽게 대용량의 말뭉치를 획득할 수 있는 최근의 기술상황에서 기존의 방법론보다 다양한 활용분야를 가질 수 있을 것으로 기대된다.

제안된 데이터 기반의 문맥 분석 방법은 생성모델(Generative Model)로써의 하이퍼그래프 모델의 특성[2]을 활용하여 기존에 존재하는 이야기를 기반으로 새로운 이야기를 만들어 내는 것뿐만 아니라 이야기의 흐름 및 등장인물간의 관계 분석 등의 의미론적인 구조 분석에 있어서도 유용한 역할을 할 수 있을 것으로 기대된다. 하이퍼그래프 모델을 이용한 문맥 분석의 효율성을 확인하기 위해서는 이야기의 문맥이 잘 표현되어 있는 데이터를 사용하는 것이 중요하다. 이를 위해서 본 논문에서는 문장의 전후 관계가 잘 드러날 수 있도록 대화문으로 구성된 말뭉치를 이용하여 실험을 진행하였다.

2. 하이퍼그래프 언어모델(Hypergraph Language Model)

자연어 안에서 문법, 구문, 단어 등에 대한 규칙성을 찾아내고 이를 이용하기 위해 만들어진 언어모델은 오랫동안 음성 인식이나 기계 번역, 문자 인식, 철자 교정 등 다양한 응용분야에서 시스템의 정확도를 높이고 수행 시간을 줄이는 데 유용한 방법으로 각광을 받아왔다. 통계적 기법을 이용하는 언어모델에서 문장 s가 나타날 확률은 다음과 같이 계산될 수 있다[9]. 여기서 n=1인 경우 bigram, n=2인 경우 trigram이 된다.

$$
\begin{aligned}
P(s) &= P(w_1, w_2, ..., w_n) \\
&= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)...P(w_n \mid w_1, w_2, ..., w_{n-1}) \\
&= \prod_{i=1}^{n} P(w_i \mid w_1, ..., w_{i-1}) \\
&\approx \prod_{i=1}^{n} P(w_i \mid w_{i-n+1}, ..., w_{i-1})
\end{aligned} \tag{1}
$$

이때의 확률분포는 학습 말뭉치에 대한 최대 우도 추정(Maximum Likelihood Estimation)을 통해서 아래와 같이 계산될 수 있다.

$$
\begin{aligned}
P_{MLE}(w_1 \cdots w_n) &= \frac{C(w_1 \cdots w_n)}{N} \\
P_{MLE}(w_n \mid w_1 \cdots w_{n-1}) &= \frac{C(w_1 \cdots w_n)}{C(w_1 \cdots w_{n-1})}
\end{aligned} \tag{2}
$$

하이퍼그래프 모델에서는 말뭉치에 존재하는 단어들의 조합으로 구성된 하이퍼에지와 하이퍼에지의 출현빈도를 표현하는 가중치로 확률분포를 표현[1]하게 되는데 이때 하이퍼그래프 모델의 에너지는 다음과 같이 정의된다[6].

$$
\varepsilon(s^{(n)}; W) = -\sum_{i=1}^{|E|} w_{i_1 i_2 \ldots i_{|E_i|}} x_{i_1}^{(n)} x_{i_2}^{(n)} \ldots x_{i_{|E_i|}}^{(n)} \tag{3}
$$

$x_{i_1}^{(n)} x_{i_2}^{(n)} \ldots x_{i_{|E_i|}}^{(n)}$ 는 말뭉치 내에 존재하는 단어들의 조합으로 구성된 하이퍼에지를 의미하고 W는 하이퍼에지의 가중치를 의미한다. 즉, n-gram 방식의 언어모델에서 사용하는 단어들의 출현빈도를 가중치의 형태로 표현하여 이를 이용하여 하이퍼그래프 모델을 표현하는 것이다.

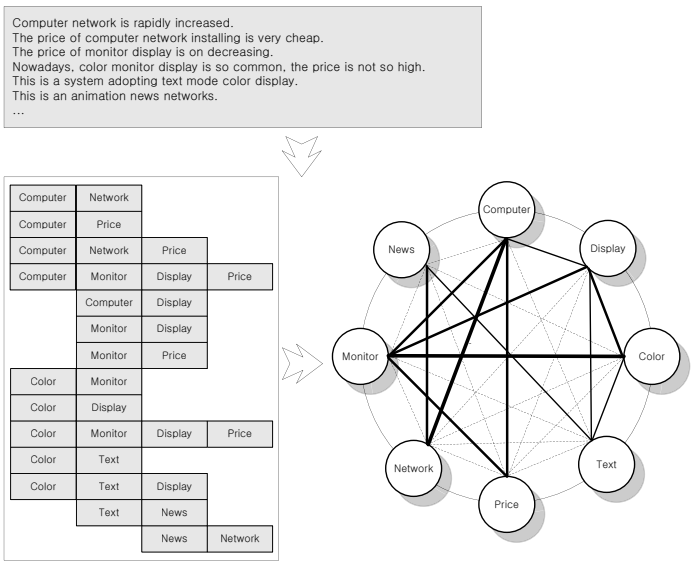

그림 1. 하이퍼그래프 모델을 이용한 언어모델 구축 예.

이렇게 만들어진 하이퍼그래프 모델에서 문장 $s^{(n)}$이 나타날 확률은 다음과 같이 깁스 분포에 의해서 주어지게 되고,

$$
P(s^{(n)} \mid W) = \frac{1}{Z(W)} exp\{-\varepsilon(s^{(n)}; W\} \tag{4}
$$

분할함수(partition function) $Z(W)$는 다음과 같이 정의된다.

$$
Z(W) = \sum_{x^{(m)}} exp\{-\varepsilon(s^{(m)}; W\} \tag{5}
$$

3. 하이퍼그래프 언어모델을 이용한 스토리모델

자동 스토리텔링을 위한 스토리모델의 구축을 위해서는 문장이 주어졌을 때, 그 문장과 가장 관련이 높은 문장을 선택할 방법이 필요하다. 즉, 주어진 문장 S_q에 대해서 결과 문장 S의 확률을 계산할 필요가 있고, 이를 위해서 $P(S|S_q)$의 계산이 필요하다. 이를 위하여 본 논문에서는 인자언어모델 (Factored language Model)의 기본 개념을 확장하여 사용하였다. 인자언어모델에서는 단어를 다음과 같이 k개의 인자를 가진 벡터로 간주하게 된다[5].

$$
w_i = \{f_i^1, f_i^2, \ldots, f_i^k\} \tag{5}
$$

이와 유사하게 제안된 모델에서는 문장을 k개의 단어를 가진 벡터로 간주하여, 이를 이용하여 문장간의 관계를 이용한 스토리모델을 구축하였다. 문서 D와 문장 S가 주어졌을 때 문서 D가 나타날 확률은 다음과 같이 계산된다.

$$
\begin{aligned}
D &\equiv \{s^1, s^2, \ldots, s^k\} = s^{1:k} \\
S &\equiv \{w^1, w^2, \ldots, w^k\} = w^{1:k} \\
P(D) &= P(s_1, s_2, \ldots s_T) = P(w_1^{1:k}, w_2^{1:k}, \ldots, w_T^{1:k}) = P(w_{1:T}^{1:k})
\end{aligned} \tag{6}
$$

이때, 식(6)은 n-gram 모델에서와 마찬가지로 확률의 연쇄법칙을 적용하고 마코프 가정 (Markov Assumption)을 이용하면 다음과 같이 정리될 수 있다.

$$
\begin{aligned}
P(w_{1:T}^{1:k}) &= P(w_1^{1:k}, w_2^{1:k}, ..., w_T^{1:k}) \\
&\approx \prod_{t}^{T} P(w_t^{1:k} \mid w_{t-(n-1)}^{1:k}, ..., w_{t-1}^{1:k}) \\
&= \prod_{t}^{T} \prod_{k}^{K} P(w_t^k \mid w_t^{1:k-1}, w_{t-(n-1)}^{1:k}, ..., w_{t-1}^{1:k})
\end{aligned} \tag{7}
$$

이때 식(7)의 계산을 위해서는 서로 다른 문장에 존재하는 단어들간의 확률을 알고 있어야 한다는 전제가 있고, 이로 인해서 기존의 n-gram 방식의 모델에서는 이를 계산하는 것이 불가능해진다. 그러나 그림 1에서 확인할 수 있듯이 하이퍼그래프 모델에서는 서로 다른 문장에 존재하는 단어들간의 확률도

하이퍼에지를 통해서 모델링하고 있기 때문에 하이퍼그래프 언어모델의 경우에는 식(7)의 계산이 가능하게 된다[3]. 서로 다른 문장에 존재하는 단어들간의 관계까지도 모델링 하는 하이퍼그래프 모델의 특성을 이용하여 만들어진 스토리모델은 기존의 언어모델과 달리 문서에 존재하는 문장간의 선후관계, 즉 문맥을 학습하는 것이 가능하게 되고, 이를 통해서 주어진 문장과 관련이 있는 새로운 문장을 찾아내는 것이 가능해진다.

4. 실험 결과

실험에 사용된 데이터 자체가 사람의 판단으로도 문맥 파악이 쉽지 않을 경우에는 실험결과가 제대로 문맥을 분석하고 있는지를 확인하는 것 자체가 어렵기 때문에, 본 논문에서는 문장의 전후 관계나 문맥이 잘 드러날 수 있는 대화문을 데이터로 사용하였고, 이를 위해서 시트콤 'Friends'의 대사로 구성된 말뭉치를 선택하여 실험을 진행하였다.

말뭉치는 전체 223개의 에피소드에 나오는 모든 대사로 구성이 되어 있고, 교착어인 한국어의 특성으로 인해서 발생하는 어휘의 증가와 전처리 과정에서 발생할 수 있는 오류를 감소시키기 위해서 영어 대사를 이용하여 실험을 진행하였고, n-gram에서의 n과 같은 역할을 하는 하이퍼에지의 차수는 2 와 3을 사용하였다. 또한 제안된 방법론에서는 단어가 아닌 문장간의 관계를 모델링 한다는 특징으로 인해서 통계적 언어모델에서 흔히 발생하는 데이터의 희소성 문제가 더욱 커지기 때문에, 확률 값의 계산 과정에서 smoothing을 위하여 generalized backoff 방법[5]을 적용하였다.

말뭉치를 이용하여 학습된 스토리모델이 문장의 전후 관계를 얼마나 잘 학습했나 확인하기 위하여 질의문장을 제시한 후에, 이를 이용하여 다음 문장을 말뭉치에서 찾아내는 실험을 수행하였다.

질의문장	선택문장
Can I get you some coffee? (Episode 0101)	Hi, sure! (Episode 0101, Episode 0421)
I have to say Tupolo Honey by Van Morrison. (Episode 0301)	Oh my god (Almost every episode – 223개)
umm we gotta get up early and catch that plane for New York. (Episode 0501)	It's a very large plane. (Episode 0424, Episode 0501)
I cannot believe you guys are talking about this! (Episode 0701)	You wanna talk about people's feelings? (Episode 0917)
I need to talk to you. (Episode 0901)	Sure! (Almost every episode – 224개)

표 1. 질의문장을 이용한 다음 문장 선택 결과

표1의 실험결과에서 알 수 있듯이 짧은 문장이나 흔한 표현들이 다음 결과로 많이 나오는 경향을 보이고 있으나, 3번째 질의문의 경우와 같이 특징적인 단어가 들어가 있는 경우에는 의미적, 시간적으로 좋은 결과를 보이는 것을 알 수 있다.

이 실험 결과를 바탕으로 학습된 스토리모델의 이야기 구성 능력을 알아보기 위해서 질의문장으로부터 시작해서 연속적으로 문장을 선택하는 실험을 진행하였다. 즉, 질의문장 S_q에 대하여 $P(S_t|S_q)$ 을 계산하여 문장 S_t를 선택하고, 이를 다시 질의문장으로 사용하여 $P(S_{t+1}|S_t)$를 계산하여 S_{t+1}을 선택하는 과정을 반복적으로 수행하여 이야기를 만들어 보았다.

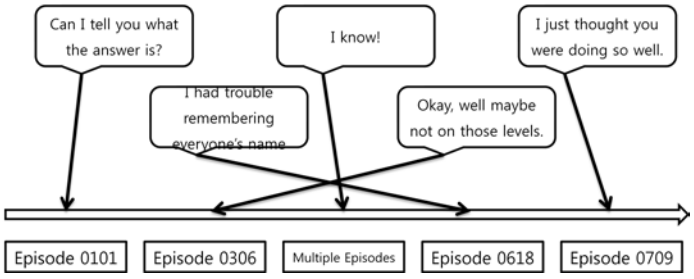

그림 2. 선택된 문장의 시간 순 배열

질의문장인 'Can I tell you what the answer is?'로부터 연속적으로 조건부 확률을 계산하여 선택된 문장들을 에피소드 별 시간 순으로 배열한 결과를 그림 2에서 보여주고 있다. 좌측에서 우측으로 배열된 문장의 순서는 문장이 선택된 순서를 보여주고 있고, 화살표 밑으로 해당 문장이 등장하는 에피소드가 표시되어 있다. 생성된 문장들은 시간 순으로 배치가 잘 되지 않은 것처럼 보이지만, "I Know"라는 자주 나오는 짧은 문장의 전후의 문장들은 에피소드 별 시간 관계에 의해서 선택되었음을 확인할 수 있다. 이러한 상황은 조건부 확률을 계산하는 과정에서 정확도가 떨어질 수 밖에 없는 짧은 문장의 등장으로 인해서 발생한 오류가 전체적인 문장의 배열에 크게 영향을 미치게 된 것으로 생각된다.

5. 결론

본 논문에서는 자동 스토리텔링을 위한 데이터 기반의 스토리모델 구축을 위한 방법론을 제안하였다. 사람이 작성한 규칙이나 태깅된 데이터 없이 말뭉치만으로 스토리텔링에 필요한 문장간의 관계 분석을 위해서 기존의 n-gram 방법과 달리 떨어져 있는 단어들 간의 관계도 모델링 할 수 있는 하이퍼그래프 언어모델을 이용하여 구축한 스토리모델을 이용하였다.

실험결과에서 확인할 수 있듯이 일부 실험에서는 짧고 자주 나오는 문장이 주로 선호되는 것을 알 수 있는데, 이러한 문제는 문장간의 관계를 모델링 하려는

시도의 특징상 데이터의 희소성이 너무 크기 때문에 발생한 것으로 판단이 된다. 그림 2에서 보여진 짧은 문장으로 인해서 조건부 확률의 계산 과정에서 정확도가 떨어지는 문제 또한 하이퍼에지의 차수를 증가시킴으로써 어느 정도 개선이 가능할 것으로 예상되나, 이러한 방법은 데이터의 희소성을 더욱 증가시키는 문제가 있기 때문에 이의 해결을 위해서는 대용량의 말뭉치 사용이나, 보다 정교한 smoothing 기법의 도입으로 문제점을 해결해야 할 것으로 생각된다.

추가적으로 현단계에서는 주어진 말뭉치에 기반해서 새로운 이야기를 만들어 내는 것이 아니라, 주어진 문장에 대해서 문맥상 전후 관계에 있는 문장들을 선택하는 것에 그쳤지만, 하이퍼그래프 언어모델의 생성 모델로써의 특징을 활용하면 이미 존재하는 이야기를 기반으로 새로운 이야기를 만들어 낼 수 있는 새로운 스토리 모델의 구현이 가능할 것이라 기대된다.

감사의 글

본 연구는 지식경제부 및 한국산업기술평가관리원의 IT산업원천기술개발사업의 일환으로 수행하였으며 (KI002138, 차세대 맞춤형 서비스를 위한 기계학습 기반 멀티모달 복합 정보 추출 및 추천기술 개발, MARS), 교육과학기술부의 재원으로 국가연구재단의 지원을 받아 수행된 연구(314-2008-1-D00377, Xtran), 지식경제부 산업원천기술개발사업(10035348, 모바일 플랫폼 기반 계획 및 학습 인지 모델 프레임워크 기술 개발) 및 교육과학기술부의 BK21-IT사업에 의해 일부 지원되었음.

참고문헌

[1] 고영길, 장하영, 김선, 장병탁, 기계번역문장 품질 평가를 위한 하이퍼네트워크 기반 언어 모델링, *2008 정보통신분야학회 합동학술대회 논문집*, pp. 277-280, 2008.11.

[2] 남진석, 장병탁, 랜덤 하이퍼그래프 메모리 모델에서 순차적 단서를 활용한 문장 생성, *2010 한국컴퓨터종합학술대회(KCC2010)논문집*, 제37권 1(A), pp. 148-149, 2010.06.

[3] 석호식, 작가멧, 장병탁, 단어 간 관계 패턴 학습을 통한 하이퍼네트워크 기반 자연 언어 문장 생성, *정보과학회논문지: 소프트웨어 및 응용*, 제37권 제3호 pp. 120-128, 2010.11.

[4] Aylett, R., Narrative in virtual environments-towards emergent narrative, *Proceedings of the AAAI fall symposium on narrative intelligence*, pp. 83-86, 1999.

[5] Bilmes, J.A. and Kirchhoff, K., Factored language models and generalized parallel backoff. *In HLT-NAACL 2003: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 4–6, Edmonton, Canada. Association for Computational Linguistics. 2003.

[6] Zhang, B.-T., Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory, *IEEE Computational Intelligence Magazine*, 3(3) pp. 49-63, 2008.

[7] Cavazza, M., Charles, F. and Mead, S., Agents' interaction in virtual storytelling, *Intelligent Virtual Agents*, pp156-170, 2001.

[8] Riedl, M., and Young, R., Story Planning as Exploratory Creativity: Techniques for Expanding the Narrative Search Space. *New Generation Computing, Computational Paradigms and Computational Intelligence. Special Issue: Computational Creativity,* 24(3) pp. 303– 323. 2006.

[9] Song, F. and Croft, W.B., A general language model for information retrieval, *Proceedings of the eighth international conference on Information and knowledge management*, pp. 316-321, 1999.

[10] Theune, M., Faas, E., Nijholt, A., and Heylen, D., The Virtual Storyteller: Story Creation by Intelligent Agents. *In Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference*, pp. 204–215. Berlin: Springer. 2003.