인과 추론 문제 해결에 적합한 자연어 처리 모델 BART 분석

조건희, 성창민, 김은총, 정윤경
성균관대학교
skate4333@g.skku.edu, tjdckdals@skku.edu, prokkec@naver.com, aimecca@skku.edu

Analyzing Natural Language Processing Model BART Suitable for Solving Causal Inference Problems

Gunhee Cho, Changmin Seong, Eunchong Kim, Yungyung Cheong
Sungkyunkwan University

요 약

자연어처리 연구에서 인과 추론 문제는 언어 모델이 문맥과 인과 관계의 개념을 이해해야 해결할 수 있어 매우 도전적인 과제이다. 인과 추론 문제를 해결 시 뉴스 생성, 스토리 생성 등 다양한 분야에 적용할 수 있다는 점에서 관심이 높은 과제이다. 따라서, 본 논문에서는 최근 다양한 자연어 처리 과제에서 좋은 성능을 내고 있는 BART 모델을 인과 추론 문제 해결에 적용하고 그 성능을 분석한 결과를 보고한다.

1. 서 론

자연어 처리 연구는 인간이 사용하는 언어에 관한 문제들을 컴퓨터, 즉 인공지능으로 해결하는 분야이다. 언어를 컴퓨터에게 이해시키기 위해서는 복잡한 구조의 모델이 필요하기 때문에 연구자들은 다양한 모델들을 개발하여 자연어 처리 문제를 해결해 왔다.

자연어 처리 문제들 중 인과 추론은 관찰된 행동으로 인해 발생할 수 있는 미래 시나리오를 예측하는 것을 목표로 한다. 예를 들어, "민수는 운동을 한다."라는 행동이 관찰되었다면 "민수는 피곤하다."라는 미래 상태나 "민수는 샤워를 한다."라는 미래 행동을 예측해야 하는 것이 인과 추론이다. 인과 추론 문제는 뉴스 생성, 스토리 생성, 대화 생성 등 다양한 자연어 생성 과제에서 발생한다. 따라서 인과 추론 문제를 해결한다면 다양한 분야의 과제들의 실마리를 풀 수 있는 기대 효과를 노릴 수 있다. [1)]

하지만 인과 추론 문제는 언어 모델이 문맥과 인과 관계의 개념을 이해해야 해결할 수 있어 매우 도전적인 과제이다. 최근 연구에서는 인과 추론 문제를 해결하기 위해서 다양한 구조의 언어 모델을 적용해보고 있다.[1]

현존하는 언어 모델들은 서로 구조가 다르기 때문에 어떤 과제에 적용하느냐에 따라서 성능 역시 다르다. 본 논문에서는 문장 생성에 뛰어난 모델인 BART[2]를 기본 모델 구조로 사용하여 인과 추론 문제를 해결한다. 그리고 BART 모델의 초매개변수(hyperparameter) 설정을 변경하면서 비교하여 가장 성능이 좋은 BART를 찾아내는 것을 목표로 한다.

본 논문은 Wikipedia 데이터셋으로 사전 학습(pre-train)한 BART 모델을 사용한다. 전 연구에서 사전 학습만 되어 있는 언어 모델들보다 특정 문제에 관련된 미세 조정 학습(fine-tuning)을 진행시킨 모델들이 특정 문제 해결에서 성능이 더 좋다고 밝혀진 바[4]가 있어, 본 논문에서도 모델에게 인과 추론 문제에 관련된 조정학습을 진행하였다. 이 과정을 위해 필요한 데이터셋은 인과 관계 QA 데이터셋을 전처리하여 구축하였다. 이후 학습된 모델의 성능을 여러 평가 지표로 평가해 보고 어떤 BART 모델이 인과 추론에 가장 적합한 모델인지 확인하였다. 그리고 초매개변수에 따라서 달라지는 성능도 평가하여 제일 성능이 좋은 초매개변수 설정을 확인하였다.

1) 이 논문은 2019년도 정부(교육과학기술부)의 재원으로 한국연구재단(No. 2019R1A2C1006316), 2019년도 정부(과학기술정보통신부) 의 재원으로 정보통신기획평가원(No.2019-0-00421, 인공지능대학원지원), 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업(IITP-2017-0-01642)의 지원을 받아 수행된 연구임.

2. 관련 연구

자연어를 처리하기 위한 딥러닝 기반 모델들 중 가장 혁신적이고 성능을 크게 향상시킨 모델이 Transformer[3] 모델이다. Transformer 모델은 기존의 LSTM 대신 attention 구조만을 사용한 언어 모델이다. Transformer 모델을 기반으로 많은 모델들이 개발되었고 그 중에서 사전 학습(pre-training)과 미세 조정 학습(fine-tuning)으로 성능을 상승시킨 GPT-1[4], 새로운 언어 모델 학습법인 Masked Language Modeling(MLM)으로 사전 학습을 진행시켜 성능을 높인 BERT[5] 등이 성능이 좋은 모델로 알려져 있다.

BART 모델은 기존의 BERT 모델에서 사용하였던 Masked Language Modeling(MLM) 기법을 심화하였다. BERT의 MLM이 문장의 단어를 MASK 토큰으로 바꾸어서 언어 모델에게 그 단어를 맞추게 하여 훈련을 진행했다면 BART는 이 훈련뿐만 아니라 문장의 단어 순서를 뒤바꾸고 언어 모델에게 원래 문장 순서를 맞추도록 하는 훈련, 단어를 아예 없앤 뒤에 언어 모델에게 그 단어와 위치까지 맞추도록 하는 훈련 등 MLM보다 더 심화한 방식들을 채용하여 더 성능 좋은 언어 모델을 만들었다. BART는 이러한 훈련 덕에 언어의 특성을 잘 파악하고 특히 요약이나 대화 생성 문제에서 좋은 성능을 보여주고 있다.

최근에는 인과 추론 문제를 해결하기 위한 모델들이 계속해서 연구되고 있다. BERT 계열의 모델들을 전이 학습(transfer learning)시켜서 인과 추론 문제를 해결하는 CausalBERT[1], 모델이 단어가 아닌 문장에 집중하게 하여 원인 문장과 결과 문장 사이의 관계를 그래프로 이해하게 만든 EGCER[6] 등이 그러

한 연구들이다. 인과 추론 문제 해결은 다양한 자연어 생성 과제 해결에 도움을 주기 때문에 앞으로도 지속적으로 연구되어야 한다.

3. 실험

3.1. 데이터셋

인과 추론 데이터셋으로 총 두 종류를 사용하였다. 인과 추론 데이터셋으로는 COPA[7], GLUCOSE[8] 두 종류의 데이터셋을 사용하였다. 두 데이터셋은 인과 추론 문제 연구에서 많이 사용되고 있는 데이터셋이다.

3.1.1. COPA 데이터셋

COPA 데이터셋은 상식 인과 추론에 대해서 질문과 대답으로 이루어진 데이터셋이고 총 1000개의 질문으로 구성되어 있다. 모델에서 사용할 수 있도록 질문과 대답으로 이루어진 것을 페어로 만들어서 하나의 데이터로 만들었다.

```
<item id="1" asks-for="cause" most-plausible-alternative="1">
  <p>My body cast a shadow over the grass.</p>
  <a1>The sun was rising.</a1>
  <a2>The grass was cut.</a2>
</item>
```

그림 1 COPA 데이터셋

[그림 1]을 참고하면, 와 사이에 들어있는 문장이 질문 문장이다. asks-for에 할당된 값이 “cause”이므로 질문 문장의 원인이 되는 문장을 a1이나 a2 중 골라야 하는데 most-pluasible-alternative의 값이 1이므로 a1이 원인이 되는 문장이 된다. 따라서 “The sun was rising.”이 원인, “My body cast a shadow over the grass.”가 결과인 두 문장으로 구성된 페어를 만들 수 있고 이것이 하나의 훈련 데이터가 된다.

3.1.2. GLUCOSE 데이터셋

GLUCOSE 데이터셋은 대규모 상식 인과 추론 데이터셋이다. 이 데이터셋에는 원인, 결과를 페어로 하는 데이터가 65522개 존재한다. [그림 2]는 원인과 결과로 구성된 GLUCOSE 데이터셋의 예시이다.

```
Sara goes to a football game > Causes/Enables > Her team wins
```

그림 2 GLUCOSE 데이터셋

3.1.3. 데이터 전처리

데이터셋은 대부분 완전하고 해석 가능한 문장으로 구성되어 있지만 언어 모델이 이해하기 쉽도록 대문자는 소문자로 변환해주고 영어가 아닌 문자들은 삭제하였다.

3.1.4. 데이터 스플릿

COPA 데이터셋과 GLUCOSE 데이터셋을 랜덤으로 섞은 다음 그 중 80퍼센트(45110개)는 train 데이터셋으로 사용하고 10퍼센트(5639개)는 validation 데이터셋, 나머지 10퍼센트(5639개)는 test 데이터셋으로 사용하였다.

3.2. 모델

기본적인 모델 구조는 BART 논문의 문장 생성 모델을 따른다. BART는 인코더와 디코더를 연결한 구조의 모델로 인코더와 디코더는 Transformer 논문의 인코더, 디코더 구조를 따른다. 인코더는 입력 문장을 해석하여 디코더에게 해석한 정보를 전달해주고 디코더는 해석한 정보를 토대로 정답 문장을 생성해낸다. 인과 추론 문제에서는 입력 문장이 원인이 되고 정답 문장이 결과가 된다.

3.3. 학습 과정

학습과정은 두 단계로 이루어진다. 첫 단계는 사전 학습 단계로 대량의 말뭉치로부터 학습을 진행하여 언어의 특성을 모델에게 가르친다. 그 후 두 번째 단계는 미세 조정 학습 단계로 특정 문제와 관련된 데이터셋으로 모델을 학습시켜 그 문제 해결에 적합한 모델이 되도록 만든다.

이 실험에서는 2)Huggingface 라이브러리부터 사전 학습된 BART 생성 모델을 불러오고 위 과정에서 만든 train 데이터셋으로 BART의 미세 조정 학습을 진행한다. 그 후 test 데이터셋으로 모델의 성능을 평가한다.

3.4. 학습률 변화

학습률은 모델이 출력한 결과 문장과 정답 문장 간의 차이(loss)가 발생했을 때 그 차이의 얼마만큼을 모델 매개변수에 영향을 미칠것인지를 결정한다. 학습률에 따라서 모델의 성능이 바뀌기 때문에 어떤 학습률(1.3-5e, 5.3-5e, 9.3-5e)을 가질 때 BART의 학습이 가장 잘 되는지를 확인하는 실험을 진행했다.

3.5. 일반화 계수 변화

BART의 학습을 진행할 때 모델 매개변수가 크게 변하지 않도록 일반화 계수(weight decay)를 설정해주는데 그 이유는 모델이 훈련 뒤 train 데이터는 정답을 잘 추출하지만 train 데이터 이외의 데이터는 정답을 잘 추출하지 못하는 현상인 과적합(overfitting)을 방지하기 위해서다. 일반화 계수의 수치에 따라서 모델의 성능의 차이가 존재하는데 어떤 일반화 계수(0.1, 0.5, 0.9)를 가졌을 때 BART의 성능이 좋은지를 확인했다.

3.6. 모델 설정

모델을 학습하는 과정에서 학습 횟수(epoch)은 5회, 배치(batch) 크기는 16, 손실함수로는 MSE(Mean Squared Error)를 사용했다.

3.7. 비교군 모델

비교군 모델로는 학습률은 9.3-5e, 일빈화 과정은 진행하지 않는 BART 모델을 사용하였다.

3.8. 평가지표

평가 지표로는 BLEU[9]와 ROUGE[10]를 사용한다. 두 지표는 정답 문장과 모델이 생성한 문장의 유사도를 계산하여 점수로 나타내준다. 두 평가 지표의 값이 크면 클수록 모델의 성능이 좋음을 의미한다.

4. 결과

4.1. 문장 생성

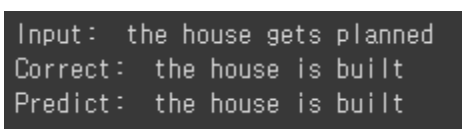

그림 3 문장 생성 예시(같음)

[그림 3]은 입력 문장으로 “the house gets planned”가 입력

2) https://huggingface.co

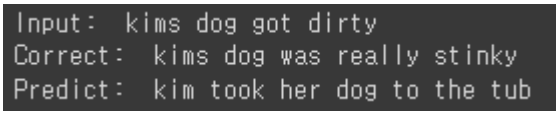

그림 4 문장 생성 예시(다름)

되었을 때 정답 문장은 “the house is built”이고 모델이 예측한 문장은 “the house is built”이다. 모델이 생성한 문장과 정답 문장이 정확히 일치한다.

또한, [그림 4]와 같이 모델이 생성한 문장과 정답 문장이 다르긴 하지만 문맥상 “kims dog got dirty”후에 “kim took her dog to the tub”은 원인, 결과 문장으로 봐도 무방하기 때문에 모델은 문맥을 이해하여 정답 문장과 다른 새로운 결과 문장도 생성할 수 있음을 확인할 수 있었다.

4.2. BART 성능 분석

[표 1]을 볼 때, 9.3-5e 학습률과 0.9 일반화 계수로 설정하여 훈련했을 때 BART의 성능이 가장 좋다는 것을 알 수 있다. 보통 과적합 현상은 일반화 계수가 클 때 잘 방지할 수 있다. 그리고 실험 결과에서 일반화 계수가 제일 클 때 성능이 좋다는 것을 알 수 있는데 이는 과적합 현상이 모델의 성능을 얼마나 해치는 지를 알려주는 결과이다. 학습률에 관련해서 모델마다 좋은 학습률이 존재하는데 일반화 계수가 0.9인 BART가 인과 추론을 할 때는 학습률 9.3-5e가 가장 적합하다는 것을 알 수 있다. 결국 이 결과를 통해서 초매개변수의 설정이 지표의 0.01 점수 차이를 만들어내는 만큼 모델의 성능에 영향을 준다는 것을 알 수 있다.

표 1 학습률과 일반화 계수에 따른 BART 성능

지표 / BART	BLEU	Rouge-1	Rouge-2	Rouge-L
lr:1.3-5e, wd:0.1	**0.44**	0.31	0.12	0.31
lr:5.3-5e, wd:0.1	0.42	0.32	0.13	0.31
lr:9.3-5e, wd:0.1	0.42	0.32	0.13	0.31
lr:1.3-5e, wd:0.5	0.43	**0.33**	0.14	0.32
lr:5.3-5e, wd:0.5	0.42	0.32	0.13	0.31
lr:9.3-5e, wd:0.5	0.42	0.32	0.13	0.31
lr:1.3-5e, wd:0.9	0.43	**0.33**	0.14	0.32
lr:5.3-5e, wd:0.9	0.41	**0.33**	0.13	0.32
lr:9.3-5e, wd:0.9	0.42	**0.33**	**0.15**	**0.33**
비교군 모델	0.43	0.32	0.13	0.31

lr=learning rate(학습률), wd=weight decay(일반화 계수)

5. 결론

본 논문은 자연어 처리에서 도전적인 과제로 여겨지는 인과 추론 문제를 BART 모델로 해결해보고 학습률과 일반화 계수에 따라 달라지는 BART 모델의 성능 변화를 확인하고 비교하였다. 그 결과로 본 논문에서 설정한 환경에서 인과 추론 문제를 가장 효과적으로 해결하는 BART 모델은 학습률을 9.3-5e, 일반화 계수를 0.9로 가지는 BART이었고 초매개변수 설정이 모델의 성능에 영향을 줄 수 있다는 것을 입증했다. 향후에는 학습률과 일반화 계수의 더욱 세밀한 교정을 통해서 더욱 성능이 향상된 모델을 개발하기를 기대한다.

참고문헌

[1] Li, Zhongyang, et al. "Causalbert: Injecting causal knowledge into pre-trained models with minimal supervision." arXiv preprint arXiv:2107.09852 (2021).

[2] Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461 (2019).

[3] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[4] Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

[5] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[6] Mu, Feiteng, Wenjie Li, and Zhipeng Xie. "Effect Generation Based on Causal Reasoning." Findings of the Association for Computational Linguistics: EMNLP 2021. 2021

[7] Roemmele, M., Bejan, C., and Gordon, A. (2011) Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford University, March 21-23, 2011

[8] Mostafazadeh, Nasrin, et al. "Glucose: Generalized and contextualized story explanations." arXiv preprint arXiv:2009.07758 (2020).

[9] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

[10] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004.