요약문 기반 문학 스타일 문장 생성

최부광○, 이은찬, 안상태*
경북대학교 전자공학부, 경북대학교 전자전기공학부
*stahn@knu.ac.kr

Generating Literature-Style Sentences based on Summarized Text

Bugwang Choe○, Eunchan Lee, Sangtae Ahn*
School of Electronics Engineering, Kyungpook National University
School of Electronic and Electrical Engineering, Kyungpook National University

요 약

최근 자연어 생성 연구는 딥러닝 기반의 사전 학습 모델을 중심으로 활발하게 연구되고 있다. 하위 분야 중 하나인 텍스트 확장은 입력 텍스트를 출력에 잘 반영하는 것이 무엇보다도 중요하다. 기존 한국어 기반 텍스트 확장 연구의 경우 몇 개의 개념 집합에 기반해 문장을 생성하도록 한다. 그러나 이는 사람의 실제 발화 길이에 비해 짧고 단순한 문장만을 생성한다는 문제점이 존재한다. 본 논문은 이러한 문제점을 개선하면서 문학 스타일의 문장들을 생성하는 모델을 제안하였다. 또한 동일 모델에 대해 학습 데이터의 양에 따른 성능도 비교하였다. 그 결과, 짧은 요약문을 통해 문학 스타일의 여러 문장들을 생성하는 것을 확인하였고, 학습 데이터를 추가한 모델이 성능이 더 높게 나타나는 것을 확인하였다.

주제어: 자연어 생성, 한국어 사전학습모델, 문학 스타일 문장 생성

1. 서론

자연어 생성(Natural language generation)은 특정한 의사 전달 목표의 달성을 위해 자연어 텍스트를 생성하는 자연어 처리 분야로, 최근 딥러닝을 중심으로 활발하게 연구되고 있다. 자연어 생성의 하위 분야로는 텍스트 요약(Abbreviation), 텍스트 확장(Expansion), 텍스트 재작성 및 추론(Rewriting and reasoning) 등으로 분류할 수 있다[1].

그 중 텍스트 확장은 여러 의미있는 단어들에 접속사나 전치사와 같은 요소들을 추가해 완전한 문장 또는 텍스트를 생성하는 작업이다[2,3,4]. 한국어에 기반한 기존 텍스트 확장 연구로는 [5]가 존재하는데, 3-6개의 개념 집합을 입력 데이터로 구성하고 평균 길이 25 이하의 문장을 타깃 데이터로 가지는 상식 추론 데이터셋을 제안하고 있다.

본 연구에서는 한국어 기반 시퀀스-투-시퀀스 사전 학습 모델을 문학 요약 데이터셋으로 미세 조정한 장문의 문학 스타일 텍스트 생성 모델을 제안한다.

2. 관련 연구

2-1. 딥러닝 기반의 자연어 생성

딥러닝 기반의 자연어 생성은 순환 신경망(Recurrent neural network, RNN) 혹은 트랜스포머(Transformer)와 같은 신경망 모델을 조건부 확률에 기반한 언어 모델링 같은 방식을 통해 완전한 텍스트를 생성하도록 학습시킨다[6].이때 언어 모델링(Language modeling, LM)은 자연어 생성의 핵심 중 하나로, 자연어를 입력하였을 때 확률에 기반하여 다음에 위치할 토큰을 추론하고 종료 토큰이 나올 때까지 동일 과정을 반복하여 자연어 시퀀스를 생성한다. 이러한 방식으로 신경망 모델에 자연어 말뭉치 데이터를 충분히 학습시키면 모델이 작문, 번역, 요약 등의 태스크를 수행할 수 있게 된다.

시퀀스-투-시퀀스는 신경망 언어 모델을 인코더와 디코더가 연결되도록 구성하여 입력 시퀀스를 조금 더 용이하게 학습하도록 설계된 모델이다. 해당 모델은 특히 기계 번역, 자연어 요약, 입력 텍스트 스타일 변환 등의 자연어 시퀀스를 목적에 맞는 다른 시퀀스로 변환시켜주는 연구에서 활발히 사용된다[7].

2-2. 한국어 사전 학습 언어 모델

한국어 기반의 사전 학습 언어 모델(Pretrained language model, PLM)은 대규모 한국어 말뭉치 데이터를 언어 모델링 방식으로 사전에 학습시킨 모델로, 전이 학습(Transfer learning)을 통해 다운스트림 태스크에서 더 높은 성능을 달성할 수 있도록 한다. 영어 기반의 PLM의 구조를 유지하되 학습 과정에서 한국어만을 학습시키는 대표적인 모델로는 KoGPT[8], KoGPT-2[9], KoBART[10]가 있다.

본 논문에서는 모델이 입력 데이터를 이해하도록 하기 위해 시퀀스-투-시퀀스 방식의 KoBART를 사용하였다.

표 1. 문학 자료 요약 데이터 예시

요약문	샘웅덩이 옆에 앉아 있던 막내 왕녀는 심심해지면 노오란 황금공을 치던지고 받는 장난을 하였는데 하루는 치던진 공을 놓쳐서 샘물 웅덩이에 퐁당 빠져 들어가 버렸다.
원문	(생략) 그 나무숲속 한가운데 커다란 노목나무 밑에 조그만 샘물이 흘러서 깊은 웅덩이가 되어 있었습니다. 그래 그 어여쁜 막내 왕녀는 언제든지 나무숲 속으로 가서 그 샘웅덩이 옆에 서늘하게 앉아 있었습니다. 그렇게 앉았다가 심심해지면 노오란 황금공(黃金球)를 하늘로 치던지고, 밑에서 두 손으로 받는 장난를 하였습니다. 하루는 왕녀님이 치던진 공를 받다가 잚소하여 놓쳐서, 풀 위에 뚝 떨어져 서는 떼굴떼굴 굴러서 그 샘물 웅덩이로 가 퐁당 빠져 들어가 버렸습니다.₩n왕녀는 놀라서 곧 뒤쫓아 가 보았으나 (생략)

KoBART는 BART와 동일한 구조의 모델을 한국어 위키백과, 청와대 국민청원 등이 포함된 40GB 이상의 대규 모 한국어 데이터셋을 사전 학습시킨 모델이다. 인코더와 디코더 각각 6개의 레이어와 16개의 헤드로 구성되고 총 124M의 모델 파라미터를 가진다.

3. 실험 환경

3-1. 데이터셋

본 논문에서는 소설 텍스트 생성을 위해 AI-HUB에서 제공하는 요약문 및 레포트 생성 데이터[11]에서 창작물 요약 데이터를 활용하였다. 창작물 요약 데이터는 문서 종류에 따라 간행물, 연설문, 문학, 나레이션으로 구분되는데 그 중에서 문학 요약 자료 총 10,800건 중 9,600건을 학습 데이터로 사용하였고 1,200건은 1:1 비율로 나눠 각각 검증 데이터와 평가 데이터로 사용하였다. 문학 이외 간행물, 연설문, 나레이션 데이터는 $KoBART_{add}$의 학습에만 사용하였다. 데이터는 표 1과 같이 요약문과 원문이 각각 입출력으로 구성된다.

3-2. 사전 학습 모델

소설 생성 모델은 입력된 문장에서 문맥을 이해하여 이를 출력 문장에 잘 반영해야 한다. 이를 위해 시퀀스-투-시퀀스 방식의 한국어 기반 사전 학습 모델인 KoBART를 사용하여 실험을 진행하였다.

3-3. 실험 세팅 및 설계

본 실험에서는 $KoBART_{base}$ 모델과 요약 데이터를 활용해 두 가지 실험을 진행하였다. 두 경우 모두 요약문을 입력 데이터로, 원문을 타겟 데이터로 사용하였다. 자연어 처리 주요 태스크 중 하나인 요약의 경우와 반대인 셈인데, 이는 문학 요약 자료의 주제라고도 볼 수 있는 요약문의 내용을 장문의 텍스트 전반에 걸쳐 풀어 나가도록 하기 위함이다.

첫 번째 실험의 경우 문학 요약 자료만 사용하였으며 두 번째의 경우 간행물, 연설문, 나레이션 요약 자료로 추가적인 학습을 진행하였다. 이는 문학 이외의 요약 데이터로 모델을 먼저 학습시키는 것이 성능 향상에 영향을 미치는지 확인하기 위한 것으로, $KoBART_{add}$로 표기한다.

학습에 사용된 하이퍼 파라미터는 다음과 같다. 문학 요약 자료에 대해 에폭은 3으로 설정하고 배치 크기는 4로 적용하였다. 옵티마이저는 AdamW, 학습률(Learning rate)은 2e-5를 사용하였다. 또한 입력 문장과 출력 문장의 최대 길이는 각각 64, 512로 설정하였다. 출력 $KoBART_{add}$ 대해 추가적으로 수행된 학습에 대해서는 배치 크기를 10으로 하였고 나머지 조건은 동일하게 두었다.

4. 결과

평가 지표 학습된 모델의 성능은 요약과 기계번역에 많이 사용되는 n-gram 기반의 평가 지표인 Recall-oriented understudy for gisting evaluation (ROGUE)와 Bilingual evaluation understudy(BLEU)와 평가 데이터로 측정하였다.

결과 분석 실험 결과는 표 2와 표 3에서 확인할 수 있다. 표 2는 평가 데이터에 대한 모델의 생성 예시를 보여준다. 생성 결과 입력된 요약문을 재구성하여 문학 스타일로 생성하는 것을 확인할 수 있다. 그러나 두 모델 모두 앞 부분에서는 요약문을 반영하고 있지만, 텍스트의 길이가 길어질수록 관련성이 떨어지는 문장을 생성하는 경향이 나타났다. 이 부분은 모델이 요약문의 의미를 텍스트 전반에 반영하여 생성할 수 있도록 개선하는 것이 필요해 보인다. 표 3은 평가 데이터에 대한 ROUGE 점수와 BLEU 점수를 나타낸다. 문학 요약 자료로만 학습된 $KoBART_{base}$와 비교해 간행물, 연설문, 나레이션의 요약 자료를 추가로 학습한 $KoBART_{add}$가 평균적으로 0.00585 만큼 높은 점수를 기록하였다. 이는 문학 이외의 요약 자료도 생성 모델의 성능 향상에 기여할 수 있음을 의미하며, 추후 연구에서 이러한 결과를 통해 실험에서 사용되지 않은 다양한 요약 데이터를 활용하면 모델의 성능 개선에 더욱 도움이 될 것이라 기대된다.

표 2. 테스트 데이터에 대한 모델의 텍스트 생성 예시

입력 요약문	광필이는 수년 있으면서도 손님방에 출입하거나 크게 말하지도 않고 어머니 모르게 걸인에게 돈이나 먹을 것을 주는 옥정이가 김 소사에게서 나왔다는 것이 의문이었다.
생성 결과	김 소사에게서 나왔단 말인가? — 광필이는 궁금하던 것이 그 후로 그의 집에 수년 있으면서도 자기 어머니 모르게 돈이나 한 푼 주지 않고, 다만 걸인에게 돈을 주는 것뿐이니 옥정이가 나와도 안되었다는 것을 생각할 때에는 어쩐지 '옥정이가'였으니까 그는 의문이었다. 그리고 또 한편으로는 이러한 일이 언제 생긴 것인지는 모르는 수수께끼였다. 자기가 이 일을 어떻게 처리할까 하고 생각해보는 동안에 무슨 일인지 생각했던 것이다. 그러나 그래도 몇 걸음 물러 나오며 제 말을 못 들은 체하며 아무 말도 못하고 그저 서성대는 것도 그만두었다. 그럴 때에, 어떤 경망한 생각을 가지고 와야 할 것인가. 그래 그런 생각이 자꾸만 생기면서 문득 이런 생각도 들었다. '내가 누구냐'고 묻자 오라버니가 먼저 대답하여 버리는 것이었으니 나는 가엾고 섭섭하였 다. '이 사람아, 어디서 온 사람이냐? 그렇다 하더라도 내가 너에게 하는 말에 동의하지 않는다. 이렇게 말하면 응하는 거나 마찬가지이니까.....' 하며 내숭을 피울 뿐이다. 사실 정초부터 사절이라고 해서 친정으로 가는 것은 아니었는데 만일 여기서 만나더라도 "얘! 네가 어디 있다는 소리인가는" 하기에 조금 조바심이 생겨서 집으로 돌아간다고 대답을 해 버린 것이라고 생각하면 과연 우연치 못할 노릇인데 저 편에서는 그것이 그렇게까지 큰 실수라고는 하겠으나 너무도 대담하게 나온 까닭이다. 김소사의 말이 들리면 그가 아는 여자요, 알든 모르든, 어쨌든 금전이라는 건 남의 재산이고 보면 결코 돈이 아니다. 그런데 왜 딴 여자를 하나 같이 찾아갔을까.

표3. 요약문 기반의 텍스트 생성 성능 비교
(Rogue-1: unigram, Rogue-2: bigram, Rogue-L: longest common subsequence)

	Rouge-1	Rouge-2	Rouge-L	BLEU
$KoBART_{base}$	0.383	0.082	0.213	0.0475
$KoBART_{add}$	0.391	0.088	0.218	0.0519

5. 결론

본 논문은 기존 요약 태스크에 사용되는 데이터셋을 활용한 문학 스타일 텍스트 생성 모델을 제안하였다. 제안 모델은 인코더에서 입력된 요약문을 이해하였고, 디코더에서 이를 반영하여 장문의 문학 스타일의 텍스트를 생성하였다. 다른 분야의 요약문을 추가적으로 활용한 실험을 통해 모델 성능이 향상됨을 보였다. 이는 데이터 부족 문제를 다른 분야의 데이터를 통해 개선시킬 수 있음을 의미한다. 향후 연구로는 생성 모델이 텍스트의 앞부분 뿐만 아니라 뒷부분까지도 요약문의 의미를 잘 담아내고 더 자연스러운 문장을 생성하도록 추가적으로 데이터 수집과 전처리를 시도할 계획이다. 또한 캡셔닝 데이터셋을 활용하여 생성 모델과 이미지 캡셔닝 모델과 융합한 이미지 기반 문학 스타일 텍스트 모델을 제안할 계획이다.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A4A1023248).

참고문헌

[1] Dong, Chenhe, et al. "A Survey of Natural Language Generation." arXiv preprint arXiv:2112.11739 (2021).

[2] Feng, Xiaocheng, et al. "Topic-to-essay generation with neural networks." IJCAI. 2018.

[3] Yang, Pengcheng, et al. "Knowledgeable Storyteller: A Commonsense-Driven Generative Model for Visual Storytelling." IJCAI. Vol. 3. No. 6. 2019.

[4] Wang, Wei, Hai-Tao Zheng, and Zibo Lin. "Self-attention and retrieval enhanced neural networks for essay generation." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and

Signal Processing (ICASSP). IEEE, 2020.
[5] Seo, Jaehyung, et al. "KommonGen: A Dataset for Korean Generative Commonsense Reasoning Evaluation." Annual Conference on Human and Language Technology. Human and Language Technology, 2021.
[6] Gatt, Albert, and Emiel Krahmer. "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation." Journal of Artificial Intelligence Research 61 (2018): 65-170.
[7] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems 27 (2014).
[8] https://github.com/kakaobrain/kogpt
[9] https://github.com/SKT-AI/KoGPT2
[10] https://github.com/SKT-AI/KoBART
[11] 요약문 및 레포트 생성 데이터 세트, https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=582