T5 모델을 활용한 반주 기반 가사 생성 기법에 관한 연구

장기태[1], 진태헌[1], 김두상[2]
[1]동서울대학교 컴퓨터소프트웨어학과 학부생
[2]동서울대학교 컴퓨터소프트웨어학과 교수
2470024@du.ac.kr, honey7844@naver.com, dskim@du.ac.kr

Research on Lyric Generation conditioned on Accompaniment using T5

Gi-Tae Jang, Tae-Heon Jin, Doo-Sang Kim
Dept. of Computer Software, Dongseoul University

요 약

본 논문은 T5(Text-To-Text Transfer Transformer) 모델을 활용한 반주 기반 가사 생성 기법을 제안하였다. 텍스트 이벤트 형식으로 변환한 정제된 반주를 "가사 생성" Task Token과 같이 T5에 적용하여 입력된 반주에 상응하는 가사를 생성하는 방식이다. 본 논문에서 제안한 방식의 성능 검증을 위해 Transformer, GPT-2, BART를 이용하여 가사를 생성한 출력물을 BLEU(Bilingual Evaluation Understudy) 값과 감정분석 일치도(Emotion Analysis Consistency) 결과값을 통해 비교 평가하였다. 본 논문에서 제안한 T5를 이용한 방식이 Transformer, GPT-2, BART를 사용하는 방식보다 우수한 결과를 얻었다.

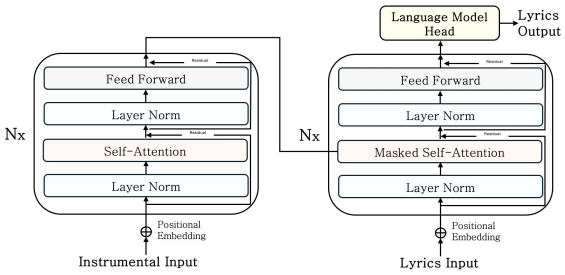
1. 서 론

최근 인공지능(Artificial Intelligence) 기술의 급속한 발전은 창작의 영역에 혁신적인 변화를 가져왔다. 특히, 음악과 가사 생성 분야에서의 인공지능 활용은 창작자들에게 새로운 가능성을 열어주고 있다. 이 중에서도 NLP(Natural Language Processing) 기술의 발전은 가사 생성에 특별한 관심을 불러일으키고 있다[1]. NLP 기술을 활용하여, 기계는 이제 인간의 언어를 이해하고, 창의적인 텍스트를 생성할 수 있게 되었다. 본 논문에서는 NLP 분야에서 주목받고 있는 T5(Text-to-Text Transfer Transformer) 모델을 활용하여 반주 기반 가사 생성 시스템을 제안하였다. T5 모델은 '텍스트에서 텍스트로의 변환'이라는 개념을 기반으로 하여, 주어진 입력 텍스트에 대해 다양한 형태의 출력을 생성할 수 있는 능력을 가지고 있다. 이는 가사 생성과 같은 창의적 작업에 있어 특히 유용하며, 음악의 반주와 같은 다양한 입력에 맞춰 개성 있는 가사를 생성할 수 있는 잠재력을 가질 것으로 판단된다. 본 논문에서 T5 모델을 활용한 반주 기반 가사 생성기를 설계하고, 이 생성기로 생성된 가사의 품질을 타 방식과 비교 평가하여 제안한 방식의 성능을 검증하였다.

2. 본론

2.1 기존 방식

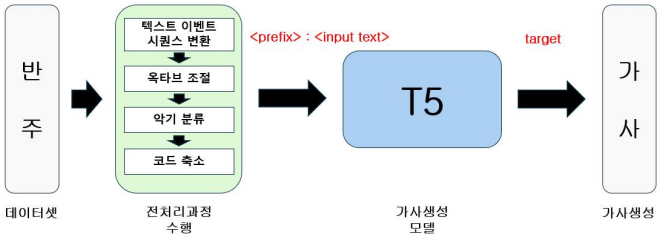
기존 반주 기반 가사 생성 연구는 트랜스포머(Transformer) 기반 모델을 활용하는 방식으로, 대규모 데이터셋에서 학습을 통해 언어의 복잡한 구조를 학습하며, 이를 기반으로 보다 정교하고 창의적인 가사를 생성하는 능력을 확인할 수 있었다[1].



(그림 1) 기존 방식의 구조도

2.2 제안한 방식

본 논문에서 제안한 방식은 가사 생성 모델을 T5 모델로 선정하였다. T5 모델은 "Text-To-Text" 접근 방식을 채택하여, 모든 자연어 처리 작업을 텍스트 입력을 받아 텍스트 출력을 생성하는 방식을 적용한다. 이 접근 방식은 다양한 NLP 작업에서 높은 유연성과 성능을 제공하는데. 일괄된 프레임워크를 통해 다양한 작업을 동일한 모델 아키텍처와 학습 접근 방식으로 처리하기 때문이다. 또한, 스팬 손상(Span Corruption) 학습 방식을 적용하여, 모델이 더 넓은 컨텍스트를 이해하고, 주어진 반주에 맞는 가사를 생성할 때 중요한 요소와 문맥을 더 잘 파악할 수 있게 하는 장점이 있다[2].



(그림 2) 제안한 방식의 구조도

3. 성능 검증 및 실험

3.1 제안한 방식의 학습 과정

학습 과정에서는 먼저 T5Tokenizer를 사용하여 반주, 가사 데이터를 토큰화한 후 모델이 텍스트를 이해할 수 있는 형태로 변환하였다. 토큰화된 데이터는 MidiDataset 클래스를 통해 입력 형태를 "가사 생성 : 반주 데이터"로 처리하고 타깃으로 "가사 데이터"로 지정하였다. T5 모델은 입력 데이터 앞에 접두사(Prefix)를 붙임으로써 데이터를 Text-To-Text 형태로 작업을 진행하기에 이 클래스는 접두사를 붙이고 데이터를 정제하는 역할을 수행한다. 이 과정을 진행한 후, 각 토큰화 된 반주와 가사를 쌍(pair)으로 모델이 처리할 수 있는 형태로 인코딩한다. 모델의 학습은 T5-small 모델을 불러와 입력된 반주 텍스트에 대해 가사 텍스트를 생성하는 과정으로 전이학습을 진행한다. 옵티마이저는 AdamW를 사용하였고, 손실 함수를 통해 예측과 실제 가사 사이의 오차를 계산한다. 제안한 방식의 학습에서 적용된 하이퍼 파라미터 정보는 표 1과 같다.

<표 1> 하이퍼 파라미터 정보

Model Parameter	값
Embedding Size	512
Batch Size	8
Epoch	50
Learning Rate	1e-4
Source Max_Len	512
Target Max_Len	512
Loss Function	CrossEntropyLoss

3.2 성능평가에 사용한 데이터셋

다양한 장르에 대해 음악 MIDI 파일과 가사, 음악 정보 등을 포함하는 Lakh MIDI 데이터셋(LMD)을 성능평가를 위한 학습과 검증 데이터셋으로 활용하였다. 데이터셋에서, 반주 데이터와 가사 데이터를 추출하고, 반주 데이터는 노트 시작, 노트 종료, 대기시간 단위로 텍스트 이벤트 시퀀스로 변환 후, 로마 숫자 분석을 통해 축소된 텍스트 이벤트 시퀀스로 생성하였고, 가사 데이터는 영어 가사만을 선택하고 메타데이터를 제거하여 표준화된 가사 텍스트를 생성하였다. 표 2는 전처리된 데이터셋 형식이다 [1].

<표 2> 반주, 가사 데이터 형식

Instrumental	Lyrics
"_DB_", "_REST_", "W_960", "_B_", "ON_1", "W_2000", "W_1360", "ON_1", "W_160", "ON_1", "W_80", "ON_1", "W_240", "_B_", "ON_V6", "W_240", "ON_V6", "W_2000"	"Give", "me", "time", "\nTo", "re", "a", "lise", "my", "crime", "\nLet", "me", "love", "and", "steal", "\nI", "have"

3.3 실험 결과

본 논문에서는 생성된 가사와 기존 가사 사이의 BLEU(Bilingual Evaluation Understudy) 값을 계산함으로써 모델의 성능을 평가하였다. BLEU 값은 모델이 생성한 텍스트의 자연스러움과 기존 가사와의 유사성을 정량적으로 측정하는 메트릭(metric)으로, 자동 텍스트 생성 작업에서의 출력물 품질을 평가하는 것으로 사용된다. 또한, 생성된 가사의 감정 일치도를 측정하여, 가사가 전달하고자 하는 감정적 내용의 정확성을 평가하였다. 본 논문에서 제안한 방식의 성능을 기존 방식과 비교 평가하기 위하여, 동일한 데이터셋을 사용하여 제안한 방식과 Transformer, GPT-2, BART를 사용한 방식을 BLEU 값과 감정분석 일치도 결과값을 통해 비교 평가하였다. 비교 평가한 결과는 표 3과 같다.

<표 3> 각 적용 모델의 검증 결과표

평가방식 / 적용모델	BLEU Score (2-gram)	BLEU Score (3-gram)	감정분석 일치도
Transformer	8.61	7.97	52.71%
GPT-2	4.18	3.68	62.19%
BART	15.12	14.50	74.49%
제안한 방식	27.20	26.48	78.50%

표 3의 결과를 보면 제안한 방식의 BLEU 값이 다른 모델 적용 방식보다 2-gram에서 18.59, 23.02, 12.08, 3-gram에서 18.51, 22.80, 11.98 값만큼 더 높은 값을 나타내는 것을 알 수 있다. 여기서 n-gram은 연속된n개의 아이템(단어, 문장 등)을 의미하는 통계적 언어 모델의 기본 단위이다 [3]. 또한, 감정분석 일치도에서도 각각 25.79%, 16.31%, 4.01% 더 높은 값을 얻었다. 결론적으로 제안한 방식은 다른 모델을 사용한 방식보다 전체적으로 향상된 성능을 나타낸 것을 알 수 있다.

-기존 가사-	-생성 가사-
You're just too good to be true Can't take my eyes off you You'd be like heaven t touch I wanna hold you so much At long last love has arrived And I thank God I'm a live You're just too good to be true Can't take my eyes off you (생략)	I've been thinking of you forever But I know it's hard to take a look at your face but it seems so wrong that you're not the right thing and then doing what I'm looking back on me this time too long chasing all my thoughts (생략)

(그림 3) 기존 가사와 생성된 가사 결과 화면

4. 결론

본 논문에서 제안한 T5 모델을 활용한 반주 기반 가사 생성 방식이 기존의 Transformer, GPT-2, BART 모델을 적용하는 방식보다 향상된 성능을 보임을 확인할 수 있었다. 이러한 결과는 T5 모델의 강력한 언어 이해 능력과 문맥적 정보 처리 능력이 가사 생성과 같은 창의적인 작업에 특히 유용함을 증명하였다. 본 논문의 연구결과를 통해 T5 모델이 음악 창작 분야에서 기존보다 창의적인 콘텐츠 생성에 기여할 것으로 기대된다. 향후 연구는 한국어 텍스트의 특화된 KoT5 모델과 한국가요 데이터셋을 통한 K-POP 장르에서의 반주 기반 가사 생성 연구를 진행하고자 한다.

참 고 문 헌

[1] Thomas Melistas et al., Lyrics and Vocal Melody Generation Conditioned on Accompaniment, 2nd Workshop on NLP for Music and Spoken Audio (NLPMusA), 2021, 18-23.

[2] Colin Raffel et al., Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Journal of Machine Learning Research (JMLR), 21, 140, 1-67, 2020.

[3] Papineni et al, BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002., 311 - 318