

내재된 편향성과 차별을 완화한 한국어 인공지능 동화 생성기 구축

안 지 인^{1*} · 이 육 섯 별^{1*} · 장 예 원^{1*} · 정 다 샘^{2*}^{1*} 서강대학교 아트앤테크놀로지학과 학사과정 ^{2*} 서강대학교 아트앤테크놀로지학과 조교수

Construction of Debiased Korean AI Fairytale Generator

Jiin An^{1*} · Saetbyeol Leeyouk^{1*} · Yewon Jang^{1*} · Dasaem Jeong^{2*}^{1*} Bachelor's Course, Department of Art&Technology, Sogang University, Seoul 04107, Korea^{2*} Assistant Professor, Department of Art&Technology, Sogang University, Seoul 04107, Korea

[요 약]

<프랭클린>은 편향성 완화 과정을 거친 언어 생성 모델을 통해 AI 모델과 동화를 공동 창작할 수 있도록 개발된 인터랙티브 서비스이다. 기존 동화들이 사회의 고정관념을 포함하고 오래된 사회상을 반영한 점, 그리고 한국어 언어 모델이 사회의 차별과 혐오를 그대로 반영한 데이터셋으로 학습된 점을 고려하여 비윤리적이고 편향적인 표현의 생성을 방지하는 개발 과정을 거쳤다. 한국어 동화 데이터셋을 구축하는 과정에서 명시적인 혐오 표현과 욕설을 삭제하고, 일부 고전 동화가 포함한 잔인하거나 비도덕적인 부분을 탐지하여 불필요한 데이터를 제거하며, 문법적인 오류를 최소화하여 크기가 작은 학습용 데이터셋의 효용성을 높였다. 언어 생성 모델이 동화의 어투를 구현할 수 있도록, 구축한 데이터셋을 토대로 모델을 미세 조정하였으며 모델의 단어 임베딩 값을 조정하여 이에 내재된 젠더 편향성을 완화하고자 하였다. 인공지능이 스토리를 가진 글을 생성할 수 있도록 맥락성을 보완하는 방식에 대한 실험을 진행하였으며, 실험 결과를 토대로 모델을 미세 조정하였다. 아이들이 학습된 모델을 이용하여 인공지능과 동화를 공동 창작할 수 있도록 웹서비스를 디자인하였으며, 동화 생성 파이프라인을 구축하였다.

[Abstract]

<Franklin> is an interactive service that enables children to become the protagonists and authors of fairy tales in collaboration with an AI language generation model. Considering the inclusion of social prejudices and outdated social norms in existing fairy tales and the fact that Korean language models are trained on datasets that reflect discrimination and hatred in society, <Franklin> was designed to prevent the generation of unethical and biased expressions. During the process of building a Korean fairy tale dataset, explicit expressions of hatred and profanity were deleted, unnecessary data were removed by detecting cruel or unethical parts included in some classic fairy tales, and grammatical errors were minimized to increase the effectiveness of the small training dataset. Based on the constructed dataset, the model was fine-tuned to implement the tone of the fairy tales, and the word embedding values of the model were adjusted to mitigate the inherent gender bias. Experiments were conducted to complement contextuality and generate coherent stories, and the model was fine-tuned based on the results. A web service was designed to enable children to create fairy tales using the trained model, which serves as a platform for providing fairy tales that reflect more diverse social norms to children.

색인어 : 인공지능, 동화, 편향성 완화, 인공지능 윤리, 공동 창작, 한국어 생성 모델**Keyword** : Artificial Intelligence, Fairytale, Debiasing, AI Ethics, Co-Creation, Korean Generative Model<http://dx.doi.org/10.9728/dcs.2023.24.7.1513>

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 31 May 2023; Revised 06 July 2023

Accepted 10 July 2023

* These authors contributed equally to this work
*Corresponding Author; Dasaem JeongTel: E-mail: dasaemj@sogang.ac.kr

I. 서론

아이들이 동화를 접하는 형식은 유튜브 등으로 미래화되었지만, 그 내용에는 여전히 편향적인 클리셰가 가득하며 편견이 남아있다. 동화를 통해 세상을 배우는 아이들이 세상의 편견을 재학습하지 않도록, 변화하는 세상의 다양성을 배우고 더 긍정적이고 다양한 사고를 할 수 있는 콘텐츠를 제공하는 것이 본 연구의 목표이다. 본 프로젝트를 통해 한국어 동화를 작성하는 모델의 성능을 발전시키는 동시에, 인간과의 상호작용 과정에서 인간에게 선한 영향력을 줄 수 있는 인공지능이 무엇인지 실험한다.

<프랭클린>은 궁극적으로 인공지능 윤리에 대한 연구를 통해 인공지능에 내재된 편향과 혐오 표현을 제거하는 기술적 목표를 가진다. 언어 모델의 자체적인 성능 향상에만 집중한 결과, 대다수의 언어 생성 모델들은 사회에 내재된 편향과 차별을 그대로 반영하고 있다. 실제로 현재 뛰어난 성능을 보이는 카카오브레인의 한국어 언어 모델 KoGPT 공식 웹사이트[1]에는, 이러한 언어 모델의 위험성이 직접적으로 언급되어 있다. <프랭클린>을 개발하는 과정 속에서 동화를 써내려가는 모델의 성능을 발전시키는 동시에, ‘올바른’ 인공지능이란 무엇인지에 대한 깊은 고민을 이어간다. 문제적인 텍스트를 포함하지 않은 동화 데이터셋을 구축하고, 언어 모델에 내재된 차별을 분석하여 그 정도를 낮추는 방안을 강구하고자 한다. 본 연구의 주요 기여는 아래와 같다.

- 한국어 동화 텍스트를 수집하고 정제하여 프롬프트에 따라 동화를 생성하기 위한 데이터셋을 완성하였다.
- 차별적인 언어 생성을 줄이기 위한 탈편향 방식으로 단어 임베딩을 중립화하는 방법을 적용하였다.
- 이야기의 문맥을 보완하기 위하여 프롬프트와 키워드를 연결짓는 방식을 도입하였다.
- 아이와 인공지능이 동화를 공동 창작할 수 있는 웹서비스를 개발하였다.

II. 관련 연구

2-1 동화에 담긴 고정관념과 편향성

한국 전래 동화는 직업과 관련한 여성과 남성에게 대한 차별적인 묘사, 여성의 외모를 강조하는 표현, 남녀의 인성과 행동 특성에 대한 차별적인 묘사를 담고 있으며, 남녀 주인공을 통하여 주제를 차별적으로 나타내는 경향[2]을 가지고 있다. 독자들의 성인지 감수성은 높아진 현대의 상황에도 불구하고, 창작 동화들은 여전히 성차별적인 표현들을 포함하고 있다. 김지은[3]에 따르면, 아동문학 속의 성차별은 대부분 적대적 성차별이 아닌 호의적 성차별[3]이며, 이들은 차별인지 아닌지 구분하기 어렵다는 특성을 지닌다. 어린이 독자들은 동화

속의 농담이 누군가에게 불편할 수도 있는 이야기라는 것을 알아차리기 힘든 문화 안에서 자라고 있다[3].

2-2 인공지능 모델의 탈편향 연구

데이터 혹은 알고리즘에 내재된 편견은 인공지능 모델이 예측하는 결과의 편향성을 초래한다[4]. 이처럼 편향된 데이터와 알고리즘을 토대로 학습한 인공지능 모델은 인간 사회가 지닌 편견을 그대로 혹은 확대하여 반영하는 결과를 야기한다[5]. 이러한 문제를 해결하기 위해 선행 연구에서는 남녀 성별을 반대로 바꾼 데이터셋을 추가로 만드는 데이터 증강을 통해 모델의 젠더 편견을 완화하고[6], 주어에 성별 태그를 붙이는 방법을 통해 번역 과정에서 모델이 자의적으로 성별을 왜곡하지 않도록 방지[7]하였다. 산업 분야에서는 IBM사가 AI Fairness 360 (AIF360) toolkit[8]을 출시하여 학생들, 연구자들 그리고 산업계에 종사하는 사람들이 사회의 전반적인 편견을 줄이는 알고리즘을 사용하여 공정한 머신러닝 개발이 용이하도록 하였다.

2-3 인공지능 창작

미디어아트 그룹 슬릿스코프는 카카오브레인의 KoGPT (Korean Generative Pre-Trained Transformer) 기반 AI 모델 ‘시아’를 개발하였다. 슬릿스코프는 KoGPT 모델에 시작법을 학습시켜 다양한 시제의 작품을 생성하였다. 또한 대학로예술극장 소극장에서 모델 ‘시아’와 함께 인공지능 시극 <파포스>[9]를 개최하여 인공지능과 사람이 함께 창작하는 예술 작품을 선보였다. Bensaïd 등[10]은 Reddit의 WritingPrompts 데이터셋을 이용해 GPT-2 모델을 미세조정하여 이야기 생성 모델 FairyTailor을 개발하였다. FairyTailor는 사용자의 프롬프트에 따라 짧은 동화와 이어 어울리는 삽화를 생성한다. Chung 등[11]은 GPT-3를 이용한 이야기 생성 모델 TaleBrush를 개발했다. TaleBrush는 사용자가 입력한 주인공의 이름과 이야기를 도입부에 따라 주인공의 서사적 감정 굴곡선(fortune line)을 그려내고 이에 맞는 이야기를 생성한다.

III. 한국어 동화 생성 모델 파이프라인 구축

2-1 한국어 동화 데이터셋 구축

1) 데이터셋 수집 기준 및 방법

데이터셋을 수집한 기준은 1) 한국어로 구성된 이야기로 전래 동화와 국내 창작 동화 그리고 외국 동화 번역본 등이 포함된다. 그리고 2) 무료로 접근 가능한 동화 자료, 3) 무료로 공개되지 않은 경우 작가 및 출판사에게 인공지능 학습 용도로 허가 받은 동화 자료이다.

표 1. UnSmile 베이스라인 모델을 통해 ‘clean’하지 않다고 평가한 동화 예시

Table 1. Examples of fairytales not evaluated as ‘clean’ by UnSmile baseline model

*Since we used Korean-based generative model, all of the stories are written in Korean

The most highly scored category	Score	Story
악플/욕설	0.39	제목 : 무엇이든 삼켜버리는 마법상자 시끄러운 소리에 아침부터 짜증이 났어. 귀찮게 매달려서 그런 건데 엄마는 또 나만 혼냈어. 공부시간은 재수없게 나만 걸러서 벌 받고, 돌아오는 길엔 개한테 물리고 모든 게 싫었어. 그때 우연히 마법상자를 주웠어. 마법 상자! 이 상자는 당신이 싫어하는 건 무엇이든 삼켜버립니다. 진짜일까? 또 생선이야? 생선 먹기 싫다니까. 옆집에서 나는 시끄러운 소리도 싫어. 나만 혼내는 선생님이랑 날 우습게 보는 친구들도 싫어. 이 상자 굉장해! 내가 싫어하는 건 진짜 뭐든지 삼켜 버리네. 이씨 또 난 동생이 싫어. 동생 편만 들고. 엄마도 싫어. 이제 난 자유다! 아무도 날 방해하지 못해. 텔레비전이 나 싫것 봐야지. 기분이 이상해. 그래. 뭘 좀 먹으면 나올거야. 모두 돌려줘. 다 다 나 때문이야. 난 내가 싫어. 미안해.
여성/가족	0.37	제목 : 소녀와 우유통 젖소 농장에서 우유를 짜는 소녀가 있었어요. 소녀는 이런저런 상상하기를 좋아했어요. "저 새는 오늘 꽃놀이 가나 봐. 기분이 아주 좋아 보이네. 시냇물이 조르르 졸졸 춤추는 것 같아. 나도 빙그르르 빙글 춤추고 싶다. 참, 우유가 신선할 때 어서 시장에 내다 팔아야 하는데." (종락) 모두 나랑 춤 추려고 할 거야." "오! 아름다운 아가씨. 나와 함께 춤추시겠어요?" "흥! 내가 그렇게 쉽게 허락할 줄 알아?" "싫어요. 그럴 수 없어요." 이때었어요. "아, 어떡해. 내 우유!" 우유를 모두 쏟은 소녀는 슬퍼하며 농장으로 돌아갔답니다.
악플/욕설	0.32	제목 : 개와 돼지 옛날 옛적에, 한 할머니가 개와 돼지를 기르셨습니다. 할머니는 개와 돼지를 무척 귀여워하셨습니다. 할머니가 개의 머리를 쓰다듬으며 말씀하셨습니다. "참 착하구나. 집도 잘 지키고..." 할머니는 낮잠만 자는 돼지도 좋아하셨습니다. "돼지야, 아프지 말고 무럭무럭 자라라." 그러나 낮잠을 자고 있던 돼지는 할머니의 말씀을 듣지 못하셨습니다. 돼지는 할머니가 개만 귀여워한다고 생각하셨습니다. (종락) 내일 아침 날이 밝으면 장에 내다 팔아야겠다.' 이튿날, 할머니가 돼지를 장에 내다 파셨습니다. 돼지는 할머니의 사랑을 더 받으려다 그만 팔려가고 말았습니다.

저작권-프리 동화가 게시된 온라인 사이트에서 이야기 파일을 다운로드하였으며, PDF 혹은 책 이미지로 구성된 파일들은 광학문자인식(Optical Character Recognition, OCR)을 이용하여 변환 작업을 진행하였다. 이야기가 적힌 페이지를 휴대전화와 태블릿 등의 전자 기기로 직접 촬영하였고, 애플 사에서 제공하는 라이브 텍스트 기능을 이용하여 이미지 속의 텍스트를 추출하였다. 텍스트 추출 과정에서 생긴 문법적 오탈자들은 저자 3인이 직접 교정하여 엑셀 파일에 정리하였다. 텍스트 파일로 다운받은 자료들은 원본 파일에서 특수문자와 특수기호를 삭제하는 과정을 거쳤다.

이와 더불어 청와대 어린이 동화 사이트[12]와 유튜브에 게시된 동화 영상들의 자막을 크롤링하였다. 청와대 동화 사이트에서 크롤링한 텍스트들은 한 문단의 이야기 소개 문구와 2~3문단의 구연동화, 그리고 한 문단의 감상으로 이루어져 있었으며 편집자의 주관적인 감상이 적힌 첫번째 문단과 마지막 문단을 삭제한 후 데이터셋에 포함하였다. 유튜브 자막 크롤링은 아이들을 타겟으로 한 동화 영상들 중 한국어 자막이 포함된 자료들을 토대로 진행하였다. 유튜브 자동 생성 기능으로 작성된 자막의 경우에 대해서는 저자 3인이 전수조사를 진행하여 전반적으로 품질이 낮은 자료들을 우선 삭제하였고, 영상 속에 자막이 이미 작성되어 있는 자료들과 함께 문법적 교정을 거친 후 사용하였다. 위 과정에서 자동 생성 자막의 품질이 좋지 않았던 동화들에 대해서는 각 영상에서 음성 파일을 추출하였으며, 네이버 클로바 노트를 사용하여 음성을 텍스트로 변환하는 작업을 1차로 진행한 후 저자 외 3인이 음성 자료에 맞도록 텍스트 변환본을 교정하는 과정을 거쳤다. 교정 결과로 전달받은 자료들을 저자 3인이 최종 검수하여 데이터셋에 포함하였다. 현대 동화 혹은 소설을 확보

하기 위해 동화 작가 및 출판사에 직접 연락하여 전달 받은 작품들도 데이터셋에 추가하였다.

위 방법을 통해 확보한 동화들을 단편 동화의 경우에는 각 동화마다, 장편 동화의 경우에는 각 챕터를 한 편으로 간주하여 저장하였다. 이러한 작업을 통해 원시 데이터로 동화 자료를 총 1600편 수집하였다.

2) 데이터셋 정제 기준 및 방법

1차 정제는 스마일게이트 사에서 공개한 한국어 혐오표현 데이터셋 UnSmile[13]을 토대로 진행하였다. UnSmile 데이터셋은 한국의 주요 온라인 포털 사이트들의 뉴스 댓글과 온라인 커뮤니티들에서 수집된 온라인 댓글들로 구성된다. 전문가 집단을 통해 멀티-레이블로 레이블링된 총 18742개의 문장들로 이루어져 있으며, 혐오표현 10139 문장, 악플/욕설 3929 문장, clean 4674 문장으로 구성되어있다[14]. 본 데이터셋에서는 혐오 표현을 “특정 사회적 (소수자) 집단에 대한 적대적 발언, 조롱, 회화화, 편견을 재생산하는 표현”으로 정의하며, 혐오 표현의 카테고리를 ‘여성/가족’, ‘남성’, ‘성소수자’, ‘인종/국적’, ‘연령’, ‘지역’, ‘종교’로 세분화하였다. 스마일게이트 사에서는 UnSmile 데이터셋을 기반으로 텍스트 자료에 포함된 혐오와 차별적인 표현을 탐지하는 모델(이하 베이스라인 모델이라 칭한다)을 제공한다. 위 베이스라인 모델은 UnSmile 데이터셋을 15005개의 training 샘플과 3737개의 test 샘플로 나누어 학습하였으며 9개의 혐오 표현 세부 카테고리에 대한 점수 연산을 수행하도록 설계되었다. 베이스라인 모델의 최대 수용 길이에 맞추어 동화 텍스트를 최대 길이 600자가 되도록 마침표를 기준으로 끊어내는 작업을 진행한 후, 각 텍스트 부분에 대하여 UnSmile 결과값을

표 2. 명시적으로 부적절한 표현을 포함하지는 않지만 전체적인 내용이 부정적인 동화 예시

Table 2. Examples of fairytales do not include explicitly inappropriate expressions but overall not educable

*Since we used Korean-based generative model, all of the stories are written in Korean

Score	Story
0.39	강아지 똥 돌리네 흰둥이가 똥을 넣어요. 골목길 담 밑 구석 쪽이었죠. 흰둥이는 조그만 강아지니까 강아지똥이었던답니다. 날아가던 참 새 한 마리가 보더니 강아지 똥 곁에 내려앉아 콧코 째면서 말했어요. "내가 똥이라고 더럽다고?" 강아지 똥은 화도 나고 서러워서 눈물이 나왔답니다. (중략) "가여워라" 그래서 이렇게 벌을 받아 달구지에 실러오다가 떨어진 거야 난 이제 끝장이야." 그때 저쪽에서 소달구지가 덜컹거리며 오더니 갑자기 멈춰버립니다. 아니 이젠 우리 발 흠이잖아 어저 갖고 오다가 떨어뜨린 모양이군 도로 밭에다 갖다 놓아야겠다.
0.37	"사자 똥이 똥! 사자가 코뿔소를 찾아왔어요. "의사 할아버지, 배가 아파요. 누가 콧코 찌르는 것만 같아요." "배가 왜 아픈지 모르겠는 걸. 아무래도 똥 검사를 해봐야겠다. 이 봉지에 네 똥을 담아 오너라." "어찌지 지금은 똥이 안 마려운데." 그때 철썩철썩 소리가 들려왔어요. "어, 코끼리가 시원하게 똥을 누고 있네! 울지. 저 똥을 내 똥 대신 가져가야겠다. 할아버지 똥 가져왔어요." 코뿔소는 입을 씹룩거렸어요. (중략) 코뿔소는 사자 똥을 이리저리 뒤져 살펴봤답니다. 사자는 이마 살을 찌푸렸죠. "어휴, 냄새 한번 고약하네. 할아버지 너 무 가까이 얼굴을 대지 마세요. 똥 묻겠어요." "으이그, 네 똥 속에 벌레가 우글우글하구나. 이 약을 먹으면 배가 안 아플 거야." 사자는 머리를 긁적거리며 약을 한입에 꿀꺽 삼켰답니다. 이제 배가 안 아프겠죠
0.32	난 잠만 잤는걸 숲 속에 심술꾸러기 곰이 살았어요. 곰은 틈만 나면 다른 동물들을 괴롭혔죠. "흐헤헤, 재밌다." "아우, 여우 살려!" 그리고 겨울이 왔답니다. 동물들은 모두 겨울잠을 자러 갔어요. 곰도 겨울잠에 빠져들었죠. 그리고 시간이 지나 드디어 따뜻한 봄이 왔어요. "아함, 잘 잤다." 곰은 기지개를 켜면서 밖으로 나왔죠. '곰은 오줌싸개. 곰은 엉덩이에 뽀났다.' 누가 벽에다가 곰의 흥을 잔뜩 써 놓았어요. (중략) 그때 연못 속에서 남생이가 기어 나왔어요. "맞아, 남생이일 거야. 내가 거꾸로 뒤집어 놓은 적이 있으니까." 곰이 남생이에게 달려갔어요. "야, 남생이 너. 내가 벽에다가 네 욕을 썼지?" 곰은 눈을 부릅 뜨고 말했어요. "어어, 나 금방 일어났어. 겨우네 물속에서 잠만 잤는 걸." 남생이가 느릿느릿 말했어요.

측정하였다. 이 모델을 사용하여 저자가 확보한 동화 속에 ‘여성/가족’, ‘남성’, ‘성소수자’, ‘인종/국적’, ‘연령’, ‘지역’, ‘종교’, ‘악플/욕설’에 해당하는 부적절한 표현들이 얼마나 포함되어 있는지에 대한 점수를 0부터 1사이의 값으로 수치화하였다. 동화 텍스트를 분리한 개수만큼 UnSmile 지표 점수를 합산하고 평균을 내어 전체 동화에 대한 카테고리별 점수를 도출할 수 있었다. 카테고리별 점수를 비교한 후 가장 높은 점수를 받은 카테고리가 동화의 UnSmile 지표가 되었으며 혐오/차별 표현이 없는 ‘clean’ 동화 데이터셋과 ‘clean’하지 않은 동화 데이터셋으로 나눌 수 있었다. 이러한 결과값을 토대로 부정적인 항목의 점수가 0.3점 이상인 이야기들을 저자 3인이 직접 검토하였고, 0.32점 이상을 받은 이야기들을 혐오 및 차별 표현을 가진 동화로 간주하여 제거하였다. 위 표 1은 UnSmile 베이스라인 모델을 통해 ‘clean’하지 않다고 평가한 동화들의 예시이다.

일반적으로 아이들이 읽는 동화는 밝고 긍정적인 내용을 담고 있을 것이라 생각하는 경향이 있다. 하지만, UnSmile 베이스라인 모델을 사용하여 동화 데이터셋을 분석한 결과, 저자가 수집한 원시 한국어 동화 문치에 차별적 표현, 적대적 표현 이외에도 폭력적이거나 비방적인 표현들 그리고 편견을 재생산하는 표현이 포함되어 있었음을 파악할 수 있었다. 본 연구에서 추구하는 동화 데이터셋을 완성하기 위해 1차 정제를 완료한 데이터셋을 대상으로 추가적인 정제 작업을 진행하였다.

2차 정제 단계에서는 감정 분석 모델을 활용하였다. 위 모델은 전체 20만개의 리뷰로 구성된 네이버 영화 리뷰 데이터 [15]로 학습된 CNN classifier 모델이다. 혐오/차별적 표현 이외에도 잔인하거나 부정적인 내용이 과도한 부분을 차지하는 동화를 제거하기 위해 영화 리뷰를 바탕으로 긍정/부정을 분류하는 모델을 사용하였다. 위 분류 모델은, 긍정적인 리뷰

일수록 1에 가까운 점수를, 부정적인 리뷰일수록 0에 가까운 점수를 부여하며, 0.2점 이하의 점수를 받은 텍스트를 negative_story라는 이름의 리스트에 저장한다. 감정 분석 모델이 읽어들이는 문장의 최대 길이를 1000으로 설정하여, 긴 동화들을 1000자 이하로 분할하는 작업을 진행하였다. 이렇게 나누어진 동화들을 토대로 도출한 감정 분석 결과에 대하여 저자 3인이 negative_story에 포함된 이야기를 모두 확인한 후, 0.15점을 기준으로 정하여 이 이하의 점수를 받은 셀들을 데이터셋에서 제거하였다. 여러 개의 셀로 분할된 동화의 경우, 하나의 셀에서라도 감정 분석 결과가 0.15 이하로 산출되면 그 동화에 해당하는 셀들을 모두 삭제하였다.

1차 정제 과정을 거친 이후에도 표 2의 동화들과 같이 명시적으로 부적절한 표현을 포함하지는 않지만 전체적인 내용이 아이들에게 교육적이거나 긍정적인 영향만을 준다고 판단하기 어려운 이야기들이 데이터셋 안에 포함되어 있었다. 어떤 분류 모델을 사용하더라도 동화에 은유적으로 표현되고 내재되어있는 고정관념들을 탐지하기는 어려웠다. 데이터셋 구축 및 정제 단계에서는 여러 종류의 분류 모델을 통해 명시적인 혐오 표현 혹은 욕설, 잔인한 표현들을 걸러낼 수 있지만, 맥락 속에 숨어있는 표현까지 모델이 포착해내기는 쉽지 않다. 본 데이터셋 구축 이후 저자는 데이터셋 이외의 부분에서도 편향성 완화를 위한 기법을 도입해야 한다는 결론을 내리게 되었으며, 그 내용은 이후 IV. 차별적인 언어 생성을 줄이기 위한 탈편향 방식의 도입에서 이어진다.

3) 데이터셋 통계

본 연구에서 구축한 한국어 동화 정제 데이터셋은 총 1197개의 동화로 이루어져 있다. 작품의 평균 토큰 수는 593개이며, 가장 긴 작품의 토큰 수는 8632개, 가장 짧은 작품의 토큰 수는 27개이다. 본 데이터셋에는 각 작품에 대하여 제목과

표 3. 구축한 한국어 동화 정제 데이터셋의 일부

Table 3. A part of the constructed, refined Korean fairytale dataset

*Since we used Korean-based generative model, all of the stories are written in Korean

Title	Story	Keywords
크리스마스 전날 밤	크리스마스 전날 밤이에요. 우리 가족들은 모두 잠들었어요. 오늘 밤엔 산타 할아버지가 선물을 주러 오신대요 동생은 꿈인형을 달라고 조르는 편지를 썼어요. 나는요? 비밀이에요. 동생은 설탕 과자와 춤추는 꿈을 가나 봐요. 빨리 잠이 들어야 산타 할아버지가 오실 텐데... 어찌죠? 잠이 오지 않아요. 아, 맞다! 엄마가 눈을 감고 양을 세면 잠이 온다고 했어요. 한 마리 두 마리 세 마리 "어 무슨 소리지?" "루돌프야! 지붕 위로 내려가자." 친구들, 지금 이건 꿈일까요? "어떡하지? 누가 머리 좀 지붕에 있나 봐요." "아니, 이게 왜 안 들어가?" "거기 누구 있어요?" "누구세요?" "안녕 꼬마야?" 별처럼 빛나는 눈, 장밋빛 뿔, 박살 그리고 눈처럼 하얀 수염, 빨간 모자와 털옷. "산타 할아버지!" "꼬마여! 나를 만난 건 우리 둘만의 비밀이란다. 지킬 수 있지?" 산타 할아버지는 익숙한 손길로 양말에 선물을 넣기 시작했어요. 산타 할아버지의 불룩한 배는 움직일 때마다 울 풍선처럼 나는 그 모습을 보고 웃음을 참느라 혼이 났어요. "어허 이걸 네 거란다. 네가 원하던 선물일 거야. 아침에 풀어보렴." "날 기다리는 아이들이 많아서 이만 가봐야겠구나." 산타 할아버지는 다시 굴뚝을 타고 올라가셨어요. "그 모습이 힘들어 보이는 건 나쁜인가요?" "자, 빨리 가지꾸나." "이래! 달빛 속에서 썰매가 훑 날아왔어요. "우와! 안녕히 가세요.산타 할아버지!" "모두, 메리 크리스마스!".	['산타', '할아버지', '눈', '크리스마스', '밤']
나는 내가 좋아	내 이름은 김아영. 새로 산 원피스! 오늘 유치원에 입고 가면 친구들이 뭐라고 할까? 그런데 친구들은 나를 보자마자 "김하영 공주 옷 입었구나! 돼지 공주!" "돼지 공주!" "난 돼지가 아니라 그냥 통통한 건데." 선생님이 짝꿍을 칭찬해요. "어머 정말 잘 그랬다. 그림도 잘 그리네." 나는 슬그머니 내 그림을 가렸어요. 나도 열심히 그렸는데 엉망이야. 엄마에게 주스를 주려고 했는데 그만 잊지르고 말았어요. "아휴, 또 말썽이야" 난 왜 잘하는 게 없을까 맨날 실수만 하고... 거울 속에 내 모습은 엉망. 내가 콩알만큼 작아진 것 같아요. 아무도 없는 곳에 꼭 숨고 싶어요. 그런데... 무슨 소리지? 환한 불빛 속에 엄마 배가 보여요. "우리 아기 어떻게 생겼을까?" "아가야. 너를 만나려고 꼬박 열 달을 기다렸단다." 다정한 엄마 아빠의 목소리. 아기가 엄마 품에 꼭 안겨 새근새근 잠들었어요. 엄마가 아기를 보며 웃어요. 아기가 작은 다리도 몸을 일으켰어요. 겨우 몇 발자국 걸었을 뿐인데 엄마 아빠 손뼉을 치며 기뻐해요. 부럽다. 아기가 아픈가 봐요. "우리 아기 얼른 낫게 해주세요. 아가야 아프지 마." 엄마가 기도해요. "아가야, 태어나줘서 고마워." "건강하게 자라줘서 고마워." 모두 축하해줘요. 나도 아이처럼 사랑받고 싶어요. 그때 아기가 고개를 돌렸어요. 어머니! 그 아이는 바로 나였어요 그 아이가 바로 나라니. 나는 보잘것없는 아이가 아니에요. 엄마 아빠의 귀한 보물. 세상에 하나밖에 없는 소중한 아이. 내가 달리기에서 넘어져서 골짜지를 했지만 그래도 괜찮아. 울지 않고 끝까지 뛰었거든. 난 내가 자랑스러워.	['엄마', '공주', '돼지', '아빠', '친구']

동화 본문, 그리고 동화의 키워드 5개가 포함되어 있다. 작품의 제목은 데이터셋 수집 시 동화 본문과 함께 저장한 자료이며, 동화의 키워드는 카카오브레인 사에서 공개한 통합 자연어처리 프레임워크 ‘PORORO(Platform Of neuRAL mOdels for natuRAL language prOcessing)’[16]를 사용하여 구성하였다. PORORO에서 지원하는 Sequence Tagging task로 개체명 인식을 수행하여 명사를 추출하고 가장 많이 등장한 상위 5개의 단어를 키워드로 간주하였다. 데이터셋에 키워드를 추가한 이유는 본 파이프라인 내의 언어 생성 모델이 보다 맥락성 있는 이야기를 생성하도록 하기 위함이며, 이는 V. 이야기의 문맥을 보완하기 위한 방법 연구 부분에 상세히 기술되어 있다. 이러한 방식으로 구성된 동화 데이터셋은 추후 한국어 언어 모델의 학습(미세조정) 데이터셋으로 활용되었다.

3-2 미세조정

1) 동화 어투 구현

본 파이프라인에서는 SK 텔레콤 사에서 huggingface에 공개한 kogpt trinity 모델[17]을 사용하였다. trinity를 포함한 한국어 언어 모델은 한국어 위키 백과, 모두의 말뭉치 그리고 청와대 국민청원 등 문어체로 쓰인 자료가 대부분을 차지하는 데이터셋으로 사전 학습되었다[18]-[20]. 이러한 이유로 한국어 언어 모델이 아이들이 읽는 동화의 문체를 따라 글을 작성하기 위해서는 한국어 동화에 대한 추가적인 학습 과정이 필요했다.

이를 위해 본 연구에서는 3.1 과정을 통해 구축한 데이터

셋을 바탕으로 제목, 키워드 5개 그리고 동화 본문을 하나의 프롬프트로 연결하여 한국어 언어 생성 모델을 미세조정하였다. 사용된 데이터셋의 일부는 표 3과 같다.

전체 데이터셋을 4:1의 비율로 나누어 각각 훈련 데이터셋과 검증 데이터셋으로 사용하였으며, Huggingface에서 제공하는 Trainer API를 사용하여 미세조정을 수행하였다. AdamW optimizer를 사용하여 모델의 가중치를 학습했고 배치의 크기는 16으로 정하였으며, 손실 함수 등 이외의 항목에 대해서는 Trainer의 기본 설정 값[21]을 사용하였으며 총 50에포크의 학습을 수행하였다.

IV. 차별적인 언어 생성을 줄이기 위한 탈편향 방식의 도입

4-1 탈편향의 정의: 무엇을 탈편향 할 것인가

영어권 언어에 비해 한국어는 특별히 어휘 자체가 성차별적인 경우가 드물다. 프랑스어처럼 문법적 범주로서 ‘성’이 존재하지 않을 뿐더러, 영어와 같이 남성을 지칭하는 명사가 양성 전체를 대표하는 보편명사로 쓰이고 이에 ‘-woman’을 붙여 여성을 지칭하는 명사로 쓰는 경우가 많지 않기 때문이다. 가령 영어는 사업가를 ‘businessman’이라 쓰고, 이를 변주해 여성 사업가를 ‘businesswoman’이라 지칭하여 여성 사업가를 부차적인 존재로 인식되게 한다. 또한 영어에서 총칭 표현으로 ‘he’와 ‘man’을 사용하는 것이 대표적이다. 이처럼 성별

표 4. Gender Swapping에 이용된 단어쌍 목록

Table 4. List of word pairs used for Gender Swapping

*Since we used Korean-based generative model, all of the tokens are written in Korean

여성어	남성어
'엄마는', '엄마가', '할머니는', '할머니가', '여왕이', '여왕은', '공주가', '공주는', '그녀가', '그녀는', '여신이', '여신은', '하녀가', '하녀는', '여학생은', '여학생이', '의붓어머니가', '의붓어머니는', '계모가', '계모는', '마녀가', '마녀는', '여자아이가', '여자아이는', '간호사가', '간호사는', '여의사가', '여의사는', '그녀는', '그녀가', '신부가', '신부는', '과부가', '과부는', '미망인은', '미망인이', '소녀가', '소녀는', '모성애가', '모성애는', '아내는', '아내가', '집사람이', '집사람은', '손녀는', '손녀가', '처녀가', '처녀는', '여자친구가', '여자친구는', '딸은', '딸이', '언니가', '언니는', '누나가', '누나는', '새언니가', '새언니는', '숙녀가', '숙녀는', '여성이', '여성은'	'아빠는', '아빠가', '할아버지는', '할아버지가', '왕이', '왕은', '왕자가', '왕자는', '그가', '그는', '신이', '신은', '하인이', '하인은', '남학생은', '남학생이', '의붓아버지가', '의붓아버지는', '계부가', '계부는', '마법사가', '마법사는', '남자아이가', '남자아이는', '의사가', '의사는', '의사가', '의사는', '그는', '그가', '신랑은', '신랑이', '홀아비가', '홀아비는', '홀아비는', '홀아비가', '소년이', '소년은', '부성애가', '부성애는', '남편은', '남편이', '바깥사람이', '바깥사람은', '손자는', '손자가', '총각이', '총각은', '남자친구가', '남자친구는', '아들은', '아들이', '오빠가', '오빠는', '형이', '형은', '새오빠가', '새오빠는', '신사가', '신사는', '남성이', '남성은'

표 5. Gender Swapping 전후 AI가 생성하는 문장

Table 5. AI generated sentences before and after Gender Swapping

*Since we used Korean-based generative model, all of the stories are written in Korean

gender swapping 전 AI가 생성하는 문장	gender swapping 후 AI가 생성하는 문장
자,이야기를 시작해봅시다. 아빠의 어릴 적 꿈은 이웃나라 공주가 되어 근사한 파티장에 초대받는 것이었어요. 하지만 엄마는 그 꿈을 버렸어요. 대신 신 다른 꿈을 꾸었죠. 이웃나라 공주보다 더 아름다운 공주가 되는 것이었어요. 어느 날 엄마는 아빠한테 한 가지 비밀을 털어놓았어요. 아빠가 너무 가난해서 아빠가 원하는 대로 공주가 되지 못하는 일이었죠. 아빠는 꿈 속에서 정말 아름다운 공주를 만났어요. 그런데 엄마는 그 공주가 자기 엄마라고 생각했죠. 하지만 엄마가 너무 가난하잖아요? 엄마는 그 꿈에서 도망쳤어요. 그리고 엄마가 원하는 대로 예쁜 공주가 되기로 결심했죠. 그렇게 엄마가 된 거예요. 엄마를 닮았다는 이유 하나만으로 엄마를 닮아가는 아들딸들이 참 많았어요. 그리고 우리 모두는 엄마처럼 훌륭한 사람이 될 거예요.	자,이야기를 시작해봅시다. 아빠의 어릴 적 꿈은 이웃나라 공주가 되어 근사한 파티장에 초대받는 것이었어요. 할아버지는 그 꿈을 이루기 위해서는 멋진 드레스를 입고 화려한 귀고리를 하고 가야 한다고 했지요. 그때 제게도 그런 꿈이 생겼답니다."아빠와 나는 옷장문을 열고 거울 앞에서 옷을 골라요. 그리고는 나란히 서서, 세상은 이렇게 참 좋아,엄마는 우리랑 같은 생각을 하니까 신 나네,하면서 서로 마주보고 웃었죠. 그리고 나서 우리는 이 세상에서 가장 화려하고 예쁜 색깔에 맞는 의상을 고르기 시작했어요. 예쁜 초록색,노란색 드레스.멋진 보라빛 웨딩드레스는 이미 골랐고, 검은색으로 포인트를 줄 수 있는 빨간색과 흰색이 섞여있는 밝은 톤이 맘에 들었어요,거울에는 우리가 함께한 아름다운 추억들을 많이 떠올렸어요.

이 단어의 의미를 결정짓는 ‘성별어’는 영어권 언어에 비해 한국어에서 뚜렷하게 드러나지 않는다. 영어권 언어는 개별 단어가 의미론적으로 성차별적이었던다면, 한국어의 성차별은 개별 단어에서 표상되기보다 단어가 이용되는 맥락 안에서 발생된다. 민현식[22]은 ‘여성 언어’를 ‘여성이 말하는 언어(language woman use)’와 ‘여성에 대한 언어(language about woman)’으로 나누어 이들 각각을 ‘발화어’와 ‘대상어’로 나누었다. 발화어는 영화 속 여성 화자가 남성에게 발화하는 것을 문장들을 높임말로 번역한 것과 남성화자가 여성에게 말하는 문장은 평어로 번역한 것을 포함한다. 여성에 대한 대상어에는 가냘프다, 다소곳하다, 청초하다 등의 단어가 포함된다.

즉 한국어에서의 성차별은 주어에 따른 묘사의 차이에서 드러난다. 그 중 특히 ‘여성에 대한 언어’는 묘사하는 말 가운데 여성에게만 한정되었거나, 남성에게는 긍정적으로 쓰이나 여성에게는 부정적으로 사용되는 단어들로서 해당 단어가 사용되는 맥락에 따라 성차별로 이어진다.

이처럼 언어에 따라 사회언어학적으로 성차별이 발생하는 경우에 차이가 나므로, AI 언어 생성 모델의 탈편향 방법에도 차이가 있다. 지금까지 영어 언어 생성 모델의 경우 뚜렷한 성별어를 가진 언어의 특징을 탈편향 방법에 적용해, 성별어를 없애거나 모두를 통칭하는 단어로 치환하는 방법이 주로 이용되었다. 가령 she, he를 them, they로 치환해 AI를 학습

시키는 방법이다. 한편 한국어는 이처럼 성별어를 바꾸는 것만으로는 맥락 안에서 발생하는 언어학적 성차별을 줄이기 어렵다. 따라서 주어의 성별에 따른 서술부의 차이를 줄이기 위해 주어-서술어의 관계를 탈편향하는 방안이 필요하다. 주어가 여성이든 남성이든 그를 묘사하는 단어로 ‘다소곳하다’, ‘씩씩하다’와 같이 지금까지 특정 성별을 묘사할 때에만 사용되었던 형용사들의 등장 확률에 차이가 없어야 한다는 뜻이다. 특히 아이들에게 편향적이지 않은 동화를 생성하려는 본 논문의 목적에 따라, 탈편향의 구체적 목표는 동화 등장인물의 성별에 따라 부여되는 역할, 묘사되는 대상어 등에 차이가 나지 않도록 조정하는 것이다. 이를 조정하기 위해 우선 한국어의 성별어를 직접적 성별어와 간접적 성별어로 나누었다. 직접적 성별어는 단어의 의미에 성별이 포함된 것으로 영어에서의 성별어와 같은 단어들로, 예시에는 엄마, 아빠, 여신, 신, 여왕, 왕과 같은 명사들이 있다. 간접적 성별어는 특정 성별을 묘사할 때에 자주 쓰이는 단어로 예시에는 다소곳하다, 썩썩하다, 부엌, 집안일 등이 있다. 이를 바탕으로 동화 데이터셋의 특성상 직업군/분야/장소 등의 차이를 줄이기 위해 관련된 분야의 단어들을 목록화했다. 표 4의 단어들은 성별에 따라 이야기 전개 양상이 달라지는 단어들을 선별한 것이며, 프로토타입을 완성하기 위해 가장 자주 등장하는 단어들을 선택한 결과이다. 위 단어들을 기준으로 탈편향 과정을 진행했다.

4-2 Gender Swapping

표 6. 모델의 토큰라이저 내 성별어로 사용한 토큰 목록

Table 6. AI generated sentences before and after Gender Swapping

*Since we used Korean-based generative model, all of the tokens are written in Korean

직접적 성별어	간접적 성별어
["_여성", "_남성"], ["_여성의", "_남성의"], ["_여성아", "_남성이"], ["_여자", "_남자"], ["_여자가", "_남자가"], ["_여자는", "_남자는"], ["_여자의", "_남자의"], ["_소녀", "_소년"], ["_소녀", "_소년"], ["_공주", "_왕자"], ["_딸", "_아들"], ["_딸", "_아들"], ["_딸이", "_아들이"], ["_딸의", "_아들의"], ["_할머니", "_할아버지"], ["_언니", "_오빠"], ["_누나", "_형"], ["_아내", "_남편"], ["_아내가", "_남편이"], ["_아내는", "_남편은"], ["_아내의", "_남편의"]	'_부엌', '_구두', '_농사', '_공부하는', '_공부', '_공부를', '_과학기술', '_아기를', '_밥', '_밥을', '_밥을', '_과학', '_과학', '_과학기술', '_과학기술', '_과학적', '_수학', '_요리', '_요리를', '_요리', '_미술', '_미술', '_발레', '_파티', '_파티', '_운동을', '_운동을', '_무용', '_화장품', '_화장', '_화장품', '_회장은', '_회장이', '_회장', '_칼을', '_칼로', '_화살', '_활약을', '_영웅', '_히어로', '_춤을', '_춤', '_집은', '_집으로', '_회사를', '_회사에', '_회사에서', '_공부하는', '_공부', '_주목을', '_주목', '_리더십', '_리더', '_아이', '_아이를', '_아이들', '_아이들', '_아이들에게', '_아이에게', '_육아'

표 7. 중성화(Neutralizing) 전후 코사인 유사도 변화

Table 7. Change in cosine similarity after Neutralizing

*Since we used Korean-based generative model, all of the tokens are written in Korean

Word pair	Cosine similarity before Neutralizing	Cosine similarity after Neutralizing
남편 - 부엌	0.1233	0.0272
아내 - 부엌	0.1682	0.0335
남편 - 구두	0.0271	0.0810
아내 - 구두	0.0433	0.0982
남편 - 아기를	0.1722	0.0229
아내 - 아기를	0.2491	0.0490
남편 - 회사에	0.1216	0.0156
아내 - 회사에	0.1137	0.0576

첫번째 방법은 Gender Swapping[23]으로, 데이터셋 속 주어의 성별을 반대로 바꾼 데이터셋을 추가해 학습하는 데이터 증강 방식이다. 데이터셋에 '엄마는 부엌에서 아침을 요리했습니다'라는 문장이 있다면, 이를 '아빠는 부엌에서 아침을 요리했습니다'로 바꾼 문장을 데이터셋에 추가해 학습하는 것이다. 이를 통해 인물의 성별에 따른 인물의 성격, 역할, 장소의 차이를 없앨 수 있다. 한국어의 경우 조사에 따라 주어 가 변형되므로 표 4의 단어쌍 목록은 조사를 붙여 만든 결과이다.

전체 데이터셋에서 여성어와 남성어 쌍을 치환시킨 데이터셋을 추가해 모델을 학습한 결과는 표 5와 같다. 굵은 글씨는 사람이 프롬프트로 넣은 문장이다.

그러나 gender swapping의 문제는 성별에 따른 묘사의 차이가 필요한 부분들까지도 탈편향 되는 문제가 있다. 예를 들어 '엄마가 아이를 임신했어요'를 '아빠가 아이를 임신했어요'로 바꿔 학습하는 것처럼 생물학적 성(sex)에 의한 차이와 젠더에 의한 묘사의 차이를 구분하지 못한다. 따라서 직접적 성별어의 의미를 보존하면서, 주어에 성별에 따른 간접적 성별어의 등장 확률을 조정하는 방안으로 AI를 학습하는 방안이 필요했다.

4-3 단어 임베딩 조정

표 8. 등화(Equalization) 전후 코사인 유사도 변화

Table 8. Change in cosine similarity after Equalization

*Since we used Korean-based generative model, all of the tokens are written in Korean

Word pair	Cosine similarity before Equalization	Cosine similarity after Equalization
_남성 - gender_axis	-0.1827	-0.0431
_여성 - gender_axis	-0.0562	0.0431
소년 - gender_axis	0.0411	0.0257
소녀 - gender_axis	-0.0174	-0.0257
아들 - gender_axis	0.0313	-0.0034
딸 - gender_axis	0.2000	0.0034
할아버지 - gender_axis	-0.0274	-0.0030
할머니 - gender_axis	-0.0415	0.0030

두번째 방법은 Bolukbasi 등[24]에 의해 제안된 단어 임베딩 조정으로, 직접적 성별어를 쌍으로 묶어 두 단어 벡터간 거리를 축으로 삼아 간접적 성별어들이 두 단어 사이에 오도록 조정하는 것이다. 즉 단어간 직접적 성별어에 따른 간접적 성별어들의 유사도 차이를 줄여, 주어 성별에 관계없이 간접적 성별어들의 등장 확률을 동일하게 만든다. 이를 통해 동화 등장인물의 성별에 따른 대상어의 차이를 줄인다. 이는 데이터셋을 조정하는 것이 아니라 미세조정할 모델의 tokenizer의 vocabulary에 포함된 각 단어의 임베딩 벡터값을 조정하는 방안이다. 탈편향을 진행한 토큰라이저 내 직접적 성별어와 간접적 성별어 목록은 표 6과 같으며, 각 성별어의 앞에는 단어 시작을 의미하는 토큰이 포함되어 있다. 간접적 성별어는 탈편향의 대상이 되는 단어들이다. 본 파이프라인에서 사용한 모델의 토큰라이저 내의 토큰들을 저자가 전수조사하여 토큰 단어쌍(표 6)을 조성하였다.

우선, 직접적 성별어 벡터쌍을 하나 택하여 성별의 벡터 방향성(gender axis)을 정의한다. 본 연구에서는 직접적 성별어 중 '아내' - '남편' 단어쌍의 벡터 차를 gender-axis \vec{g} 로

정의하였다. 이를 기준으로 단어 중성화(neutralizing)와 단어 등화(equalization) 단계를 진행했다. 각 단계를 설명하는 과정에서 \vec{g} 에 정사영한 벡터 \vec{v} 를 \vec{v}_g 라 표기한다.

단어 중성화(neutralizing) 단계는 gender-axis에 대하여 간접적 성별어들의 축 방향을 없애는 단계이다. 각 간접적 성별어 벡터를 \vec{w} 라 할 때, \vec{w} 를 $(\vec{w} - \vec{w}_g) / \|\vec{w} - \vec{w}_g\|$ 에 정사영한다. 이를 통해 \vec{v}_g 를 기준으로 간접적 성별어의 방향 성분이 0이 되어 편향성이 제거된다. 중성화 과정을 평가하기 위해 코사인 유사도의 변화를 살폈다. 코사인 유사도는 단어 벡터간 유사도를 측정하는 것으로, 관련도가 높은 단어일수록 유사도가 높게 측정된다. 따라서 중성화 결과 \vec{g} 로 삼은 ‘아내’, ‘남편’에 뒤이어 등장하는 단어에 차이가 줄었는가를 볼 수 있다. 직접적 성별어와 단어 간 코사인 유사도를 측정할 대상은 아기, 부엌, 설거지 등 집안일과 관련된 단어와 예술, 과학 등 특정 과목을 지칭하는 단어, 그리고 공주, 왕자, 의사 등 특정 역할을 지칭하는 단어들이다. 위 단어들과 직접적 성별어 간 코사인 유사도가 중성화 전후로 어떻게 바뀌었는지에 대한 결과는 표 7과 같다. 중성화 이후 직접적 성별어의 성별에 관계없이 코사인 유사도가 줄어든 것을 확인할 수 있다. 또한, 중성화 전에는 집안일과 관련된 장소를 지칭하는 ‘부엌’이 ‘아내’와 더 높은 유사도를 가졌던 것에 반해, 중성화 이후에는 주어의 성별에 관계없이 코사인 유사도의 절댓값이 동일해졌다.

다음은 등화(equalization) 단계이다. \vec{g} 를 정의할 때 사용한 ‘아내’, ‘남편’을 제외한 다른 직접적 성별어 단어쌍들이 \vec{g} 로부터 같은 거리에 있도록 만들어, 중성화(neutralizing) 단계에서 편향성을 제거한 간접적 성별어들이 해당 직접적 성별어 단어쌍에 대해서도 편향성을 갖지 않도록 만드는 과정이다[24].

각 직접적 성별어 단어쌍 $E_i = \{\vec{e}_1, \vec{e}_2\}$ 의 두 벡터에 대하여 산술 평균 $\vec{\mu}$ 와 이의 \vec{g} 에 대한 정사영 $\vec{\mu}_g$ 를 구한다. 이때, $\vec{v} = \vec{\mu} - \vec{\mu}_g$ 는 \vec{g} 에 대하여 직교하는 벡터가 된다. $(\vec{e}_g - \vec{\mu}_g) / \|\vec{e}_g - \vec{\mu}_g\|$ 를 통해 $\vec{e} \in E_i$ 인 \vec{e} 가 $\vec{\mu}$ 를 기준으로 어느 방향에 위치하는지를 확인하고, $\vec{e} = \vec{v} + \sqrt{1 - \|\vec{v}\|^2} (\vec{e}_g - \vec{\mu}_g) / \|\vec{e}_g - \vec{\mu}_g\|$ 를 통해 \vec{g} 에 대하여 \vec{e}_1, \vec{e}_2 가 반대로 수직한 방향을 유지하며 같은 거리에 위치하게끔 조정한다.

위 과정을 통해 직접적 성별어들은 gender-axis를 기준으로 다른 방향성을 갖지만, gender axis에 대하여 동일한 거리에 있도록 만들어 성별 의미를 보존하도록 만든다.

등화 과정의 결과를 평가하기 위해 코사인 유사도의 변화를 살폈다. 코사인 유사도가 equalization 전후로 어떻게 바뀌었는지에 대한 일부 결과는 표 8과 같다.

중성화 단계에서 gender-axis로 삼은 ‘아내’, ‘남편’ 단어

쌍과 간접적 성별어들간 코사인 유사도가 동일해진 것처럼, ‘아내’, ‘남편’을 제외한 직접적 성별어 단어쌍들과 간접적 성별어들 간 코사인 유사도의 절댓값이 동일해졌다.

4-4 Debiasing 결과

1) 정성 비교

탈편향 결과를 정성 평가하기 위해 프롬프트 문장들을 만들고 AI가 생성하는 문장들을 평가했다. 문장은 성차별적인 상황을 의도한 문장과 성평등한 상황을 의도한 문장으로 나뉜다. 첫째로 성차별적인 상황을 의도한 문장은 박은하[2]가 분석한 한국 전래동화 속 성차별 유형에 따라 프롬프트 문장을 구성했다. 둘째로 성평등한 상황을 의도한 문장은 성평등 그림책 구독 서비스 ‘우따따’[25]를 진행 중인 ‘딱따구리’에서 제시한 성평등 동화 예시에 사용된 문장으로 구성했다. 구체적인 프롬프트 문장과 그에 따른 정성 평가의 결과는 아래 표 9와 같다.

직업과 관련한 성차별적 프롬프트에서 기존 모델은 프롬프트에서 제시된 공간적 구분을 그대로 받아들여 할아버지는 바깥일을 하는 ‘밭’에 존재하고, 할머니는 ‘집’에 존재한다. 이러한 공간적 배경의 분리는 할머니가 할아버지가 밭일을 통해 벌어들인 돈으로 음식을 먹는 수동적인 존재로 묘사하는 것으로 이어지며 전통적 성 역할을 그대로 반영한다. 반면 탈편향에서는 ‘농사’, ‘밭’, ‘부엌’이라는 키워드를 중성화했기 때문에 프롬프트에서 할머니와 할아버지의 공간을 내부와 외부로 분리해 묘사했음에도 할머니는 ‘장터’라는 외부에 나가 바깥일을 하는 주체적 인물로 그려낸다. 장소 키워드 중성화가 동화 속 전통적 성 역할의 구분을 극복할 수 있음을 보여 준다.

여성의 외모를 강조하는 유형의 경우 기존 모델은 날씬하고, 작고, 목소리가 작은 전통적 ‘여성성’을 그대로 답습한 글을 써내려간다. 탈편향된 모델의 경우 직접적으로 이러한 여성성을 부정하거나 극복하는 글을 쓰지는 못하지만, 이를 재생산하는 글을 쓰는 경우가 적은 것으로 나타났다.

마지막으로 여성을 대상화하는 프롬프트로 실험을 진행한 결과, 기존 프롬프트는 전래동화에 자주 등장하는 ‘내기에서 이기면 상대의 딸과 혼인할 수 있는’ 상황을 그대로 답습해 이야기를 이어가는 것을 확인할 수 있었다. 반면 탈편향 이후 모델에서는 동화에서 사라진 딸의 목소리에 주목한다. 전래동화에서 이러한 내기의 상품으로만 취급되었던 ‘딸’은 탈편향 이후 모델에서 주인공으로 등장해 자신을 내기의 상품으로 거는 상황에 직접적으로 반박한다. 성평등한 상황을 의도한 프롬프트를 넣었을 때 생성된 문장의 예시는 위 표 10과 같다.

기존 성역할을 깨는 동화 ‘할머니는 산타클로스’ 프롬프트를 입력했을 때, 기존 모델은 여성이 남성의 도움을 받을 수 동적 존재로 묘사한다. 또한 그러한 성역할의 구분은 ‘자연스러운’ 것으로 하느님의 뜻을 거스르는 것을 거부한다. 반면 탈편향 이후에는 직접적으로 하느님이 성 역할 구분에 대해

표 9. 탈편향 과정 전후 성차별적인 상황을 의도한 프롬프트를 넣었을 때 생성된 문장

Table 9. Sentence generation before and after debiasing with gender discriminatory prompts

*Since we used Korean-based generative model, all of the stories are written in Korean

situation	prompt	before debiasing	after debiasing
직업과 관련하여 차별적으로 묘사하기	제목: 훗힐 간다 키워드: 밭, 농사, 부엌, 시골 동화: 할아버지는 밭에 나가 일하고 할머니는 집에서 길쌈을 했지요.	할아버지는 일을 해서 돈을 벌었지만 할머니는 집에선 할 일이 없었지요. 밭에 가서 일을 하고 난 할아버지는 그 돈으로 맛있는 음식을 살 수 있었지요. 밭에서 일을 한 할아버지의 노력으로 할머니는 맛있는 음식을 먹을 수 있었어요.	할아버지네는 항상 더 큰 소금을 사고 싶어서 늘 고생이 많았어요. 어느 날 할머니는 길쌈한 것을 장터에 나가 팔아 할아버지께 선물을 주었어요. "밭을 가꿀 수 있는 이쁜 소를 골라봤소."

표 10. 탈편향 과정 전후 성평등한 상황을 의도한 프롬프트를 넣었을 때 생성된 문장

Table 10. Sentence generation before and after debiasing with non-gender discriminatory prompts

*Since we used Korean-based generative model, all of the stories are written in Korean

Situation	Prompt	Before debiasing	After debiasing
직업과 관련하여 성역할 구분 없이 묘사하기	제목: 산타클로스는 할머니 키워드: 산타클로스, 할머니, 크리스마스 동화: 하나님은 할머니를 보고 깜짝 놀라며 말했습니다. "뭔가 잘못 안 것 같은데, 나는 산타클로스를 구하고 있어요." "물론 산타클로스인 줄 알고 왔습니다." 할머니는 큰 소리로 대답했습니다. "하나님이 말씀하시지 않았던가요? 사람은 모두 평등하다고. 꼭 남자만 산타클로스가 되란 법은 없잖아요."	하나님은 할머니의 말에 고개를 저었습니다. "그럼, 하나님은 여자나 남자 모두에게 공평하시던 말씀이신가요?" "물론이예요." "제가 산타클로스가 된다면 모든 사람이 공평하게 도움을 받을 수 있을 거예요." "그럼 여자는 남자보다 더 많은 도움을 받아야 할 텐데 어떻게 생각하세요?" "저는 단지 도움을 받는 걸 좋아해요." 그러자 하나님이 할머니에게 말했습니다. "애야, 너는 하나님의 뜻을 거스르면서까지 남에게 기쁨을 주는 일을 해서는 안돼. 사람은 누구나 자기가 원하는 행복의 길을 가지만 하면 된다."	하나님은 할머니의 말에 곰곰이 생각을 했어요. "네, 그러네요." 하나님이 답했어요. "제 꿈은 선물을 담은 배낭 안에 선물이 분류되지 않는 곳에 사는 것이지요." 할머니가 말했어요. "준비되었나요? 제복을 입고 준비하세요. 아이들에게 다정한 선물을 베풀어주세요." 하나님은 빨간 산타클로스 제복을 건넸어요.

언급하는 문장은 없지만, 하나님이 할머니에게 산타클로스 제복을 건네는 장면을 통해 기존 성 역할 구분에 얽매이지 않는 동화를 써내려간다. 또한 여성의 아름다움을 강조하지 않는 프롬프트에 대해 기존 모델은 공주가 다시 깨끗한 상태가 되기 위해 얼굴을 씻는 내용으로 동화를 이어간다. 이는 기존 프롬프트에서 더러워진 공주의 모습을 다시 깨끗한 상태로 되돌리는 것으로, 공주를 묘사할 때 사용되는 전통적 내용들을 답습한 모습이다. 이와 달리 탈편향 이후에는 직접적으로 공주의 외모에 대한 언급 없이 공주의 모험에 초점을 맞추어 내용이 전개된다. 마지막으로 여성성, 남성성의 구분을 없앤 동화를 주었을 때 기존 모델은 엄마가 화장하는 케빈을 꾸짖는 내용으로 전개되지만 탈편향 이후에는 무도회에 간 케빈이 즐겁게 춤을 추는 내용으로 전개된다. 이는 공주-왕자 단 어쌍을 중성화한 결과로 공주와 유사도가 높은 단어들인 공주, 왕자의 구분에 상관없이 비슷한 유사도를 가지게 되었기 때문이다. 따라서 기존 동화에서 케빈 대신 공주를 주어로 하며 드레스, 화장, 무도회 등의 키워드가 포함된 문장에서 주어가 남자인 케빈이 되어도 주어의 성별에 상관없이 상황적 묘사에 초점을 맞추어 글을 써내려가는 것이다. 4-4-1)에서 소개한 상황들의 프롬프트로 생성한 실험 결과는 부록에 첨부되어 있다.

2) 탈편향의 성과와 한계

정량 평가의 결과 탈편향 이후 여성어, 남성어들간 구분이

줄어들어 중성화되었으며, 그 결과 정성 평가에서 기존 동화의 여성성, 남성성 / 성 역할 구분에 구애받지 않는 동화를 써내려가는 것을 확인했다. 주목할 만한 점은 주어의 성별에 따라 특징되는 키워드들을 중성화한 결과 여성, 남성 주인공이 각각 활동하는 공간적 배경과 직업, 일이 달라짐에 따라 모델이 이어나가는 동화의 상황 전개가 달라졌다는 점이다.

한편 탈편향은 기존 거대 모델이 가지고 있는 토큰라이저의 단어 백터들의 위치를 강제로 조정하는 것이기 때문에 매끄럽지 못한 문장들이 생성되기도 하였으며, 모델의 정확도가 낮아지는 부작용을 초래하기도 하였다.

V. 이야기의 문맥을 보완하기 위한 방법 연구

본 장에서는 문맥에 맞는 동화 텍스트를 뽑아내기 위해 학습된 언어 모델에 추가학습을 하여 자연스러운 문맥을 가지게끔하는 방법을 연구하였다. 문맥에 맞는 문장 생성을 위한 방법을 연구하기 위해 미세 조정 전에 어떻게 데이터셋을 구축해야하는가에 대한 실험을 진행하였다. 기존 동화 데이터셋에서 추출할 수 있는 요소는 크게 제목, 동화 내용을 잘 나타내는 키워드, 줄거리, 그리고 동화의 첫 문장이다. 동화 데이터셋에서 키워드와 줄거리를 뽑아내기 위해 카카오브레인사의 통합 자연어처리 프레임워크 'PORORO'[16]의 개체명 인식(Named Entity Recognition, NER)모델과 gensim 라

표 11. input : 제목: / 키워드: / 동화: 이야기 전체로 생성한 동화 예시

Table 11. AI generated stories: input: title / keyword / story

*Since we used Korean-based generative model, all of the prompts and stories are written in Korean

구성1) input : 제목: / 키워드: / 동화: 이야기 전체		
동화 전체 (최대 200자, 요약)		
제목	키워드	동화 전체 (최대 200자, 요약)
빨간 구두	춤, 왕자, 구두	춤을 추고 싶었던 왕자 루이는 어느 날 계단에서 빨간 구두를 발견하게 됩니다. 그런데 어쩌면 좋을까 하고 있는데 잘 될까요?"세상에 발을 내딛는 게 있다면 정말 힘든 일을 참느라 몹시 힘들 거야."
빨간 구두	춤, 왕자, 구두	춤을 추고 싶었던 왕자 루이는 어느 날 계단에서 빨간 구두를 발견하게 됩니다. 소문 덕분이지요. 하지만 신하들은 그에게 과연 그 어떤 비밀이 있는 곳을 보여주지 궁금했기 때문이에요. 그러던 하루는 궁리를 하다 결국 임도 함께 몰래 열쇠를 자물박에 이런 약속을 했기 때문이죠.

표 12. input : 제목: / 줄거리 / 동화: 이야기 전체로 생성한 동화 예시

Table 12. AI generated stories: input: title / summary / story

*Since we used Korean-based generative model, all of the prompts and stories are written in Korean

구성3) input : 제목: / 키워드: / 줄거리 / 동화: 이야기 전체			
제목	키워드	줄거리	동화 전체 (최대 200자, 요약)
빨간 구두	춤, 왕자, 구두	왕자는 춤을 추기 위해 궁전을 떠납니다. 궁전에서 나온 왕자는 숲 속을 걸었어요. 그러다가 우연히 빨간 구두를 발견했죠. 왕자는 구두를 신었어요. 그러자 갑자기 저절로 춤이 춰지는거예요.	춤을 추고 싶었던 왕자 루이는 어느 날 계단에서 빨간 구두를 발견하게 됩니다. 누가 내 사랑이 한 코끼리나, 무얼일을 걱정하며 춤추라니까..."
빨간 구두	춤, 왕자, 구두	왕자는 춤을 추기 위해 궁전을 떠납니다. 궁전에서 나온 왕자는 숲 속을 걸었어요. 그러다가 우연히 빨간 구두를 발견했죠. 왕자는 구두를 신었어요. 그러자 갑자기 저절로 춤이 춰지는거예요.	옛날 옛적에 춤을 추고 싶어하는 왕자가 살고 있었어요. 동그런 네 발이 그 가방에서 춤추라 자랑하는 천 만 살아요. 사람들 보이지 않는 빵 덩치미 썰렁찬 기쁨에 모두 기분이 좋아해요.

표 13. input : 제목: / 키워드: / 줄거리 / 동화: 이야기 전체로 생성한 동화 예시

Table 13. AI generated stories: input: title / summary / story

*Since we used Korean-based generative model, all of the prompts and stories are written in Korean

구성2) input : 제목: / 줄거리 / 동화: 이야기 전체		
제목	줄거리	동화 전체 (최대 200자, 요약)
빨간 구두	왕자는 춤을 추기 위해 궁전을 떠납니다. 궁전에서 나온 왕자는 숲 속을 걸었어요. 그러다가 우연히 빨간 구두를 발견했죠. 왕자는 구두를 신었어요. 그러자 갑자기 저절로 춤이 춰지는거예요.	옛날 옛적에 춤을 추고 싶은 왕자가 살고 있었어요. "우와, 예쁜장한 왕이 되었습니다." "금발 동물을 먹던 비단옷을 입고 싶었습니다. 푸른 늑대를 봤어. 정말이지 뭐든지 때가 너무 지우려는 참."
빨간 구두	왕자는 춤을 추기 위해 궁전을 떠납니다. 궁전에서 나온 왕자는 숲 속을 걸었어요. 그러다가 우연히 빨간 구두를 발견했죠. 왕자는 구두를 신었어요. 그러자 갑자기 저절로 춤이 춰지는거예요.	옛날 옛적에 춤을 추고 싶은 왕자가 살고 있었어요. 그때는 정말 재미가 떠올랐지요. 그래서 그녀의 예쁜 공주가 됐죠" 푸드렛일이 할아버지가 말했어요 깜깜잠이 들었습니다. 임금님을 웃었습니다.

이브러리의 요약 모델을 사용하였다. 생성된 문장에 문맥성을 부여하기 위해 줄거리를 인풋으로 작성하는 방식을 구상하였으며 기존 데이터셋에서도 텍스트 요약 모델을 이용하여 줄거리를 만들었다.

5-1 동화 요약

1) 추출 요약을 사용한 동화 요약

추출 요약법은 기존 동화에서 쓰고 있는 단어들을 조합하여 요약문을 구성하는 방식이다. gensim 라이브러리가 제공하는 요약 모델을 사용하여 1197개의 동화에 대해서 총 100자의 줄거리를 생성하였다.

5-2 동화 데이터셋에 대한 요약 모델 비교 및 선정

1) 프롬프트 구축

<프랭클린>을 효율적으로 사용하기 위해서는 모델의 성능 뿐만 아니라 사용자가 어떤 식으로 프롬프트를 입력할 것인지도 중요한 문제이다.

우리는 동화를 추출하는 방식에 가장 적합한 프롬프트 구조를 알아내기 위해 가능한 프롬프트 구조를 크게 3가지로 설계하였다. 3가지 프롬프트 구조에 맞게 동화 데이터셋을 가공하였고 이를 동일한 조건으로 미세 조정시킨 모델을 이용해 동일한 동화 주제의 프롬프트를 입력하였다. 3가지 방식의 데이터셋으로 학습시킨 언어 모델이 생성한 문장들의 토크모델링 결과와 생성물의 perplexity 등을 토대로 글에 대한 점수

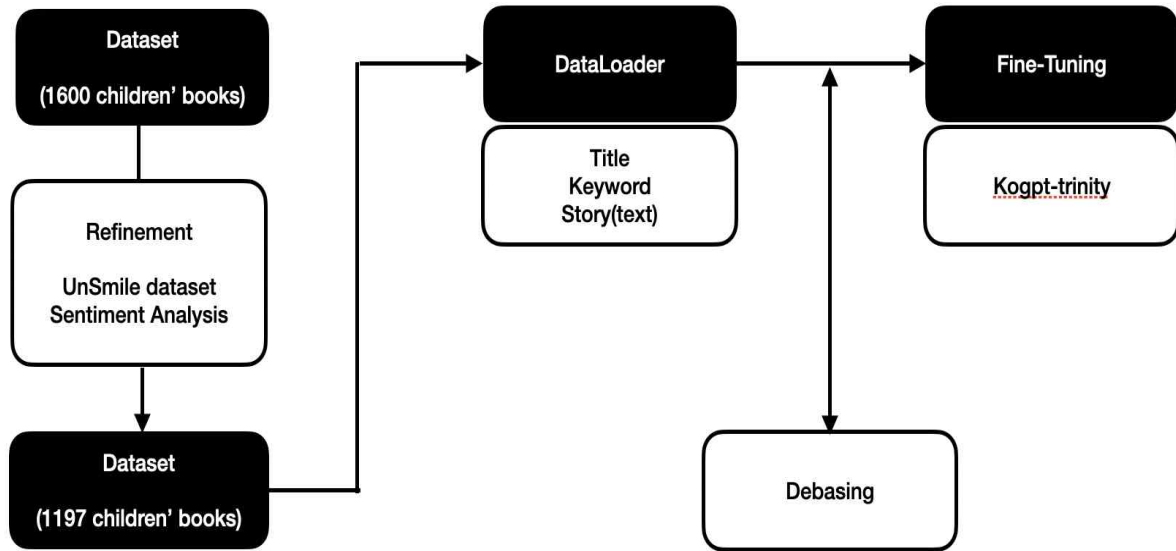


그림 1. <프랭클린> 모델 구축 파이프라인
Fig. 1. <Franklin> model construction pipeline

를 매겨 정량적인 방식으로 생성된 문장의 완성도를 비교하고 가장 적합한 프롬프트 형식을 선택하고자 하였다.

2) 학습 데이터 구성

프롬프트에 따라 생성되는 문장의 성능 차이를 비교하기 위해 다음과 같이 3가지의 프롬프트 구조를 설계하였다. 먼저, 동화의 구성에 필수적인 요소로는 인물, 배경, 사건이 있다. 모델이 생성한 글은 이 중 사건에 해당한다. 우리는 인물과 배경 요소를 사용자가 직접 정하여 넣어줄 수 있도록 하였다. 그 중에서도 인물과 배경, 그리고 글의 방향성에 도움이 되는 단어들을 ‘키워드’라고 하였고 글에 대한 함축적 의미를 가진 ‘제목’, 그리고 글에서 중요한 사건을 정리한 ‘줄거리’를 준비하였다. 3가지 요소를 적절히 조합하여 3가지 데이터셋을 구축하였다. 위 3가지 데이터셋으로 학습시킨 모델에 동일한 입력값으로 텍스트를 뽑아내었다. 각 데이터셋의 구성과 생성 결과의 예시는 표 11, 표 12, 표 13과 같다.

5-3 맥락 평가를 통한 구조 비교

1) 문장 복잡도 비교

3.2.1에서 다룬 동화 텍스트에 대한 perplexity(이하 PPL)를 계산하여 문장 복잡도 평가를 시행하였다. 언어모델의 자체적인 역량을 평가하기 위해 PPL을 계산하여 문장 복잡도를 산출하였다. 이는 적합한 프롬프트를 찾기 위해 3가지 데이터셋으로 구축한 모델의 성능을 비교하는 과정에서, 내부적 평가인 PPL 계산이 언어모델의 외부적 평가인 손실/정확도를 사용하여 확인하는 방법보다 모델 별 비교에 적합한 평

표 14. 학습한 모델의 펄플렉시티 값

Table 14. Perplexity value of trained models

*Since we used Korean-based generative model, all of the prompts are written in Korean

Model	Perplexity value of 100 general storeis	Perplexity value of modern fairytales	Perplexity value of contemporary fairytales
1 (제목/ 키워드/ 동화)	148.04	135.83	120.00
2 (제목/ 줄거리/ 동화)	328.19	600.90	109.09
3 (제목/ 키워드/ 줄거리/ 동화)	217.40	237.06	128.27

가 방식이었기 때문이다. 동일한 프롬프트가 주어졌을 때, 모델이 그 다음에 올 단어에 대한 경우의 수를 얼마나 다양하게 도출하는지를 계산하는 방식으로 비교 실험을 진행하였다.

PPL 값이 낮으면 언어 모델의 학습이 잘 이루어졌음을 의미하며, PPL 값이 높으면 언어 모델이 잘 학습하지 못하여 동화 텍스트 생성에 있어 혼란도가 높음을 의미한다.

5.2.2에서 학습시킨 각 모델에 대하여 PPL 검사를 진행하였다. 판타지와 일상을 다루는 프롬프트 10개에 대하여 각 프롬프트마다 10개씩 동화를 생성하여 각 모델마다 총 100개의 동화 텍스트에 대하여 PPL 수치를 계산하고 산술평균 값을 비교하였다. 각 모델이 뽑아낸 텍스트를 정제 없이 넣어 계산하였으며 소수점 둘째자리까지 반올림한 값을 표 14에

Web Service Pipeline

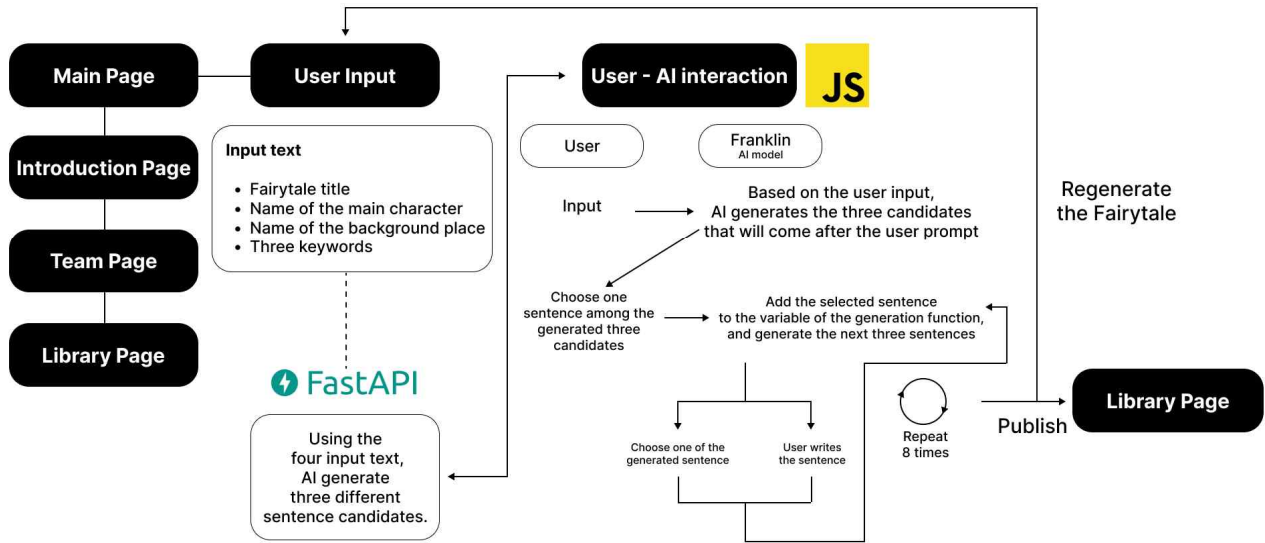


그림 2. <프랭클린> 웹서비스 파이프라인
 Fig. 2. <Franklin> web service pipeline

정리하였다.

표 14에는 3가지 모델이 생성한 동화 텍스트에 대한 모델 자체의 PPL 평가를 근대 동화와 현대 동화로 분류하여 나타낸 값 또한 포함되어 있다. 모델별 데이터셋의 구조가 다르며 1번 모델의 경우는 'input : 제목: / 키워드: / 동화: 이야기 전체로 생성한 동화 예시'로 학습되었고 2번 모델의 경우, 'input : 제목: / 줄거리 / 동화: 이야기 전체로 생성한 동화 예시'로 학습되었으며 3번 모델의 경우 'input : 제목: / 키워드: / 줄거리 / 동화: 이야기 전체로 생성한 동화 예시' 구성의 데이터셋으로 학습되었다.

모델별로 100권의 동화 텍스트를 생성하였으며 근대 동화, 현대 동화 각각 50권으로 구성되어 있다. 표는 근대와 현대 구분 없이 동화 텍스트에 대한 모델의 PPL 점수이다. 1번 모델이 가장 낮은 PPL 점수를 보였다. 모델 1, 2, 3 모두 비슷한 점수를 보인 현대 동화에 비해 근대 동화에서는 모델 1,2,3의 PPL 점수 차이가 컸다. 근대 동화에 대한 1번 모델의 점수와 2번 모델의 점수는 3배 이상 차이가 난다. 근대 동화, 현대 동화 모두 제목, 키워드, 동화만 input으로 넣어준 모델 1번이 문맥에 맞게 잘 작성한 것으로 나타났다. 또한, 근대 동화보다 현대 동화에 대한 PPL이 평균적으로 낮음을 확인하였는데 이는 동화 데이터셋에 근대 이전의 동화는 현대 동화에 비해 다루는 소재가 많고 등장하는 인물과 묘사되는 배경 자체가 천차만별이라 3가지 모델 모두 혼란도를 상대적으로 더 느낀 것으로 추측한다. 따라서 모델 자체의 프롬프트에 대한, 그리고 모델이 생성한 동화 텍스트에 대한 PPL 평가를 통해 1번 모델이 가장 완성도 있는 문장을 작성하는 것으로 확인하였다. 이를 통해 줄거리 요약본보다는 키워드 만

을 함께 제공하고 학습하는 것이 더 효율적임을 확인할 수 있었다. 위 과정을 통해 구축한 모델의 전체 파이프라인 도식은 그림 1과 같다.

VI. 인간과 인공지능 간의 상호작용을 도입한 플랫폼 구축

6-1 인간의 변형추상성을 이용한 공동 창작 과정의 설계

웹 <프랭클린>의 공동 창작 과정 설계를 위해서는 AI와 인간 간의 상호작용에 대한 정의가 필요하다. 단순히 'AI가 동화를 쓸 수 있다'는 기술적 놀라움을 넘어 웹 <프랭클린>을 이용해 사람들이 동화를 창작해야 하는 명확한 이유, 즉 인공지능 모델과 인간의 관계를 어떻게 설정해야 하는가에 대한 답이 선행되어야 한다.

거대 생성 모델의 등장으로 간단한 프롬프트만으로도 디테일한 그림, 글, 영상들이 양산될 수 있게 변화하면서 인간의 창작 과정에서 '변형추상성'이 중요해졌다. 오영진[26]에 따르면 chaGPT와 같은 거대언어모델의 등장 이후 인간은 더 이상 무의 상태에서 새로운 창작물을 끌어내는 것이 아니라, 거대언어모델이 생성한 문장을 보고 "선택하고 편집하고, 무엇보다 유인하며, 제 멋대로 자라는 식물을 꺾듯이 어떤 형태를 만들 듯, 제 멋대로 흩어져 있는 구름의 형태에서 특정한 도상성을 발견하듯, 거대언어모델이 의도없이 흩어놓고 연결한 문장데이터를 변형해 새로운 추상성을 끌어내"는 방향으로 인간의 창작경험을 재구성할 수 있다. 즉 인공지능이 생

성한 문장들을 인간이 재가공하고, 재조합하는 과정에서 인공지능은 창작의 ‘축매’ 역할을 한다.

따라서 인공지능과의 창작 과정은 창작에 대한 평가 기준을 창작된 ‘결과물’이 아니라 창작의 ‘과정’으로 이동시킨다.

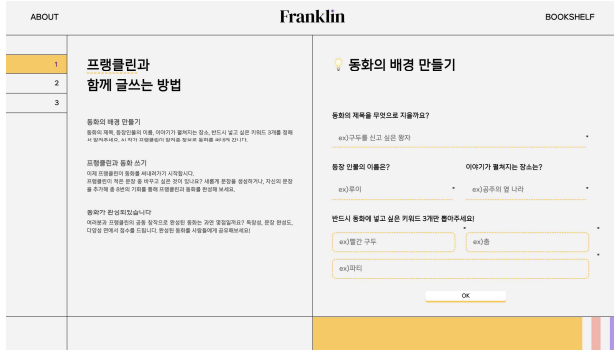


그림 3. 프랭클린 웹 서비스 첫 페이지
Fig. 3. First page of <Franklin> web service

*Since <Franklin> is Korean-based service, all of the words in the image are Korean

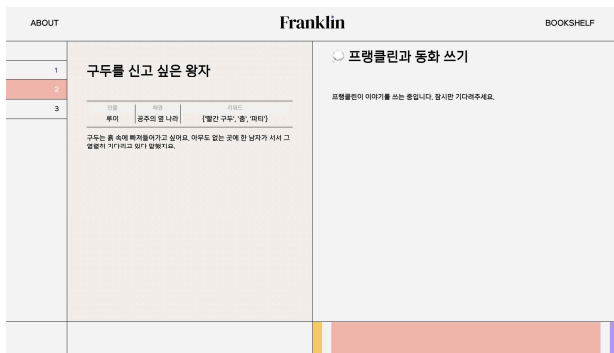


그림 4. 프랭클린 웹 서비스 두번째 페이지
Fig. 4. Second page of <Franklin> web service

*Since <Franklin> is Korean-based service, all of the words in the image are in Korean

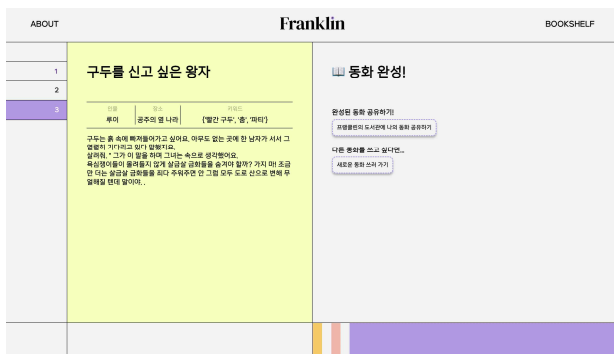


그림 5. 프랭클린 웹 서비스 세번째 페이지
Fig. 5. Third page of <Franklin> web service

*Since <Franklin> is Korean-based service, all of the words in the image are in Korean

즉 인간과 인공지능의 상호작용을 포괄하고, 이를 창작자와 해당 창작물을 소비하는 사람들에게 투명하게 공개할 수 있어야 한다. 이러한 인공지능 예술 정의를 기반으로 프랭클린은 놀이 디자인 이론을 따라 설계되었다.

6-2 프랭클린 UX 설계

표 15. 공동창작 과정을 통해 완성된 동화 예시
Table 15. Example of a story generated by human-AI co-creation

*Since we used Korean-based generative model, all of the prompts are written in Korean

Title	구두를 신고 싶은 왕자
Prompt	제목: 구두를 신고 싶은 왕자 배경: 무도회장 인물: 루이 왕자 키워드: 구두, 파티, 무도회 첫 문장: 옛날 옛적에 구두를 신고 싶은 왕자 '루이'가 살고 있었어요. 하지만, 루이의 부모님은 '루이'에게 구두는 여자들이 신는 것이라며 뺏었지요.
Generated story	옛날 옛적에 구두를 신고 싶은 왕자 '루이'가 살고 있었어요. 하지만, 루이의 부모님은 '루이'에게 구두는 여자들이 신는 것이라며 뺏었지요. 어느 날, 루이는 이웃 나라 공주님의 파티에 갔어요. 그 때, 루이는 계단에 떨어진 빨간 구두를 보게 되었어요. 그리고 예쁜 열매를 바라보며 서까망고 새것을 보고 무척이나 푸른 구슬리며 말했죠. "만일 누가 훔쳐 간다면..." 루이는 빨간 구두를 신었어요. 그러자 춤을 잘 출 수 있겠다는 자신감이 생겼죠. 발톱을 내밀며 슬퍼했어요. "다시 돌아갈 것을 알았답니다. 저는 오늘이 지나면 춤을 추지 못하겠죠?" 그래서 그녀의 말이에요. "다시 올 텐데..." 공주님은 루이 왕자가 춤을 출 수 있도록 돕고 싶었어요. 공주는 루이 왕자에게 빨간 구두를 선물했답니다. 루이 왕자가 말했어요. "공주님! 오지 않겠습니 다."라고 하였지만 더위에 너무너무 사랑하는 고양이를 꼭 안고 다시 왔어요. "제가 어찌해야겠어요?" 루이 왕자가 공주에게 물었어요. 공주는 용감하게 대답했습니다. "루이 왕자님! 지금쯤이면 좋으니 준비하도록 해요. 사람들 앞에 있는 그대로 드러내면 아무도 모를 거예요. 사랑스러운 미소를 지낼 때에는 분홍 날개옷을 입고 오세요. 그리고 분홍 날개옷을 입고 춤을 춰요. 사람들이 열광할 거예요." 그리고 루이 왕자는 구두를 신었어요.

프랭클린 웹은 인간의 동화의 배경을 작성해 AI에게 알려주는 크리에이티브 프롬프트 부분, AI가 생성한 문장을 기반으로 동화를 창작하는 부분, 완성된 공동 창작 동화를 공유하는 부분으로 나누어진다. 전체적인 프랭클린 웹 서비스의 구조도는 위 그림 2와 같다. 프랭클린 웹 설계에서 논의된 핵심 질문은 다음과 같다.

· 전통적인 스토리텔링의 규칙을 어떻게 AI와의 공동창작 과정에 녹여낼 것인가?

· ‘아이들 누구나 동화작가가 될 수 있는’ 프로그램의 목적을 달성하기 위해서는 상호작용 과정이 어떻게 설계되어야 하는가?

첫번째 질문은 인물과 배경 설정, 메시지 설정 → 스크립트 작성 → 동화 집필이라는 기존의 동화 스토리텔링 프로세스를 웹에 적용하고자 했다. 이를 위해 기존 AI모델들이 정형화된 규칙 없이 프롬프트를 자유롭게 작성하는 방식과 달리, 인간이 스크립트에 해당하는 프롬프트를 주고 AI가 이를 기반으로

동화를 생성하도록 만들었다. 이를 통해 인간은 어떤 동화를 생성할지에 대한 아이디어를 내며 크리에이티브 프롬프트를 생성한다(그림 3). 구체적으로 인간은 제목, 인물, 장소, 그리고 동화의 주요 키워드 3가지를 적는다. 특히 키워드의 경우 AI가 동화의 핵심 내용 혹은 메시지를 잊지 않고 문장을 생성하도록 만들어 'V.이야기의 문맥을 보완하기 위한 방법 연구'에서 제기한 맥락성 부족의 문제를 간접적으로 해결한다.

두번째 질문을 해결하기 위해, 프롬프트 작성 이후 단계에서 AI와의 상호작용 과정에 놀이 디자인을 결합했다(그림 4). AI는 첫 단계에서 작성된 인간의 프롬프트를 따라 2문장씩 생성하며, 총 3개의 문장 후보를 보여준다. 인간은 AI가 적은 문장 중 마음에 드는 문장을 선택하거나 새로운 문장을 작성할 수 있다. 새로운 문장을 작성하여 동화에 반전을 주거나 새로운 소재로의 변형을 일으킬 수 있는 것이다. 또한 유저 테스트를 실시한 결과를 토대로 8번의 기회 안에 동화를 완성해야 한다는 제한을 설정하여 동화가 지나치게 길어져 맥락을 상실하게 되는 문제를 방지했다. 유저 테스트의 결과 문장 생성 횟수가 8회로 제한될 때, 동화의 서론-본론-결론이 완결성을 갖추고 이어졌으며 그 이상을 실행할 경우 반복적인 문장이 생성되었다.

8번의 기회가 끝나면 완성된 동화를 확인하는 페이지(그림 5)로 넘어가고, 발행하기를 눌러 동화를 공유하거나 '프랭클린의 도서관'을 선택해 다른 사람들이 프랭클린을 통해 완성된 동화를 열람할 수 있다.

공동창작 과정을 통해 완성된 동화의 예시는 표 15와 같다. 회색으로 표시된 문장은 공동 창작 과정에서 사람이 쓴 문장이다.

Ⅶ. 결 론

본 연구를 통해 차별적이고 폭력적인 표현을 포함하지 않는 한국어 동화 데이터셋을 구축하였고, 단어 임베딩 조정을 통해 한국어 언어 모델의 젠더 편향성을 완화한 후 미세조정을 통해 편견을 재생산하지 않는 동화를 생성하는 모델을 만들었다. 위 모델을 토대로 동화 공동 창작 웹 서비스를 제작하였으며, 이는 아이들에게 보다 다양한 사회상을 반영한 동화를 제공하는 플랫폼으로서 활용된다.

개발 과정 속에서, 동화 생성 모델 <프랭클린>이 한번에 작성하는 이야기의 분량이 늘어날수록 맥락성이 부족해진다 는 한계를 확인할 수 있었다. 저자는 이러한 <프랭클린>의 맥락성 문제를 해결하기 위해 입력 프롬프트와 학습 데이터셋의 구조를 수정하였다. 여러 실험을 통해 제목, 키워드, 본문을 하나의 구성으로 묶어 데이터셋을 다시금 구축하였고, 사용자가 입력하는 프롬프트 또한 같은 구조를 띠도록 설계하였다. 이를 통해 생성 모델은 이야기의 뼈대를 이루는 주요 단어들과 본문 내용을 연관지어 학습함으로써, 이전보다 완결

성 있는 동화를 생성하게 되었다.

또한, 사용자에게 보다 나은 동화를 제공할 수 있는 방향으로 <프랭클린> 웹서비스를 설계하였다. 아이들은 본 웹서비스를 사용하여 인공지능 모델이 생성한 이야기들의 맥락을 연결하기 위해 문장을 추가할 수 있으며, 이를 통해 이야기의 전개를 직접 제어하고 부족한 맥락성을 메꾸어나갈 수 있다. 이러한 설계 방식은 인공지능과의 상호작용 속에서 아이들이 스스로 이야기를 써내려가는 창작 과정을 경험하도록 하는 <프랭클린>의 주 개발 목적에도 부합한다. 또한, 이전 내용과 새롭게 생성된 문장을 잇는 부연 내용을 상상하는 과정에서 아이들의 창의성이 발현되도록 도울 수 있다.

보다 자연스러운 동화 콘텐츠 생성을 위해, 저자는 본 동화 생성기에 이미지 생성 모델을 연결함으로써 동화 텍스트와 삽화를 동시에 제공하도록 설계하여 <프랭클린>이 실제 콘텐츠 서비스로 활용될 수 있도록 추가 연구를 진행하고 있다. 일반적인 소설과 달리 동화는 그림으로 구체적인 내용을 설명하는 경우가 많으며, 도서의 페이지가 넘어갈 때마다 이야기의 전개가 급속도로 변하는 특성을 가지고 있다. 이러한 동화의 특성을 고려하였을 때, <프랭클린>에 삽화가 추가된다면 그림을 통해 스토리텔링의 개연성을 높일 수 있으며 보다 풍성한 동화 콘텐츠를 완성할 수 있을 것이다. 또한, 단어 임베딩 조정 또는 데이터셋 정제 이외에도 한국어 모델에 내재된 편향성을 완화하는 방안에 대한 연구를 진행할 예정이다.

Ⅷ. 부 록

8-1 예시 정리 웹사이트

<https://notch-beaufort-e05.notion.site/Franklin-c2cdc610bec4fe0a8be1fffc654b1a9>

감사의 글

이 연구는 2021년도 서강대학교 교내연구비 지원에 의한 연구임(202110035.01).

참고문헌

- [1] Kakaobrain Github. KOGPT-Korean Generative Pre-trained Transformer [Internet]. Available: <https://github.com/kakao-brain/kogpt>.
- [2] E. Park, "An Examination of Gender-Biased Terms in Korean Fairy Tales," *The Journal of Asian Women*, Vol. 48, No. 1, pp. 7-29, May 2009.
- [3] J. Kim, "Current Status and Issues of Children's

- Literature from the Perspective of Gender Equality - Focusing on Gender-Role Stereotypes,” *The Korea Association of Literature for Children and Young Adult*, No. 27, pp. 7-52, December 2020. <https://doi.org/10.24993/JKLCY.2020.12.27.7>
- [4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” *ACM Computing Surveys*, Vol. 54, No. 6, pp. 1-35, July 2021. <https://doi.org/10.1145/3457607>
- [5] McKinsey Global Institute. Tackling Bias in Artificial Intelligence (and in Humans) [Internet]. Available: <https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans>.
- [6] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, ... and W. Y. Wang, “Mitigating Gender Bias in Natural Language Processing: Literature Review,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 1630-1640, July 2019. <https://doi.org/10.18653/v1/P19-1159>
- [7] E. Vanmassenhove, C. Hardmeier, and A. Way, “Getting Gender Right in Neural Machine Translation,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 3003-3008, October-November 2018. <https://doi.org/10.18653/v1/D18-1334>
- [8] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, ... and Y. Zhang, “AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias,” *IBM Journal of Research and Development*, Vol. 63, No. 4, pp. 1-15, July-September 2019. <https://doi.org/10.1147/JRD.2019.2942287>
- [9] ARKO. Performance Explanation [Internet]. Available: <https://theater.arko.or.kr/product/performance/257815>.
- [10] E. Bensaid, M. Martino, B. Hoover, and H. Strobel, “Fairytaylor: A Multimodal Generative Framework for Storytelling,” arXiv:2108.04324, September 2021. <https://doi.org/10.48550/arXiv.2108.04324>
- [11] J. J. Y. Chung, W. Kim, K. M. Yoo, H. Lee, E. Adar, and M. Chang, “TaleBrush: Sketching Stories with Generative Pretrained Language Models,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*, New Orleans: LA, pp. 1-19, April 2022. <https://doi.org/10.1145/3491102.3501819>
- [12] Presidential Archives of Korea. Former Presidents’ Website Records [Internet]. Available: http://18children.president.pa.go.kr/mobile/our_space/travles.php.
- [13] Smilegate-ai Github. Korean Unsmile Dataset [Internet]. Available: https://github.com/smilegate-ai/korean_unsmile_dataset.
- [14] T. Kang, E. Kwon, J. Lee, Y. Nam, J. Song, and J. Suh, “Korean Online Hate Speech Dataset for Multilabel Classification: How Can Social Science Improve Dataset on Hate Speech?” arXiv: 2204.03262, 2022. <https://doi.org/10.48550/arXiv.2204.03262>
- [15] Github. Naver Sentiment Movie Corpus v1.0 [Internet]. Available: <https://github.com/e9t/nsmc>.
- [16] Kakaobrain Github. PORORO: Platform Of neuRAl mOdels for natuRAl language prOcessing [Internet]. Available: <https://github.com/kakaobrain/pororo>.
- [17] SKT Huggingface. ko-gpt-trinity-1.2B-v0.5 [Internet]. Available: <https://huggingface.co/skt/ko-gpt-trinity-1.2B-v0.5>.
- [18] SKT-AI github. KoBART Dataset Explanation [Internet]. Available: <https://github.com/SKT-AI/KoBART#data>.
- [19] Kakaobrain Github. Kogpt Dataset Explanation [Internet]. Available: <https://github.com/kakaobrain/kogpt>.
- [20] SKT-AI Github. KoGPT2 Dataset Explanation [Internet]. Available: <https://github.com/SKT-AI/KoGPT2#data>.
- [21] SKT Huggingface. Trainer [Internet] Available: https://huggingface.co/docs/transformers/main_classes/trainer#transformers.TrainingArguments.
- [22] H. S. Min, “A Study of Women’s Language in the Korean Language,” *The Journal of Asian Women*, Vol. 34, pp. 7-64, December 1995.
- [23] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans: LA, pp. 15-20, June 2018. <https://doi.org/10.18653/v1/N18-2003>
- [24] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. Kalai, “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*, Barcelona, Spain, pp. 4356-4364, December 2016.
- [25] Woodpecker. Wooddada Website [Internet]. Available: <https://wooddada.com/>.
- [26] RIKS. Webzine [Internet]. Available: <https://www.webzine.riks.or.kr/post/인공지능으로-문학하기-1-오영진>.



안지인 (Jiin An)

2019년 ~ 현 재: 서강대학교 아트엔테크놀로지학과, 컴퓨터공학과 학사과정

※ 관심분야 : 자연어 처리, 인공지능 윤리, 딥러닝, 인간-컴퓨터 상호작용



이육셋별 (Saetbyeol Leeyouk)

2021년 ~ 현 재: 서강대학교 아트엔테크놀로지학과 학사과정

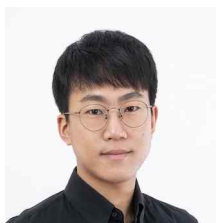
※ 관심분야 : 딥러닝(Deep Learning), AI 생성 모델(Generative AI), 인공지능 윤리, 인공지능경험디자인



장예원 (Yewon Jang)

2020년 ~ 현 재: 서강대학교 아트엔테크놀로지학과 학사과정

※ 관심분야 : 인간-컴퓨터 상호작용, 인터랙티브 미디어아트, 물리



정다샘 (Dasaem Jeong)

2015년 : KAIST 문화기술대학원(공학석사)

2020년 : KAIST 문화기술대학원(공학박사)

2020년 ~ 2021년: SK 텔레콤

2021년 ~ 현 재: 서강대학교 아트&테크놀로지학과 조교수

※ 관심분야 : 딥러닝(Deep Learning), 음악 인공지능(MusicAI), 음악 정보 검색(Music Information Retrieval)