

SMTR :

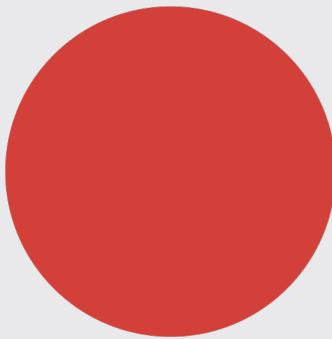
Show Me The Rhyme

간정현, 백인진, 용지호

자연어처리와 딥러닝

<https://github.com/eenzeenee/SMTR.git>

Contents



01

연구주제

- 주제 선정
- 의의

02

선행연구

- 어플리케이션
- 패러프레이즈
- sBert
- Bart

03

방법론

- 데이터 수집
- 모델 설명

04

결과

- 생성 결과
- 한계점

.....

“연구 주제

트랜스포머 기반 텍스트 스타일 변환
평서문을 가사 스타일 문장으로 변환

“의의

작사가들의 창작에 도움
텍스트 스타일 변환 태스크에서 트랜스포머 모델의 활용 가능성을 보임

66

선행 연구 0

AI - RAPSTAR (2018)



한글 자모 분리와 LSTM, Markov Chain을 활용한 가사 생성

KoGPT2 – FineTuning (2020)

KoGPT-2 기반 파인튜닝 가사 생성기 개발



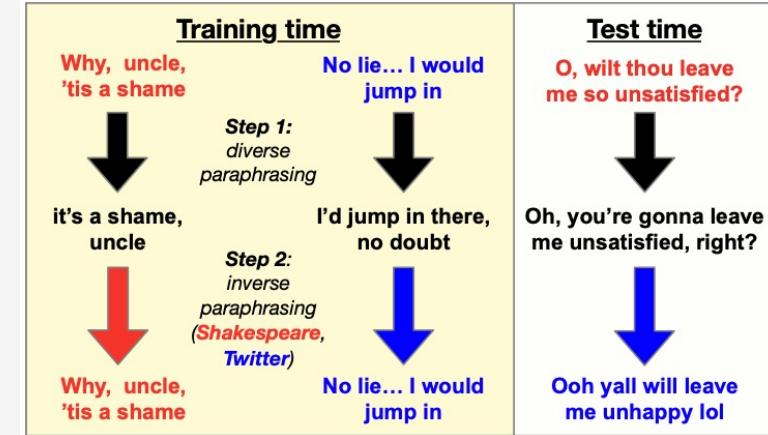
- 1) 생성 모델
 - 2) 단어의 반복 발생

“선행 연구 1”

Krishna, Kalpesh et al. (2020)

“Reformulating Unsupervised Style Transfer as Paraphrase Generation.”

텍스트 스타일 변환을 패러프레이즈 생성 문제로 보고 접근
패러프레이징 모델을 활용하여 pseudo-parallel 학습 데이터 생성
해당 학습 데이터를 활용하여 변환 모델 학습
GPT 사용한 모델



“

선행 연구 2

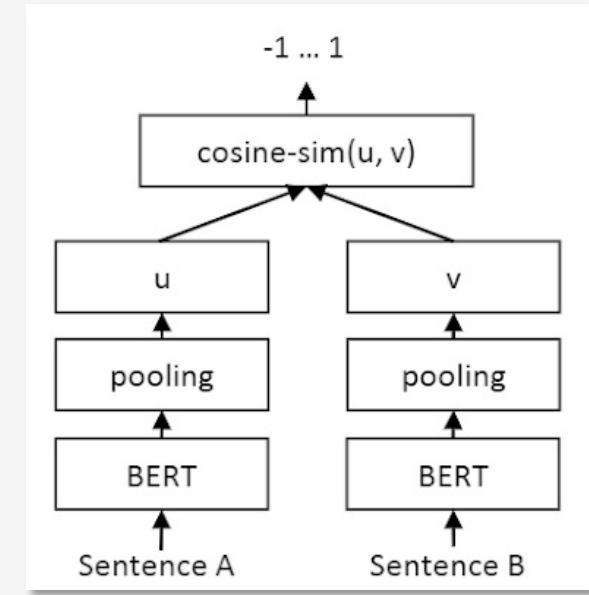
Reimers, Nils and Iryna Gurevych. (2019)

“Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.”

의미 있는 문장 임베딩을 도출하는데 특화된 BERT 모델

문장 유사도를 빠르게 구할 수 있기 때문에 paraphrase mining에서 유용

-> 평서문 – 가사 pseudo-parallel 데이터셋 구축

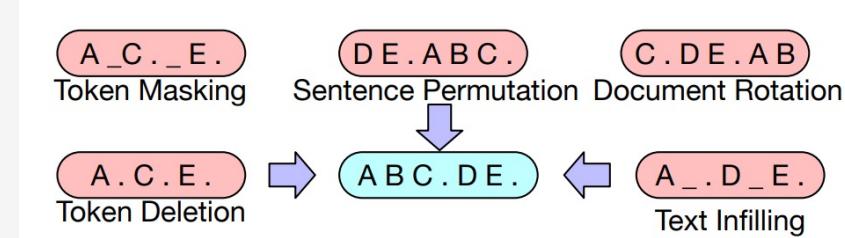


“

선행 연구 3

Lewis, Mike et al. (2020)

“BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.”



임의로 훼손된 입력 시퀀스를 원래의 시퀀스로 복원하도록 학습된 트랜스포머

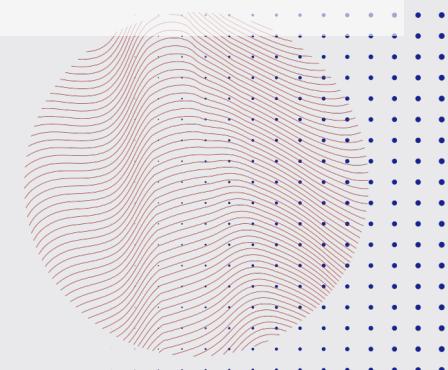
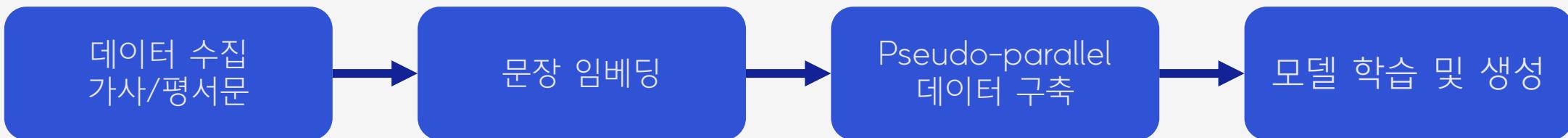
기존 연구 GPT 사용

BART 기반 모델 사용 결정

: 인코더를 통한 보다 많은 원천 문장에 대한 정보 추출 기대

방법론

- 가사 / 평서문 데이터 각각 수집
- sBERT 모델을 활용하여 평서문과 가사의 **pseudo-parallel** 데이터셋 구축
- KoBART 모델을 파인튜닝: **seq2seq** 지도학습



데이터 수집

벅스 가사 크롤링

- 국내 음악 11개 장르의 가사 데이터 수집
- 약 15만 건의 한국어 가사 데이터 수집
- 문장 분리 후 중복 제거
- 약 145만 개의 문장 데이터

The screenshot shows a mobile application interface for a music app. At the top, there is a search bar with the text "밤생작업조차 분위기있게, GroovyNights #6". Below the search bar, the title "Bugs!" is displayed in red. The main content area shows the lyrics for the song "눈이 오잖아(Feat.헤이즈)" by 이루진 (Feat. Hayez). On the left side, there is a list of genres under the heading "장르 | 발라드". The genres listed are: 국내 (Ballad), 해외 (Pop), 기타 (OST, Hip-Hop/Rap, Folk/Acoustic, Rock, Electronic, Metal, Indie, Classical, New Age, Jazz, CCM/Chorus, Children's, Religious, K-Pop, World Music, Country, Latin American, Folklore, etc.). The "국내" section is highlighted with a red box. The lyrics for "눈이 오잖아" are listed on the right side.

장르	발라드	해외	기타
국내	Ballad	Pop	OST
댄스/팝	Hip-Hop/Rap	Hip-Hop/Rap	Classical
포크/아쿠스틱	Folk/Acoustic	Folk/Acoustic	New Age
아이돌	K-Pop	Electronic	J-POP
랩/힙합	Rap/Hip-Hop	Rock	World Music
알앤비/소울	Alley-Bee/Soul	Metal	Country
일렉트로닉	Electronic	Indie	CCM/Chorus
락	Rock	Religious	Children's
알앤비/소울	Alley-Bee/Soul	Heavy Metal	Taiwanese
일렉트로닉	Electronic	Rock	CCM/Chorus
락/힙합	Rock/Hip-Hop	Indie	Armenian
재즈	Jazz	Christian	Tagalog
락/메탈	Rock/Metal	Indie	Catalan
재즈	Jazz	Christian	Georgian
인디	Indie	Christian	Welsh
성인가요	Adult Contemporary	Christian	Welsh

눈이 오잖아(Feat.헤이즈)

이루진 (Feat. Hayez)

눈이 오잖아

한 달 좀 덜 된 기억들
주머니에 넣은 채
걷고 있어 몇 시간을
혹시 몰라 네가 좋아했던
코트를 입은 채
나온 번화가 그때 마침
찬바람 막아줄
네가 이젠 없으니까
추울 때 따스히
안아줄 이가 없으니까
친구들이 불러도
나갈 수 없어 난
창문 너머
그저 바라봐 그때 마침
눈이 오잖아

데이터 수집

KorNLI 데이터셋 (Choe et al., 2020)

- 카카오브레인에서 공개한 한국어 이해 데이터셋의 일부
- Natural Language Inference 태스크 지도학습 데이터
- 텍스트만을 가져와 사용

Examples	Label
<p>P: 너는 거기에 있을 필요 없어. “You don’t have to stay there.”</p> <p>H: 가도 돼. “You can leave.”</p>	E
<p>P: 너는 거기에 있을 필요 없어. “You don’t have to stay there.”</p> <p>H: 넌 정확히 그 자리에 있어야 해! “You need to stay in this place exactly!”</p>	C
<p>P: 너는 거기에 있을 필요 없어. “You don’t have to stay there.”</p> <p>H: 네가 원하면 넌 집에 가도 돼. “You can go home if you like.”</p>	N

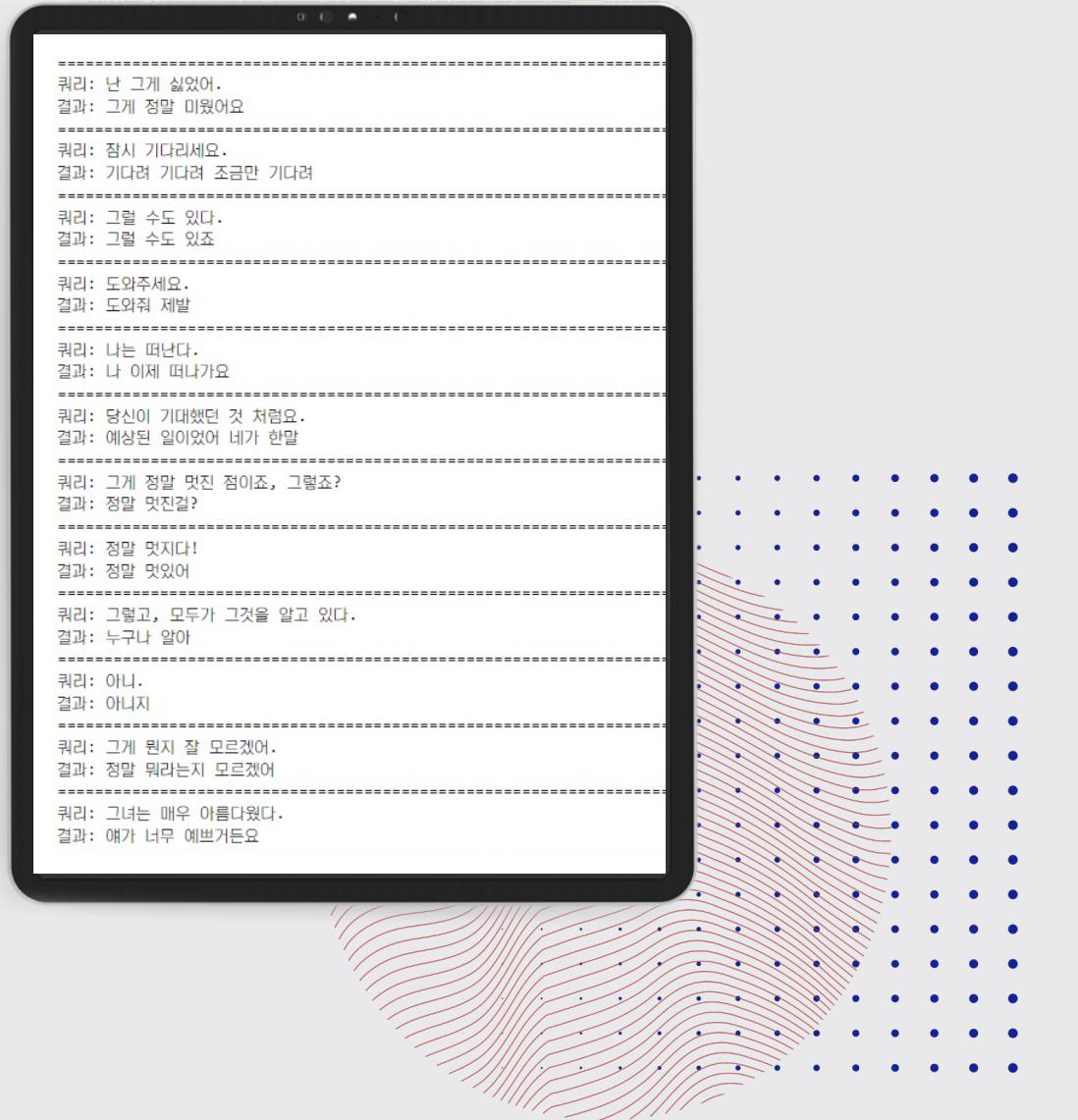


데이터 가공

sBert pretrained model

: *paraphrase-multilingual-mpnet-base-v2*

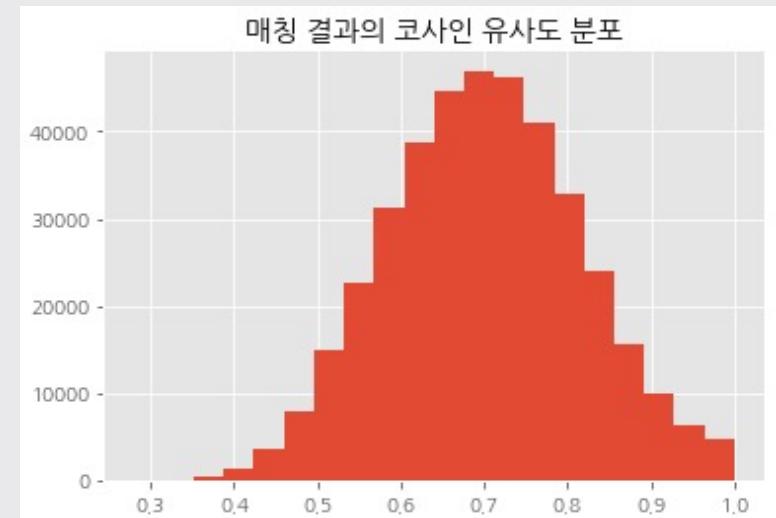
- 다국어 사전학습 모델 활용 (한글 데이터 적용 위해)
- 평서문 – 가사 유사도 계산
- 평서문에 대해 가장 유사한 가사를 매칭
- pseudo-parallel 데이터셋 구축



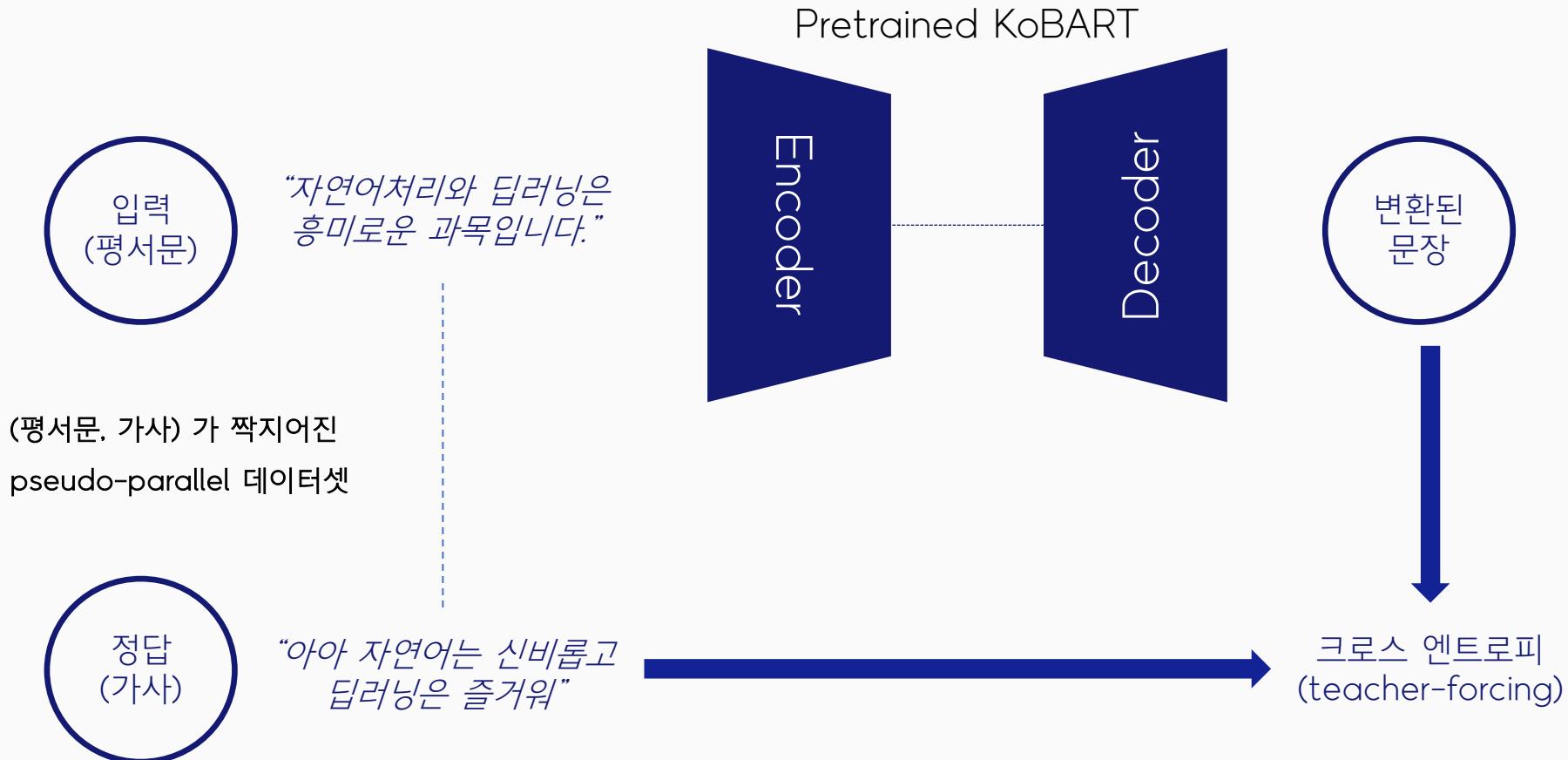
데이터 가공

- 약 39만 건의 평서문에 대해 유사한 가사를 매칭
- 매칭 결과 코사인 유사도 평균 0.7
- 데이터셋 구축 시 다양한 유사도 임계치 사용

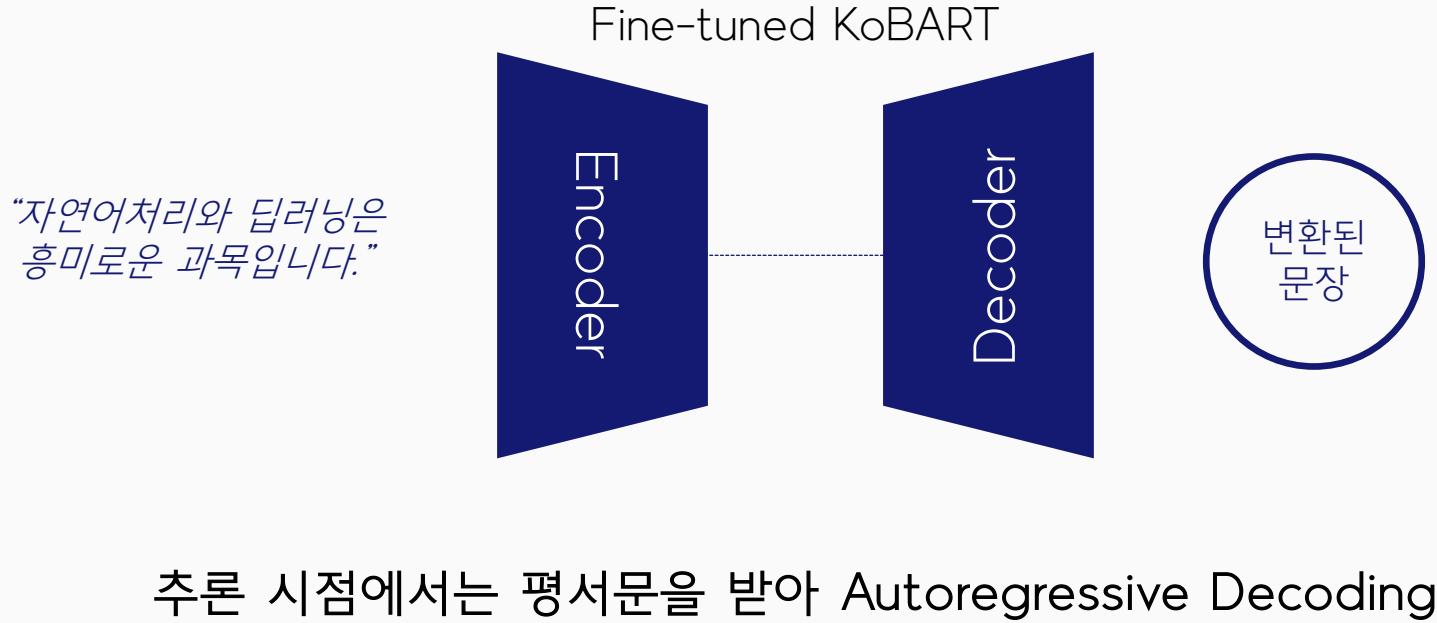
평서문	가사	cos_sim
너는 나보다 훨씬 나이가 많다.	넌 나보다 나이도 많으면서	0.979366
노인들은 매년 보험에 가입하는데 약 1,200달러가 든다.	산재보험받고 그 돈으로	0.647200
감사는 주가가 달성될지 확신할 수 없다.	성공은 장담 못해요	0.817735
GAO는 한 집에만 영향을 미친다면 요청자들과 협력할 것이다.	한방에 집을 찾아요	0.568887
나는 과거에 정신적 충격을 받은 사건으로부터 결코 감정을 느끼지 못한다. 바래져버린 내 기억들은 누구에게도 닿질 않아		0.709107



학습 목표(objective)



스타일 변환



하이퍼 파라미터 조정

학습

```
--batch_size  
64  
--max_epochs  
2  
--max_seq_len  
256
```

생성

```
--min_length  
16  
--max_length  
64  
--num_beams  
5  
--repetition_penalty  
5.0
```

생성 결과

source sentence: 기관들은 고위 간부들이 성취하도록 도와야 할 일련의 성과 목표를 가지고 있다.

generated sentence: 간부들의 뭐니 뭐니 해도 목표 달성을 위해 노력하는 게 중요하지

Top 5 most similar sentences in corpus:

뭔가 잘하려하고 끝없이 노력을해도 (Score: 0.8476)

변함없이 목표를 좇아서 (Score: 0.8399)

아무리 열심히 해도 (Score: 0.8278)

치여서 살지만 목표 라는 목적은 여전히 변함없이 존재 해 (Score: 0.8277)

열심히 노력해 포기하지 말고 매순간 난 성공해야 해서 (Score: 0.8254)



기존에 존재하는 가사와의 유사도 확인

source sentence: 저는 답지를 외우지 않았습니다.

generated sentence: 난 답도 못 외웠어 (정답은) 정답이 없었어

Top 5 most similar sentences in corpus:

아무런 대답도 할 수 없었지 (Score: 0.9013)

아무리 생각해도 답하지 못한 나 (Score: 0.8996)

난 알수가 없었으니 (Score: 0.8895)

가늠도 못했었던 나 (Score: 0.8884)

난 알 수가 없었어 (Score: 0.8860)

답지에 없는 데이터를 bart를 이용해 생성

생성 결과

source sentence: 우리는 젊고 돈도 많습니다

generated sentence: 우린 젊고 돈도 많아서 잘나가네 빡도 많아서 잘나가고 돈도 많아서 잘나가네

Top 5 most similar sentences in corpus:

돈도 많고 여유 많은 사람들도 많으면서 (Score: 0.8081)

부자가 되게 됐지 uh 우리사이 (Score: 0.7865)

흉부자부자부자 우리가 흉부자 (Score: 0.7660)

젊고 멋진 클럽 우리는 돈 되는 걸 좋아하잖아 (Score: 0.7582)

돈 많이 버소서 행복하소서 (Score: 0.7576)

source sentence: 모든 것들은 작게 시작한다.

generated sentence: 아주 작게 시작해봐도 소용없을거야 모든 게 작은 것들뿐이잖아

Top 5 most similar sentences in corpus:

작은것부터 시작해 모든일에 정답은 없어 (Score: 0.8484)

아주 작은 존재가 되면 (Score: 0.8459)

작은 것도 거슬리는지 (Score: 0.8290)

조금 무심한 그대가 작은 것, 사소한 것 (Score: 0.8226)

작은 것까지 (Score: 0.8211)

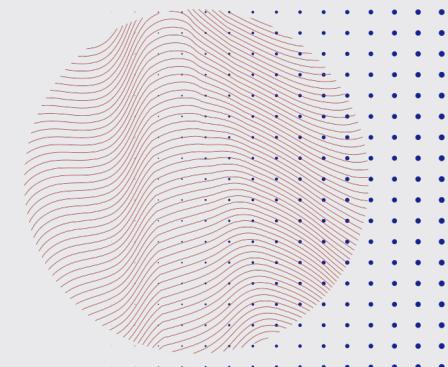
•
•
•
•
• “**한계점**

데이터셋 구축 측면

- 가사 데이터: 문장 단위 분리가 쉽지 않음 / 장르 구분의 필요성
- KorNLI 데이터: 비문, 지나친 번역체 문장이 많음
- sBERT 모델: 다국어 모델 활용

모델 측면

- 명확히 스타일을 분리해냈다고 볼 수 없음

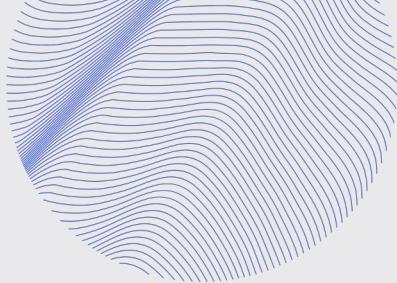




The background features a white surface with abstract elements: a large dark blue rectangle in the center containing the text, surrounded by red and blue wavy line patterns, red and blue circles, and a grid of blue plus signs.

Thanks !

참고문헌



- Krishna, K., Wieting, J., & Iyyer, M. (2020). Reformulating unsupervised style transfer as paraphrase generation. *arXiv preprint arXiv:2010.05700*.
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Ham, J., Choe, Y. J., Park, K., Choi, I., & Soh, H. (2020). Kornli and korsts: New benchmark datasets for korean natural language understanding. *arXiv preprint arXiv:2004.03289*.

