

가천 X 세종

연합 학술제

연구 파트 보고서

Prefix-Tuning 기법과 KoGPT-2 모델 기반 한국어 시 생성 모델 제안

Abstract

본 연구는 스트리밍의 확산으로 가사 접근성이 높아지고 가사가 독자적 텍스트로 조명되는 변화, 그리고 대화형 인공지능의 부상을 배경으로, 노래 가사와 시의 경계를 가로지르는 한국어 시 생성의 가능성을 탐색한다. 한국어 사전학습 언어모델 KoGPT-2를 기반으로 Prefix-Tuning 기법을 적용해 시 데이터와 장르별 가사 데이터를 학습시켜 3개의 모델을 구축하여 장르적 특성이 시 생성에 반영되는지 검증하는 것을 목표로 한다.

Perplexity와 Train/Eval loss를 측정하여 모델의 안정적인 수렴과 학습 성능을 확인하였으며, 생성된 시는 sBERT 기반 문장 임베딩을 활용한 유사도 비교와 정성적 분석을 통해 평가하였다. 각 모델은 동일한 데이터셋에 대하여 서로 다른 유사도 Top1 시를 선택하였으며, 이를 통해 장르적 표현 양식, 정서적 경향이 시 생성 과정에 뚜렷하게 반영됨을 확인하였다. 나아가 문학적 평가에서는 장르별 정서적 어휘와 운율적 특성이 시적 맥락 속에 자연스럽게 드러나, 장르에 따라 구별되는 창작 양상을 보여주었다.

이러한 시도는 시와 가사의 장르적 경계를 허물고, 운율적 언어의 새로운 표현 양식을 탐색하는 동시에, 인공지능 기반 문학 창작의 새로운 방향성을 제시한다.

팀 이름: UN-LAC(언락)

제출일 2025.08.22

팀원

가천대학교
남경민

팀원

세종대학교
이민기

Introduction

2010년대를 기점으로 음반 산업의 중심축이 실물 앨범에서 스트리밍 서비스로 이동함에 따라, 대중은 노래 가사에 더욱 쉽게 접근할 수 있게 되었다(고정민, 2024). 이러한 변화는 노래 가사가 단순한 음악의 부속물이 아니라 독자적인 텍스트로서 주목받게 되는 계기가 되었으며, 문학적 가능성 또한 함께 논의되기 시작하였다. 임수진(2018)은 노래 가사를 멜로디 없이 낭송하는 시적 활동이 창의적 표현과 문학적 사고의 확장으로 이어질 수 있음을 제시하며, 가사가 시의 영역으로 확장될 수 있는 잠재성을 보여주었다.

한편, 인공지능 기술의 발전은 문학 창작의 영역에도 새로운 가능성을 제시하고 있다. 최근 대화형 인공지능을 활용한 시 창작 사례들은, 기존의 창작 주체 개념을 재고하게 만든다. 박성준(2023)은 ChatGPT를 활용하여 윤동주의 작품 세계를 모작한 실험을 통해, AI가 인간 시인의 시적 메커니즘과 사유 방식을 일정 부분 재현할 수 있다는 점을 강조하였다. 이는 단순한 문체의 복제를 넘어서, 인공지능이 문학적 정동에 접근할 가능성을 시사한다.

본 연구는 이러한 배경 속에서 노래 가사와 시의 경계를 넘나드는 창작 모델의 가능성을 탐색하고자 한다. 기존의 시 생성 모델들이 정형화된 시 문학에 집중되었던 반면, 본 연구에서는 한국어 사전학습 언어모델 KoGPT-2를 기반으로 하여, 시와 장르별 노래 가사를 함께 학습시키는 시 생성 모델을 설계하였다. 특히 가사는 운율, 비유, 정서 등의 문학적 특성을 보이고 있음에도 불구하고 시 생성 연구에서는 상대적으로 소외된 텍스트였기에, 이를 시적 텍스트로 전환하는 실험은 창작 영역의 확장을 의미한다.

또한 본 연구는 발라드, 힙합, 인디 음악의 장르별 가사 데이터를 개별 모델에 따로 학습시켜, 장르적 특성이 반영된 시 생성 결과를 비교하고자 한다. 이를 통해 단순한 시 생성 모델을 넘어서, 장르적 정체성이 녹아든 시 생성 가능성을 탐색하고자 한다.

Related work

한국어 시 생성 연구

김명선 외(2023)는 한국어 시 생성에 특화된 인공지능 시스템을 구축하고, 생성 품질과 창작 지원 효과를 다각도로 평가하였다. 이 연구는 SKT의 KoGPT-2

언어모델을 기반으로 서정시, 산문시, 소설의 세 장르에 대해 파인튜닝을 수행하였으며, 사용자 입력 키워드 기반의 시 생성, 제외어 설정 기능, 유사도 검증 등 실용적인 기능을 포함한 전체 창작 파이프라인을 제안하였다.

특히 학습 데이터에 대한 과적합 문제를 방지하기 위해 sBERT 기반의 문장 유사도 평가 모듈을 탑재하였고, 생성 결과의 창의성과 맥락 일관성을 동시에 확보하기 위해 Temperature, Top-p 등 다양한 디코딩 하이퍼파라미터를 실험적으로 적용하였다. 평가 지표로는 문장 유사도, PPL(Perplexity), 사용자 설문 등을 함께 활용하여 모델의 성능을 정량적·정성적으로 검증하였다.

정용재 외(2023)는 KoGPT-2를 활용해 한국 근대시와 현대시 데이터를 기반으로 시를 생성하고, KoBERT를 통해 생성된 시의 시대 구분을 수행하였다. 단어 빈도 분석을 통해 시대별 언어 특성을 파악한 뒤, 이를 반영한 시 생성 실험을 진행하였으며, 주제어(예: “사랑”, “마음”)를 중심으로 생성된 시가 특정 시대의 표현 경향과 일치하는지를 정량적으로 평가하였다. 생성 시에 대한 분류 정확도는 99% 이상으로 나타났다.

한국어 가사 생성 연구

박찬솔 외(2021)는 한국어 랩 가사 생성에 있어 문맥 정보와 운율 정보를 동시에 반영한 단어 임베딩 기법과 어텐션 기반 인코더-디코더 구조를 제안하였다. 이 연구는 한국어 음절의 구조적 특성(초성·중성·종성)을 고려하여 단어를 문자 단위로 분해한 후, 운율 정보를 추출하여 벡터로 표현하였다. 해당 운율 벡터는 문맥 임베딩과 결합해 문맥-운율 통합 단어 임베딩을 구성하며, 이를 입력으로 하는 seq2seq 모델과 어텐션 메커니즘을 통해 운율 구조를 반영한 가사 생성을 가능하게 했다.

국주현 외(2023)는 멜로디 정보를 반영한 한국어 가사 생성을 위해 SeqGAN 기반 모델을 제안하였다. 해당 연구는 3,090개의 곡에서 멜로디-가사 쌍 데이터를 구축하고, MIDI 형식의 멜로디에서 추출한 여러 특성과 형태소 단위로 전처리된 가사를 입력으로 활용하였다. 평가에서는 BLEU score와 감정 분석을 사용하여 생성된 가사의 품질을 검증하였다.

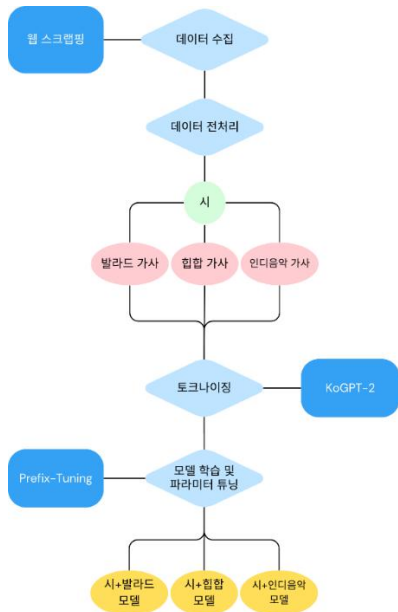
최광희(2021)는 기존 가사 생성 연구에서 Rhyme 요소를 반영한 차별적인 한국어 랩 가사 생성 시스템을 구축하였다. 해당 연구에서는 일부 마스킹 된 랩 가사를

입력 받아, 마스킹 된 부분에 들어갈 가사 후보를 여러 개의 가사 생성기를 통해 생성한 뒤, Rhyme Filter와 Lyrics Ranker를 통해 유의미한 결과를 선별하고 정렬하는 구조를 제안한다.

Rhyme Filter는 문장 간의 최대 공통 하위 문자열을 활용하여 라임 밀도(Rhyme density measure)를 정의하고, 라임 밀도가 낮은 단어는 후보에서 제외한다. Lyrics Ranker는 모델의 자동회귀 특성을 기반으로 실제 말뭉치에서 등장할 확률이 높은 순으로 가사 후보를 제시한다.

Methodology

1. 학습단계



<그림 1> 모델 학습 Architecture

학습 단계에서의 전체적인 Flowchart는 <그림 1>과 같다.



<그림 2> 학습 데이터 통계

본 연구에서는 시와 노래 가사 총 2종류의 데이터를

수집한다. 시 데이터는 장르의 구분 없이 크롤링을 통해 총 4,040개의 데이터 수집을 진행했다. 가사 데이터는 국내 최대 음원 사이트인 멜론에서 분류하는 장르에 따라 발라드(517개), 인디음악(567개), 힙합(600개) 총 3가지의 장르별로 각각 크롤링을 통해 총 1,684개의 가사 데이터를 수집하였다.

1.1 시 데이터

BeautifulSoup 기반 웹 스크래핑으로 시 데이터셋은 온라인 시 커뮤니티에 게시된 사용자 생성 시를 수집하였다. 해당 게시글들은 정해진 양식 없이 자유로운 형식으로 구성되어 있어 전체 데이터셋에 대한 전처리와 Column 별('제목', '내용') 전처리로 나누어 진행하였다.

전체 데이터셋에 대하여 크게 세 가지 전처리를 적용하였다. 첫째, 웹 스크래핑 시 데이터 로딩이 10초 이상 지연되면 timeout 처리("요청 실패")로 기록되었으며, 이를 결측치로 간주하여 제거하였다. 둘째, '제목'과 '내용'이 일치하는 38개의 중복 데이터 제거하여 독립적으로 구성된 시 데이터를 확보하였다. 셋째, 학습에 불필요하고 완전한 한국어 변환이 어렵다고 판단되는 제주 방언이 포함된 시를 제외하였다. 이는 학습의 일관성을 유지하고, 방언으로 인한 성능 저하를 방지하기 위함이다.

이후 Column 별 전처리를 진행하였다. 먼저 '제목'의 불필요한 정보를 제거하였다. 시인의 이름 또는 한문 번역이 함께 기재된 경우, 시 내용과 직접적인 관련이 없다고 판단되어 제거하였다.

다음으로 '내용'에 대하여 4가지 정제를 적용하였다. 첫째, 시 본문 앞뒤에 불필요한 공백과 줄 바꿈 기호를 제거하여 문장 구조를 정돈하였다. 둘째, 본문의 다국어 버전을 삭제하였다. 시의 영문, 일문 번역이 포함된 경우, 구분 선(---, === 등)을 기준으로 원문과 번역이 구분되었다. 이러한 구조적 특징을 이용하여 조건문 필터링을 적용하고, 불필요 번역을 제거하였다. 셋째, 시 본문 하단에 작성된 날짜 정보('2025.08.22')는 모델 학습과 무관한 메타데이터로 판단해 제거하였다. 마지막으로, 시의 내용 외적인 데이터(제목, 시인 이름)를 제거하였다. 특히 제목이 반복 기재되거나 시인의 이름이 다양한 형태(괄호 포함, 한문 단독, 한글+한문 병합 등)로 삽입된 경우가 많았다. 이에 대해 제목과 내용의 불필요 데이터가 완전히 일치하는 경우, 제목과 한문이 함께 작성된 경우, 한문만 포함된 경우,

괄호 또는 기호와 함께 작성된 경우 등 5가지 패턴을 정의하여 제거하였다.

1.2 가사 데이터

본 연구는 가사 데이터 수집 단계에서 다음의 정제 절차를 거쳤다. 가사가 제공되지 않는 연주곡은 수집 대상에서 제외하고, 전체 텍스트 중 한국어 비율이 50% 이하인 곡 역시 연구 목적과의 적합성이 낮다고 판단해 배제하였다. 또한 노래 제목에 관행적으로 포함되는 'feat.', 'prod.' 등의 표기는 의미 분석 기여도가 낮아 제거하였다. 한편, 힙합 장르에 대해서는 욕설·비속어 표현을 수집해 빈도와 표기 형태를 분석하고, '시발/씨발/x발/ㅅㅂ/ㅈㅂ' 등 동의 표기를 정규표현식으로 통합하였다. 탐지된 항목은 맥락 보존을 위해 삭제하지 않고 중립적 마스킹 토큰 '[욕설]'로 치환하였다.

2. 생성단계

2.1 파라미터 튜닝(Prefix-Tuning)&모델(KoGPT-2)

본 연구에서는 사전 학습된 KoGPT-2 모델을 사용한다. GPT 모델은 트랜스포머의 디코더를 스택 형태로 구성하고, 입력 시퀀스로부터 다음 토큰의 확률 분포를 예측하여 이를 입력 시퀀스에 순차적으로 재입력해 전체 문장을 생성하는 자기회귀 구조다.

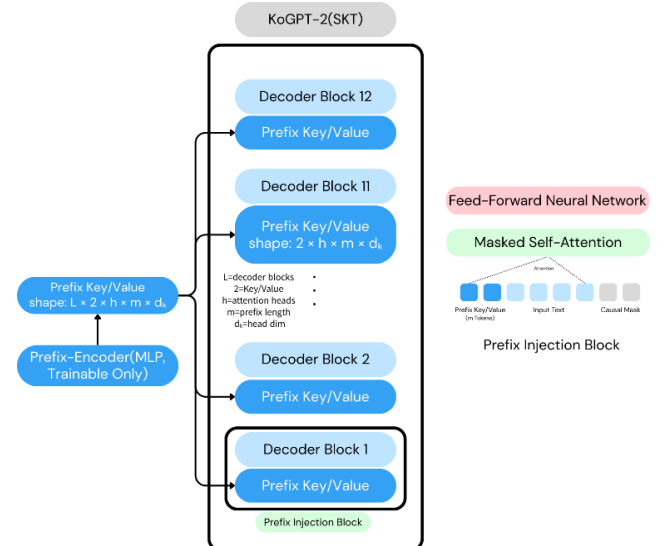
이를 기반으로 약 40GB 한국어 데이터셋을 학습한 KoGPT-2는 총 12개의 디코더로 구성되어 한국어 문장 생성 및 이해에 최적화되어 있다. 해당 모델은 여러 분야의 말뭉치로 사전 학습되었기에 파라미터 튜닝을 통해 출력을 연구 목적에 맞게 조정할 필요가 있다.

전통적인 파인튜닝(fine-tuning) 방식은 모델의 모든 파라미터를 시·가사 데이터에 맞게 조정함으로써 높은 표현력을 가진 출력을 생성할 수 있다는 장점이 있다. 그러나 이 방식은 막대한 컴퓨팅 자원과 긴 학습 시간을 요구한다는 단점이 있다.

반면, Prefix-Tuning은 기존 모델의 파라미터를 그대로 유지한 채, Prefix Key/Value를 생성하여 모델의 출력을 학습시킨 데이터에 맞게 튜닝하는 방식이다. 이 방법은 학습 속도가 빠르고, 연산 자원이 적게 소모되며, 학습된 Prefix-Encoder는 동일한 모델 기반의 다양한 응용 과제에 바로 적용할 수 있다는 점에서, 여러 모델을 효율적으로 생성해야 하는 본 연구에 적합한 접근법이다.

또한, Large GPT-2 기준 전체 파라미터 개수의 약 2% 수준의 파라미터를 학습시켜 기존 파인튜닝과 유사한

수준의 성능을 달성할 수 있다는 연구 결과에 기반하여(Xiang Lisa Li 외, 2021), 본 연구에서는 사전 학습된 KoGPT-2 모델을 기반으로 한국어 시·가사 데이터를 이용해 Prefix-Tuning을 수행한다. 전체 모델 구조는 <그림 3>에 제시하였다.



<그림 3> Prefix-Tuning 적용한 KoGPT-2 모델 구조

Prefix Encoder는 m개의 가상 임베딩 토큰을 입력으로 받아 소규모 MLP로 각 디코더 블록·헤드용 Prefix Key/Value를 생성한다(전체 shape: $L \times 2 \times h \times m \times d_k$, per-layer: $2 \times h \times m \times d_k$). 생성된 K/V는 각 디코더 블록의 self-attention K/V 시퀀스 앞에 결합해 모든 시점의 Q가 이를 참조한다. 학습할 수 있는 파라미터는 Prefix Encoder뿐이고, 기반 모델 파라미터는 모두 고정된다.

2.2 하이퍼파라미터 셋팅

본 연구에서 적용한 하이퍼파라미터의 조정 범위는 크게 모델 학습 단계와 문장 생성 단계로 구분된다.

모델 학습의 하이퍼 파라미터는 모델의 수렴 속도와 학습 효율을 제어한다.

모델 학습 단계에서의 하이퍼 파라미터 셋팅은 에폭(epoch)은 5, 학습률(learning rate)은 $1e-4$, 배치 사이즈 (batch size)는 8이며, 학습 입력의 토큰(token) 개수는 256개로 설정한다.

모델 학습 단계에서 적용한 하이퍼파라미터는 사전 학습되지 않은 Prefix-Encoder와 학습 데이터의 특성, T4 GPU 리소스를 고려하여 효율적으로 설정하였다.

생성 단계에서의 하이퍼 파라미터는 주로 생성 결과에 포함될 단어를 선택하는 기준점을 제시한다. repetition penalty는 기존 생성된 단어에 로그 확률을 감소시키는 방식을 적용하여, 동일한 단어의 반복

생성을 제약할 수 있다. 단어의 확률분포를 조절하는 또 다른 방법으로는 Temperature 파라미터가 있다. Temperature는 softmax 기울기를 변화시킴으로써 단어별 확률 분포의 극단성을 조절하고 다양성을 통제한다. 값이 1에 가까울수록 확률 분포는 동일하게 유지되어 생성 결과에 포함되는 단어의 다양성이 늘어나게 되고, 0에 가까울수록 기존 확률이 큰 단어에 집중되어 생성 결과의 일관성이 강해지게 된다.

샘플링을 통해서도 생성 결과에 포함될 단어의 다양성을 제한할 수 있다. 일반적으로 많이 쓰이는 방법으로 Top-K 샘플링과 Top-P 샘플링이 있다.

Top-P 샘플링은 누적 확률 분포 값이 p 이내인 단어들만을 후보로 선별한다. 선별한 후보 단어 내에서 무작위 선택을 진행하기에 확률적 다양성을 확보하면서도 극단적인 확률을 가지는 단어의 포함을 방지할 수 있다. Top-K 샘플링은 확률이 높은 k개의 단어 토큰만 샘플링해 단어의 분포를 제한한다. k의 값이 낮을수록 선행 단어와 연관성이 낮은 단어가 나오는 것을 방지할 수 있다.

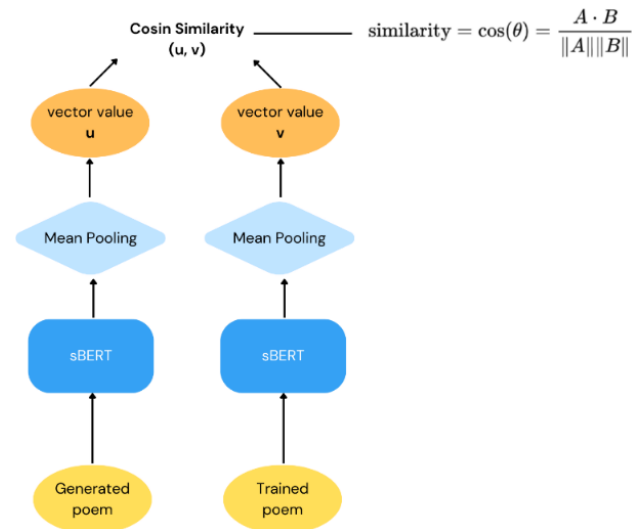
시 생성은 일반적인 문장 생성보다 표현의 창의성과 어휘의 감성적 함의가 중요한 만큼, Temperature나 Top-P 값을 적절히 조정하여 의미 왜곡 없이 다양성을 확보하는 것이 필요하다.

최종적으로 각 모델의 출력 결과 비교를 위해 결정한 하이퍼 파라미터 셋팅은 temperature는 0.8, p는 0.9, k는 50, repetition penalty는 1.2로 설정한다.

3. 평가단계(유사도 체크)

모델의 훈련된 정도 및 성능을 파악하기 위해 생성된 시와 학습 시 데이터의 의미상 유사도를 비교한다. 이를 위해서는 문장들을 토큰나이징 후 벡터로 임베딩하는 과정이 필요한데, 이 과정에서 두 문장 간의 관계 학습을 통해 BERT의 문장 임베딩 성능을 우수하게 개선한 sBERT(SentenceBERT)를 활용하였다.

sBERT는 각 문장에 대한 함축 정보를 통해 데이터를 학습하며, 각각의 문장 벡터 정보를 활용하여 문장 간 유사도를 분석하는데 효율적인 모델이다. 문장 또는 단어를 고차원 벡터로 임베딩한 뒤, 코사인 유사도를 통해 의미적 유사성을 정량화하는 STS(Semantic Textual Similarity, 문장 쌍 회귀 태스크로 의미 유사성 측정) 방식을 제공하기 때문에 문장 간 의미 구조 비교에 널리 활용된다(Reimers, N. 외, 2019). sBERT의 구조는 <그림 4>와 같다.



<그림 4> sBERT 모델 구조

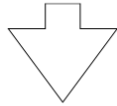
학습에 사용한 전체 시 데이터셋과 모델이 생성한 시를 각각 BERT의 입력으로 넣은 후 mean pooling을 통해 임베딩 벡터 u와 v를 계산한다. 그 후 u 벡터와 v 벡터 기반 코사인 유사도를 측정하여 문장 의미 간 유사도를 측정한다.

Experiment

언어 생성 모델에서는 입력 프롬프트의 품질이 결과물을 좌우한다. 본 연구가 사용하는 KoGPT-2는 디코더 전용(decoder-only) 구조로, 프롬프트가 사실상 유일한 조건 신호로 작동하므로 그 선정이 특히 중요하다.

본 연구에서는 출력에서 시적 표현과 장르적 특성을 최대한 반영하기 위해 발라드, 인디음악, 힙합, 시의 네 데이터셋에서 대표 어휘를 추출하였다. 우선 각 데이터셋에서 출현 빈도가 높은 단일 어절(단어)을 수집하고, 동의어·중복 표현을 정규화하여 정제 목록을 구성하였다. 이어 정제 목록에서 두 개의 단어를 복원추출 방식으로 결합해 2어절 후보 프롬프트를 생성하였다. 마지막으로 연구진이 문법적 타당성, 의미적 자연스러움의 기준으로 수작업 검토하여 실제로 사용 가능한 후보만을 선별하였다. 힙합 가사 데이터의 최종 입력 프롬프트 생성 과정을 예시로 <그림 5>에 명시하였다.

['다시', '사랑', '그대', '이젠', '정말', '함께', '눈물', '그냥', '모두', '우린', '지금', '혼자', '결에', '제발', '어떻게', '아직', '모습', '오늘', '매일', '그리워', '마음', '좋은', '하루', '안돼', '마지막', '그게', '이미', '나와', '돌아와', '너무나', '잊을', '아픈', '작은', '멀리', '이대로', '아주', '때문에' ...]



Prompts = ['지금 우린', '제발 이젠', '우린 언제나', '다시 함께', '가끔 만나', '지금 살아', '함께 전부', '항상 위로', '이미 사랑', '모두 사랑', '사랑 다시', '사랑 미안', '다시 못해', '전부 위로', '살아 오늘', '만나 절대', '여전히 여기' ...]

<그림 5> 힙합 가사 기반 bigram 입력 프롬프트 생성

이렇게 도출한 장르별 프롬프트 목록은 모델의 초기 컨텍스트로 제공되어, 생성 과정에서 장르적 감수성과 시적 표현을 보다 안정적으로 유도한다. 또한 입력 프롬프트가 시의 “제목”임을 명시하기 위해 실제 모델 입력은 “제목: [입력프롬프트] \n내용:” 형식으로 구성하였다. 모델에 전달한 최종 입력 프롬프트 목록은 <그림 6>에 제시한다.

빈도 수 기반 bigram 입력 프롬프트	
시	발라드
'노을 미소', '햇살 바람', '차가운 바람', '푸른 꿈', '달빛 춤', '노을 하늘', '달빛 바다', '차가운 미소', '아름다운 바람', '시린 바람', '다시 미소', '푸른 파도', '겨울 별', '노을 바다', '차가운 겨울', '아름다운 별', '달빛 노래', '바람의 춤', '노을 춤', '아침 물방울', '달빛 미소', '시린 물', '별빛 춤', '푸른 별', '푸른 숲', '당신의 미소', '작은 호수', '산 속에', '아름다운 물결', '아름다운 바다', '햇살 웃음', '아침 햇살', '별빛 하늘', '겨울 노래', '햇살 파도', '푸른 미소', '물방울 호수', '아침 미소', '겨울 파도', '아름다운 꿈', '차가운 별', '당신의 봄', '푸른 하늘', '별빛 노래', '다시 미소', '겨울 하늘', '작은 파도', '아름다운 호수', '가을 노래', '바다 노래'	'자꾸 그리워', '좋은 영원', '좋은 모습', '우리 모습', '매일 그리워', '마지막 눈물', '그냥 나와', '안녕 오늘', '혼자 이대로', '이대로 사랑', '모두 돌아와', '자꾸 미안', '나와 나와', '그대 안녕', '아주 그리워', '나에게 안녕', '어떤 사람', '이미 혼자', '이런 사랑', '작은 사랑', '다시 돌아와', '지금 돌아와', '그대 하나', '좋은 하루', '지금 우리', '사랑 때문에', '마지막 세상', '항상 이란', '우리 결에', '너무나 아픈', '아픈 사랑', '그래 돌아와', '매일 그대', '마지막 모습', '하루 멀리', '아주 멀리', '하지만 안돼', '이미 그리워', '오늘 나와', '제발 나와', '좋은 마지막', '항상 멀리', '매일 결에', '다시 결에', '이런 사람', '아픈 하루', '오늘 하루', '이런 눈물', '그래 안녕', '슬픈 사랑'
인디음악	힙합
'다시 눈물', '그게 우리', '좋은 사실', '그대로 이대로', '아직 처음', '다른 나에게', '그대 그대', '정말 혼자', '그대 그대', '결에 바람', '오늘 눈물', '그대로 오늘', '매일 그대', '아름다운 오늘', '사랑 처음', '아름다운 지금', '이대로 혼자', '그대로 결에', '모두 그대', '좋은 매일', '그냥 혼자', '그저 매일', '아직 좋아', '그저 언젠가', '사랑 매일', '어떤 눈물', '처음 혼자', '어떤 시간', '좋은 순간', '나에게 매일', '지금 시간', '매일 매일', '아름다운 우리', '항상 하나', '함께 그대', '항상 사랑', '지금 함께', '그래도 아파', '슬픈 눈물', '이런 영원', '우리 결에', '슬픈 그대', '매일 함께', '그저 언젠가', '아름다운 언젠가', '우리 영원', '언젠가 오늘', '슬픈 하루', '처음 같이', '아름다운 그대'	'모두 위로', '홀려 항상', '지금 우린', '계속 위로', '이미 사랑', '정말 몰라', '그때 위로', '그냥 믿어', '매일 만나', '위로 너도', '정말 멀리', '여전히 사랑', '너도 못해', '믿어 다시', '모두 사랑', '믿어 매일', '계속 멀리', '사랑 다시', '마음 위로', '이건 싫어', '위로 하지만', '몰라 미안', '언제나 만나', '제발 이젠', '자꾸 만나', '전부 항상', '다시 여기', '싫어 미안', '우린 언제나', '빨리 믿어', '다시 하나', '바로 오늘', '우린 계속', '홀려 자꾸', '다시 함께', '그냥 사랑', '홀려 다시', '가끔 만나', '몰라 아직', '사랑 몰라', '지금 살아', '함께 전부', '항상 위로', '사랑 미안', '다시 못해', '전부 위로', '살아 오늘', '만나 절대', '여전히 여기', '빨리 필요'

<그림 6> 빈도 수 기반 최종 bigram 입력 프롬프트

시 생성 결과 평가 및 분석

본 연구에서는 전체 시 데이터셋과 힙합, 발라드, 인디음악 장르별 가사 데이터를 각각 학습한 세 가지 모델의 결과를 비교·분석하였다. 입력 프롬프트는 두 가지 경우로 나누어 실험을 진행하였다.

1. 공통 입력 프롬프트 기반 생성

첫 번째 경우는 세 모델 (a), (b), (c)에 동일한 입력 프롬프트를 주었을 때의 결과이다. 생성된 bigram 입력 프롬프트 중 ‘달빛 미소’를 세 모델에 공통 입력으로 준 결과를 <그림 7>에 나타내었다. 생성된 시 결과에서 동일한 프롬프트를 입력하였음에도, 시의 문장 스타일부터 감정까지 서로 다르게 구성된 것을 확인하고 장르별 가사 데이터 학습의 영향을 실감할 수 있다.

입력 프롬프트(공통): “제목: 달빛 미소\n내용:”

<각 장르별 모델이 생성한 시>

아름다운 햇살처럼 빛나는 눈물과 사랑으로
이제야 깨달아가는 순간
내 가슴은 뜨겁게 **타오르고**
그대로 **멈춰 선**
달빛이 되어 다시 켜고 싶어하는
마음에서 우러나오는 따뜻한 숨결로
사랑에 대한 그리움이 있는 듯
삶의 기쁨과 행복, 그리고 삶의 아름다움을 느낄 수 있도록
어쩌면 이토록 아름답다는 것은
인생이라는 깊은 곳에 새겨져 있을지도 모르겠다.
한낮의 태양 속에 잠들어 있는
강물이 되고 싶지만
순수의 시간이 오도록
봄을 기다리는 꽃들의 **기도**

<훈련 시 데이터 중 유사도 Top1 시>

제목: **어쩌면 풀잎처럼**
순수의 뒷모습을 바라보았을 때
하얀 눈송이들이

보내온 천사의 한쪽 날개로 이루어진 마음에
그리움의 언덕 바람을 쌓을 수 있는 순결함이
이토록 아침 햇살이 곱게 비치는 곳이 있다면
순수 또한 녹아내리고 있음을
두 손으로 모아진 기도는
눈물방울처럼 기다리는 시간에도...

(a) 시 / 힙합 가사 훈련 모델

그대는 달빛을 받고 있다.
내 가슴은 비바람을 맞고 있는 것만 같다.
너를 사랑한 것은 마음이고 **사랑은 슬픔이다.**
달빛이 내리쬐는 밤하늘에 비친 달은
자꾸만 달빛의 빛이 빛난다.
아름다운 그대야
사랑하는 너의 마음을 알고
그대를 떠나보낸 건 나였나?
고독한 기다림이 남아 있으면
더 깊은 사랑이 되리라.
마음속 그리움을 잊으면
누구라도 행복해진다....
우리가 아름다운 날을 보내주길 빈다.
이토록 아름다운 날에
오래도록 우리 곁에 있으리라....

제목: **도라지 꽃**

휘어진 길모퉁이마다
그리움이 서려
희미한 달빛 아래 홀로 선 그대 모습
사랑의 흔적을 따라 나는 조용히 눈을 감네.
바람에 실려 오는 보랏빛 향기
그대와 추억이 가슴에 머물러
눈물에 젖은 밤하늘을 바라보며 사랑은 ...

(b) 시 / 발라드 가사 훈련 모델

그리운 밤하늘에 핀 별들이
별빛이 되어 내 가슴속에
바람이 되고 날갯짓을 하고 있습니다.
그렇게 나는 늘 나의 곁에 있어주고 싶어요
내가 언제나 나를 사랑해 주고 싶어요
마음속에 품고 있는 마음속의 따뜻한 사랑을
사랑으로 나누어가고 싶습니다
오늘도 행복하게 잘 살아보세요라고
내 가슴에 간직해주었으면 좋겠어요
달빛은 내게 가장 중요한 것
밤하늘에서 빛나는 것입니다
누구라도 다 가질 수 없는 작은 일들
아무도 가지 않은 곳에서
작은 나비가 날아오네요

제목: **아름다운 사람**

그대는 아름다운 사람
별을 사랑하고
하늘을 사랑하고
먼 곳 슬픔까지 사랑하는 그대는
아름다운 사람
아름다운 사람에게선 꽃 같은 향기가 난다
늘 곁에 있고 싶고
멀리 있으면 그리워진다
언제나 눈으로, 눈으로 사랑을 말...

(c) 시 / 인디음악 가사 훈련 모델

<그림 7> 공통 입력 프롬프트 생성 결과

힙합 가사를 학습한 모델 (a)의 생성 결과는 리듬감 있는 구절의 반복과 직설적인 어휘 사용이 두드러진다. “타오르고”, “멈춰 선”, “기도”와 같은 표현은 시적 감수성 위에 힙합 장르만의 강렬하고 직접적인 어감을 더한다. 이는 힙합 장르 특유의 직설적 서술 방식과 정서적 고조를 반영한다고 볼 수 있다. 또한 ‘삶의 기쁨과 행복, 그리고 삶의 아름다움’과 같은 구절을 통해 입력 프롬프트로부터 긍정적인 감정을 해석하여 시 내용을 통해 명확히 전달하려는 경향을 확인할 수 있다. 이는 힙합 가사에서 흔히 발견되는 자기 표현적 성향과 유사하다.

발라드 가사를 학습한 모델 (b)의 생성 결과는 서정성과 감정의 내적 흐름이 뚜렷하게 발견된다. “사랑은 슬픔이다”, “고독한 기다림”과 같은 구절은 발라드 가사에서 반복되는 이별과 회상의 정서를 잘 반영한다. 감정의 밀도와 정서적 서술이 길게 이어지는 방식, 온점의 반복된 사용(‘...’)은 발라드 장르 특유의 서사적 정체성을 더 극대화하며, 문장 전개 또한 비교적 부드럽고 유려하게 느껴진다. 이러한 정체성을 기반으로 해당 모델은 앞선 힙합 모델의 시 생성 결과와 달리 비교적 쓸쓸함과 그리움의 정서가 부각된다는 특성을 보인다.

인디음악 가사를 학습한 모델 (c)의 생성 결과는 일상적 소재와 감성적 은유가 조화롭게 결합해 있다. “작은 나비가 날아오네요”와 같이 비유적이면서도 서정적인 색채를 띠는 표현은 특정 장르적 클리셰에 갇히기보다 자유로운 감수성을 드러낸다. 특히 “누구라도 다 가질 수 없는 작은 일들”과 같은 문장의 사용은 인디 음악에서 자주 발견되는 소박하고도 독립적인 세계관을 반영한다.

또한 주목할 점은, 세 가지 모델 모두 동일한 전체 시 데이터셋에 대해 평가를 진행했음에도 하나의 입력 프롬프트(‘제목: 달빛 미소\n내용:’)로 생성된 시가 각각 서로 다른 시와 가장 높은 유사도를 기록했다는 사실이다. 이는 세 모델이 같은 입력 프롬프트를 받더라도 학습 과정에서 가사 데이터의 장르적 특성이 반영되어 모델 표현력에 차이를 만들었음을 의미한다. 이처럼 동일한 입력 조건에서 서로 다른 유사도 Top1 시를 뽑아낸 평가 결과는, 장르별 학습 데이터가 시 생성의 표현 방식에 유의미한 영향을 미친다는 것을 시사한다.

2. 장르별 입력 프롬프트 기반 생성

제목: 여전히 여기

멀리서 바라본 하늘은 그리움이 가득한
고요함 그 자체라고 생각합니다.
그렇게 홀로 남겨진 당신은
무언가를 꿈으로 간직하려고 하는 당신의 모습입니다.
그리움의 순간,
다시 한 번 더 돌아보세요.
당신의 마음이 떠나간 곳이면
아름다움 속에 들어있는
초록색의 아름다운 마음은
사랑하는 당신이 간직하고 싶은 사랑입니다.
그런 마음으로 당신을 바라보십시오.
누군가 당신에게 부탁하고 있는 것
그리고 이대로 떠난다면
이제는 내 마음속에 남아 있을 그대에게
당신에게는 이미 사라져버린 소중한 기억

(a)

제목: 슬픈 사랑

그대를 만나고 있는 순간
가슴이 떠나가는 길
살아갈 날도 그리 멀지않은데도
그리움을 그리워하는 나의 마음
눈물이 고이는 내 가슴엔
슬픔에 대한 애투함이 서려있다.
가난과 절망, 죄와 슬픔의 뿌리 깊은 곳에
진실과 사랑이 자리 잡고 있다.
우리가 마주해야할 이 세상에는
나름대로의 따뜻한 손길이 있을 것이다.
마음속의 아픔을 함께 나눌 수 있도록
사랑이라는 무거운 짐을 내려놓고
나의 외로움이 아닌 우리 곁으로 가자.
잊어버릴 때 잊혀지지 않도록

(b)

제목: 좋아 순간

나를 사랑했던 그 날
사랑을 나누며
오랜 시간 함께한 내 마음
그런 기억 속에
내 안의 슬픔이
아름다움과 나약함의 미련으로 남아
나쁜 것만 한단 말인가..
지난 봄은 그렇게 아름다웠는데도
봄의 꽃은 지고
다시 피는 꽃들의 노랫소리는
더욱 깊고 쓸쓸해져서
내가 사랑하는 그대에게 닿았을 때
나를 미워하지 않았던 그 날의
미련은 나에게 온 것 같고
나는 너를 그리워하며
함께 웃었던 그날
마음을 위로해주는 달콤한 미소
슬픔

(c)

<그림 8> 장르별 입력 프롬프트 생성 결과

두 번째 경우는 각 모델에 장르별 가사 데이터를 기반으로 구성된 경우다. 장르별 가사 데이터셋을 기준으로 TopK 단어를 추출하였으며, k=50으로 설정하였다. 추출 단어 리스트로부터 랜덤한 두 단어 조합을 50개 생성하여 장르별 bigram 입력 프롬프트를 마련하였다. 모델 (a)는 '여전히 여기', 모델 (b)는 '슬픈 사랑', 모델 (c)는 '좋아 순간'을 입력받았고, 그 결과는 <그림 8>에 나타내었다.

힙합 가사를 학습한 모델 (a)의 생성 결과는 강한 자기 정체성과 현실적 묘사를 드러낸다. "멀리서 바라본 하늘", "그런 마음으로 당신을 바라보십시오" 등의 어구는 현실적 장면 묘사와 자기 확신을 결합하여 청자에게 직접적으로 메시지를 전달하는 특징을 보인다. 또한 반복적 구절과 명령형 어투는 힙합 특유의 리듬 구조와 청중 지향적 특성을 반영한다.

발라드 가사를 학습한 모델 (b)는 사랑과 이별이라는 주제를 중심으로 감정의 점층적 전개를 보여준다. "눈물에 고이는 내 가슴", "가난과 절망, 죄와 슬픔"과 같은 표현은 발라드 가사의 전형적인 정서 구조를 따른다. 특히 감정을 고조시켜 결론에 도달하는 방식은 곡의 클라이맥스에 해당하는 발라드적 정서 전개를 충실히 담고 있다고 볼 수 있다.

인디음악 가사를 학습한 모델 (c)는 서정적이면서도 독창적인 감수성을 드러낸다. "다시 피는 꽃들의 노랫소리", "내 안의 슬픔이 아름다움과 나약함의 미련으로 남아"와 같은 표현은 기존 대중가요와 차별화된 정서를 보여주며, 인디음악 가사 특유의 개성적이고 실험적인 분위기를 반영한다. 또한 단순한 사랑·이별 감정의 구도에서 벗어나 삶의 순간을 독창적 시각에서 풀어내는 특성을 보인다.

모델별 시 생성 10회의 유사도 점수 비교 및 분석

시 생성 10회의 유사도 점수 평균			
	시+발라드	시+힙합	시+인디음악
sim_top1	0.803861	0.802584	0.801846
sim_top5_mean	0.784444	0.784929	0.783354
novelty	0.196139	0.197416	0.198054

sim_top1: 생성 시와 가장 유사한 훈련 시 1편과의 유사도
sim_top5_mean: 생성 시와 유사한 훈련 시 상위 5편 평균과의 유사도
novelty: 생성 시의 신생성(참신함)의 대략치

<그림 9> 모델별 Embedding 평균값

<그림 9>은 세 모델의 시 생성 결과 유사도 점수 및 참신성 수치를 비교한 것이다. 모델별 총 50개의 최종 입력 프롬프트를 10회 반복 실행하여 도출한 유사도 점수의 평균값을 계산하였다.

공통적으로 세 모델은 높은 유사도와 일정 수준의 참신성을 보여주며, 각 장르 데이터의 경향성 반영은 유사도 점수의 근소한 수치 차이로 확인할 수 있다.

발라드 모델은 `sim_top1`에서 가장 높은 수치를 기록하였다. 이는 발라드 가사 특유의 서정적이고 정형화된 표현 방식이 모델 학습 과정에서 강화되어, 새로운 시 생성을 할 때도 기존 시 데이터와 가장 가까운 유사도를 보이게 된 결과로 해석할 수 있다.

반면, 힙합 모델은 `novelty`에서 상대적으로 가장 높은 수치를 기록하였다. 이는 힙합 가사에 자주 등장하는 직설적이고 변주적인 표현 양식이 새로운 조합을 시도하도록 유도했음을 시사한다.

마지막으로, 인디 모델은 `sim_top5_mean`에서 가장 높은 수치를 보였다. 이는 인디 음악 가사가 지닌 특유의 자유로운 서술 방식과 동시에 일정한 정서적 흐름이 모델에 반영된 것으로, 가장 유사한 시 한 편에 집중되기보다 상위 다섯 편에 고르게 유사성을 분포시키는 경향성을 보인다.

세 모델의 절대적인 수치 차이는 크지 않으나, 발라드 모델은 안정성과 정형성, 힙합 모델은 참신성과 변주성, 인디음악 모델은 균형성과 다양성이라는 장르적 특성을 반영한 점이 잘 드러난다.

Conclusion

본 연구에서는 문학의 한 장르인 시에 노래 가사가 어떻게 접목될 수 있는지를 탐색하기 위해, 한국어 시와 장르별 가사 데이터를 함께 학습시킨 모델의 시 생성 결과를 비교해 보았다. 모델의 출력 성능을 보장하기 위해 사전 학습 언어모델 KoGPT-2를 기반으로 Prefix-Tuning 기법을 적용하여 발라드·힙합·인디음악 가사 데이터를 각각 학습시킨 세 개의 모델을 구축하였다. 학습한 모델의 출력 결과에 대해 sBERT를 통한 원본 시 데이터와의 유사도 비교를 진행하여 모델의 성능을 평가하였다. 또한 모델별 출력 결과에 대한 문학적 평가를 진행하여 장르별 특성이 시 생성 결과에 반영되는 양상을 분석하였다.

최종적으로 세 모델은 각각의 모델이 학습한 가사

데이터 기반의 입력 프롬프트뿐만 아니라, 동일한 입력 프롬프트에서도 학습한 장르 데이터의 특성이 뚜렷이 반영된 결과를 산출하였다. 힙합 모델은 직설적이고 리듬감 있는 표현을, 발라드 모델은 서정적이고 감정에 몰입된 흐름을, 인디음악 모델은 소박하고 개성적인 감수성을 중점적으로 시 생성 결과를 산출한다. 이는 곧 장르별 데이터 학습이 생성 모델의 시적 산출물에 직접적인 영향을 미친다는 사실을 내포한다. 따라서 본 실험은 장르별 데이터셋 학습이 생성 모델의 문체적 편향을 유도하며, 특정 장르적 감수성을 창작물에 반영하는 중요한 요소임을 시사한다.

다만 본 연구에도 한계는 존재한다. 본 연구에서 부족한 점들을 이야기하며 향후 연구가 발전되기를 기대한다.

첫째, 시간·환경적 제약으로 모델을 충분히 학습시키지 못했다. GPU 자원 한계로 학습 입력 한도(최대 토큰 수)를 낮춰야 했고, 그 결과 데이터셋의 상당 부분이 모델에 반영되지 못했다. 특히 힙합 장르는 곡당 평균 생성 토큰 수가 학습 입력 한도의 두 배를 넘어서는 경우가 많아 문맥 손실과 성능 저하의 가능성이 컸다.

둘째, Prefix-Tuning의 이점을 충분히 활용하지 못했다. Prefix-Tuning에서 Prefix는 MLP를 통해 생성되며 이 과정에서 MLP의 은닉층에 드롭아웃을 적용하여 과적합을 완화할 수 있으나, 본 연구의 학습 단계에서는 이를 적용하지 못했다.

셋째, 데이터 수집이 제한된 환경에서 이루어졌다. 가사는 다양한 주제·가수의 텍스트에 비용 부담 없이 접근할 수 있었지만, 시는 자체의 경제적 가치로 인해 폭넓은 접근이 어려웠다. 이에 따라 말뭉치 구성의 불균형이 발생했고, 모델의 일반화에 제약이 있었다.

향후 연구에서는 긴 문맥 학습을 지원하는 모델을 구축하기 위한 자원을 확보하고, 정식 라이선스를 통한 데이터 확충과 학습 최적화를 병행하여 보다 안정적이고 일반화된 성능을 확보할 필요가 있다.

본 연구의 의의는 크게 세 가지로 살펴볼 수 있다. 첫째, 한국어를 대상으로 한 시 생성 연구라는 점에서 기존 영어 중심 연구와 차별성을 가진다. 둘째, 시와 가사의 경계를 가로지르는 새로운 시도로서, 장르별 가사 데이터가 시 창작의 양식을 어떻게 변주하는지를 실험적으로 검증하였다. 셋째, sBERT 기반 유사도 비교를 통한 정량적 평가와 정성적 문학 분석을

결합하여, 인공지능 창작물을 다각적으로 해석할 수 있는 분석 틀을 제시하였다. 이를 바탕으로 인공지능을 통한 문학 창작이 더욱 활발해지게 되길 바란다.

References

고정민 (2024). "플랫폼 형태에 따른 음악산업의 변화 연구". 문화산업연구, 24(1), 1-4.

국주현 (2023). "SeqGAN을 활용한 멜로디 기반 가사 생성", 한국정보과학회, 265-267.

김명선 외 (2023). "인공지능 기반 한국어 시 생성 시스템 개발 연구". 경영정보학연구, 25(3), 1-15.

박성준 (2023). "ChatGPT를 활용한 AI 시인 구현(1)-'AI 윤동주'의 생성과 시 창작의 가능성을 중심으로", 세명대학교 인문사회과학연구소 공동 학술대회, 제58호, 17-29.

박찬술 (2021). "한국어 랩 가사 생성을 위한 운율 단어 임베딩과 어텐션", 한국통신학회, 46(12), 10-32.

임수진 (2018). "대중가요 매체를 활용한 한국어 시 쓰기 활동의 교육적 효과". 국어교육연구, 제42집, 16-40.

정용재 외 (2023). "한국 근현대시의 단어 빈도 분석 및 인공지능을 활용한 시 생성 및 분류 기법", 한국통신학회 하계종합학술발표회, 354-355.

최광희 외 (2021). "딥러닝 기반 언어 모델들을 활용한 라임 기반 한국어 랩 가사 생성 시스템", 한국정보과학회 학술발표논문집, 1339-1341.

Xiang Lisa Li 외 (2021). "Prefix-Tuning: Optimizing Continuous Prompts for Generation", Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 4582-4597.

Reimers, N. 외 (2019). "Sentence-BERT: Sentence embeddings using Siamese BERT-networks", arXiv preprint arXiv:1908.10084.