Journal of KIIT. Vol. 23, No. 4, pp. 145-159, Apr. 30, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.4.145

KoBART 기반 지문 패러프레이징 기법을 적용한 배리어프리 영상 콘텐츠용 화면해설 생성 시스템 구현

이혜준*, 박준형**, 윤태원***, 조재하****, 허나영*****, 황대은******, 유길상*******

Implementation of an Audio Description Generation System for Barrier-Free Video Content using KoBART-based Script Paraphrasing

Haejune Lee*, Junhyeong Park**, Taewon Yoon***, Jaeha Jo****, Nayoung Heo*****, Dae-eun Hwang******, and Gil Sang Yoo********

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2023-00246191)

요 약

최근 글로벌 OTT서비스의 활성화로 인해, 장애인의 영상콘텐츠 접근성을 높이기 위한 배리어프리 영상 콘텐츠 제작의 필요성이 높아지고 있다. 본 연구에서는 시각장애인을 위한 배리어프리 영상 콘텐츠 제작 자동화를 위한 화면해설 생성 시스템을 제안하였다. 본 시스템은 영상의 주요 지문을 효과적으로 음성으로 전달할 수 있도록 Sentence-BERT와 KoBART 기반의 패러프레이징 기법을 제안하고, 이를 통해 시나리오 지문을 적절한 화면해설 형태로 패러프레이징 하는 모델을 구축하였다. 제안하는 시스템은 영상 콘텐츠의 다양한 상황에 맞춰 대사의 흐름을 방해하지 않으면서 자연스러운 화면 해설을 제공하며, 기존의 수작업 화면해설 생성 방식에 비해 시간과 비용을 절감할 수 있다. 구현 결과, 본 시스템은 높은 정확성과 유용성을 보여주었으며, 향후 배리어프리 콘텐츠 제작에 기여할 수 있을 것으로 기대한다.

Abstract

The recent global expansion of OTT services has increased the need for creating barrier-free video content to enhance accessibility for individuals with disabilities. This study proposes an automated audio description generation system for producing barrier-free video content for visually impaired individuals. The proposed system utilizes a paraphrasing technique based on Sentence-BERT and KoBERT to effectively convert key script elements into audio descriptions. By constructing a model capable of paraphrasing scenario scripts into appropriate audio description formats, the system provides natural and context-sensitive descriptions that do not disrupt the flow of dialogue in the video content. Compared to traditional manual methods of creating audio descriptions, the proposed approach significantly reduces time and cost. Implementation results demonstrate high accuracy and practicality, suggesting that this system can contribute to the future production of barrier-free content.

Keywords
barrier-free, audio description, movie description, paraphrasing

* 고려대학교 미디어학부
- ORCID: https://orcid.org/0009-0000-3877-8369
** Stony Brook University Korea Computer science
- ORCID: https://orcid.org/0009-0004-6913-7862
*** 고려대학교 지능정보 SW아카데미
- ORCID: https://orcid.org/0009-0003-1511-2996
**** 건국대학교 경제학과
- ORCID: https://orcid.org/0009-0000-7938-733X
***** 서울과학기술대학교 전기정보공학과
- ORCID: https://orcid.org/0009-0008-9725-9479
****** 상명대학교 융합전자공학전공
- ORCID: https://orcid.org/0009-0004-0120-7457
******* 고려대학교 정보대학 정보창의교육연구소(교신저자)
- ORCID: https://orcid.org/0009-0002-1085-5355

· Received: Jan. 20, 2025, Revised: Mar. 12, 2025, Accepted: Mar. 15, 2025
· Corresponding Author: Gilsang Yoo
Creative Informatics & Computing Institute, 145 Anam-ro, Seongbuk-gu, Seoul, Korea
Tel.: +82-2-3290-1674, Email: ksyoo@korea.ac.kr

Ⅰ. 서 론

국내 영상 산업은 최근 몇 년간 급격한 성장을 이루었다. 특히 OTT(Over-The-Top) 서비스 산업의 성장으로 시간과 공간의 제약 없이 감상 가능한 콘텐츠가 다양해졌지만, 시청각 장애인은 콘텐츠 접근성에서 소외되고 있다. 시각정보나 청각정보 중 하나의 정보가 없으면 시청각 매체인 영상을 온전히 이해하기 어렵기 때문이다. 이러한 문제를 해결하기 위해 '배리어프리(Barrier-Free) 콘텐츠'가 중요하게 주목받고 있다. 배리어프리 콘텐츠는 시청각 장애인을 위한 폐쇄 자막과 화면해설(Audio description)로 구성된다[1]. 화면해설은 원본 영상에 성우의 음성을 추가하여 시각적 정보를 전달하며, 전맹뿐 아니라 저시력자의 원본 영상 감상에도 도움을 준다. 그러나 국내 배리어프리 화면해설 콘텐츠의 보급은 매우 미흡한 수준이다.

한국 영화 진흥 위원회의 영화관 입장권 통합 전산망에 따르면, 2023년 한 해간 2D 영화 1,168편이 상영되었지만, 그 중 배리어프리 영화(디지털 가치봄)는 22편에 불과하여, 전체의 0.02%에도 미치지 못하는 것으로 보고되었다[2]. OTT 플랫폼에서도 일부 플랫폼만이 제한적으로 화면해설 콘텐츠를 제공하고 있다. 특히 음성 화면해설의 경우, 폐쇄 자막보다도 보급 수준이 열악하다[3]. 2023년 한국콘텐츠진흥원에 제출된 컬처미디어랩의 연구에 따르면 국내 주요 OTT서비스 사업자인 넷플릭스, 디즈니 플러스, 왓챠, 쿠팡 플레이, 웨이브, 티빙 모두 폐쇄 자막 서비스를 지원하고 있으나, 화면해설 서비스는 넷플릭스와 디즈니 플러스에서만 지원하고 있다. 이마저도, 콘텐츠 수가 폐쇄 자막 대비 현저히 제한적인데, 2024년 4월 기준 넷플릭스에서는 200여편의 제한적인 수의 콘텐츠에서만 음성 화면해설 서비스를 지원하고 있다[4].

화면해설 콘텐츠의 제작은 전문 작가가 대본을 작성하고 이를 음성화하여 영상에 편집하는 과정을 포함하기에, 시간과 비용이 많이 소요되는 작업이다. 더욱이 현재 국내 화면해설 작가의 수는 약 50명에 불과해 콘텐츠 수요를 충족하기 부족하다[5]. 이러한 상황에서 인공지능을 활용한 자동화 시스템 도입이 점차 시도되고 있다. 음성화 단계에서의 자동화가 일부 이루어지고 있으나, 대본 작성 단계부터 비용과 시간을 절감할 수 있는 자동화 기술 개발이 필요하다[6].

본 연구는 자연어 처리 기술을 적용하고, KoBART(Korean Bidirectional and Auto-Regressive Transformers)와 Sentence-BERT(Bidirectional Encoder Representations from Transformers)기반 패러프레이징 기법을 활용하여 음성 화면해설을 자동화하는 방법론을 제안하였다. 방송통신위원회와 Netflix등 글로벌 OTT 플랫폼에서는 화면해설 제작을 위한 규정을 제시하고 있다. 본 연구는 그 중, 전맹을 기준으로, 대사를 방해하지 않으면서, 간결하고 명확한 설명이, 영화의 흐름에 맞추어 제공되어야 한다는 주요한 가이드라인을 기반으로 하였다.

본 논문의 구성은 다음과 같다. 제2장에서는 화면해설 기술과 대본 패러프레이징 관련 기존 연구를 설명한다. 제3장에서는 KoBART 기반 지문 패러프레이징 기법과 구현 과정을 다룬다. 제4장에서는 제안된 시스템의 성능을 평가하고 학습 및 검증 결과를 분석한다. 마지막으로, 제5장에서는 본 연구의 결론과 향후 과제를 제시하였다.

Ⅱ. 관련 연구

2.1 영화 화면 해설

영화 화면해설은 영화의 시각적 요소를 설명하는 나레이션으로, 영화 이해 능력을 보완하는 역할을 수행한다[1]. 화면해설은 주로 시각장애인 관객을 위해 개발된 것으로, 숙련된 화면해설 작가가 영화를 분석하고 시각적 정보를 서술하는 방식으로 제작된다. 최근 제공되는 영화 화면해설 콘텐츠의 양이 증가하고 있는데, 이는 시각장애인을 위한 사회적 지원 확대와 법적 요구사항 준수의 결과로 볼 수 있다.

화면해설은 시각적 장면, 이전 화면해설 내용, 영화 속 자막 등 다양한 맥락에 영향을 받으며, 시각장애인이 영화를 이해할 수 있도록 돕는 문장들로 구성된다. 이 과정에서 화면해설은 간단한 묘사 위주의 이미지 또는 비디오 캡셔닝과 차별화된 몇 가지 특징을 가진다[7].

첫째, 화면해설은 시간의 흐름에 따라 중요한 시각적 요소를 밀도 있게 설명해야 한다. 둘째, 화면해설은 원래의 음성 트랙과 별도로 제공되며, 이를 보완하는 역할을 한다. 예를 들어, 대사나 배경음과 같이 음성 트랙으로 전달되는 정보는 설명할 필요가 없으며, 대사와 겹치지 않는 시간 간격 내에서 설명이 제공되어야 한다. 셋째, 화면해설은 밀도 높은 비디오 캡셔닝과 달리 스토리텔링을 목표로 하며, 등장인물의 이름, 감정, 행동 등을 포함하는 설명을 제공한다.

영화 화면해설의 장기적인 목표 중 하나는 장편 영화의 내용을 더욱 효과적으로 이해할 수 있도록 하는 것이다. 이를 위해 등장인물의 얼굴 및 음성을 통한 식별[8]-[10], 행동 및 상호작용 인식[11], 관계 파악[12], 3D 자세 분석[13] 등 다양한 연구가 진행되고 있다. 그러나 스토리 맥락을 완전히 이해하는 궁극적인 목표를 달성하기 위해서는 여전히 많은 어려움과 해결해야 할 과제가 존재한다.

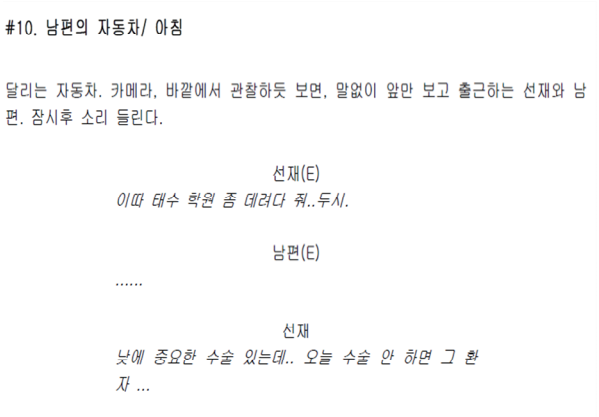
화면해설은 영화의 시청 경험을 확장하고 장애인 관객의 접근성을 높이는 데 기여한다. 그러나 대사와 음향 효과, 배경음악 등 기존의 청각적 요소를 해치지 않으면서 시각적 정보를 효과적으로 전달해야 한다는 제약이 따른다. 또한, 화면해설 제작 과정은 높은 수준의 전문성을 요구하며, 시간과 비용 부담이 크다는 한계가 있다. 이에 따라, 화면해설의 제작 효율성을 높이고 품질을 유지하기 위한 자동화 기술 개발의 필요성이 강조되고 있다.

2.2 영화 시나리오 패러프레이징 연구

그림 1에서와 같이, 영화 시나리오는 캐릭터의 대사와 지문으로 구성되어 있으며, 스토리의 맥락과 감정 표현을 전달하는 중요한 텍스트 자원이다. 자연어 처리 분야에서 패러프레이징(Paraphrasing)은 입력 문장의 의미를 보존하면서 단어와 구조를 변형하여 새로운 출력 문장을 생성하는 작업을 의미한다. 페이스북에서 개발한 BART는 Seq2Seq (Sequence-to-Sequence) 학습을 수행하는 모델로, 손상된 텍스트를 원래대로 복원하는 과정을 통해 학습되며 패러프레이징 작업에 효과적으로 활용된다.

시나리오 패러프레이징은 영화 시나리오의 표현을 변경하되, 원래의 의미를 유지하면서 문장을 재구성하는 작업이다. 이는 번역, 요약, 또는 창작 활동에서 중요한 역할을 하며, 특히 다음과 같은 분야에서 효과적으로 활용될 수 있다[14].

(1) 중복 콘텐츠 제거: 동일한 정보를 다양한 표현으로 제공하여 콘텐츠 중복을 방지한다.
(2) 다국어 자막 제작: 영화의 감정과 맥락을 유지하면서 다양한 언어로 번역한다.
(3) 대화 모델 개발: 영화 대사를 기반으로 문맥 기반의 자연스러운 대화 생성한다.


#10. 남편의 자동차/ 아침

달리는 자동차. 카메라, 바깥에서 관찰하듯 보면, 말없이 앞만 보고 출근하는 선재와 남편. 잠시후 소리 들린다.

선재(E)
이따 대수 학원 좀 데려다 줘..두시.

남편(E)
.....

선재
낮에 중요한 수술 있는데.. 오늘 수술 안 하면 그 환자 ...

그림 1. 영화 시나리오 예시 <분홍신>中
Fig. 1. Example of part of movie scenario from <The Red Shoes>

기존 연구에서는 영화 시나리오를 분석하여 주요 주제나 캐릭터 간의 관계를 추출하거나 요약하는 데 중점을 두었다. 특히, 자연어 처리 기술을 활용하여 다음과 같은 작업이 주로 이루어졌다[15].

(1) 감정 분석: 대사 및 지문에서 감정을 추출하여 캐릭터 심리와 스토리 분위기를 분석한다.
(2) 문맥 기반 대화 모델 생성: 영화 대사를 학습하여 자연스러운 대화 생성 및 캐릭터 간 상호작용을 모방한다.
(3) 시나리오 요약: 긴 대본을 간략화하여 주요 사건과 흐름을 파악한다.

이와 같은 연구는 주로 영문 대본을 대상으로 이루어졌으나, 최근 한국어 대본 처리에 대한 관심이

증가하고 있다. 특히, 한국어 특유의 문법 구조와 문화적 맥락을 반영한 패러프레이징 연구가 중요하게 다뤄지고 있다. 본 연구는 이러한 흐름을 기반으로, 영화 시나리오의 패러프레이징 기술을 발전시키기 위한 구체적인 방법론을 제안하였다.

2.3 KoBART를 활용한 자연어 생성 연구

KoBART는 Facebook의 BART(Bidirectional and Auto-Regressive Transformers)를 기반으로 한 한국어 자연어 처리(NLP) 특화 사전학습 언어 모델로, 문장 재구성, 요약, 번역 등 다양한 자연어 생성(NLG, Natural Language Generation) 작업에서 강점을 보인다. KoBART는 대규모 한국어 코퍼스를 활용한 사전학습을 통해 복원 능력과 생성 품질이 뛰어나며, 다양한 응용 분야에서 활용 가능성이 높다.

기존 연구에서 KoBART는 크게 다음 세 가지 주요 영역에서 활용되고 있다[14].

(1) 문서 요약: 뉴스 기사나 논문과 같은 긴 텍스트에서 핵심 정보를 추출하고 이를 간결하게 요약하는 작업에 활용한다.
(2) 대화 생성: 챗봇 응답 생성과 같이 문맥에 적합한 자연스러운 대화를 생성한다.
(3) 패러프레이징(Paraphrasing): 원문과 동일한 의미를 가진 새로운 문장을 생성하여 데이터 증강이나 자연어 생성 성능을 향상시키는 작업에 기여. 이 작업은 특히 데이터 부족 문제를 해결하거나 모델의 학습 성능을 강화하는 데 중요한 역할을 수행한다.

이 중 패러프레이징과 관련된 연구는 문장 수준에서 긴 문단 단위에 이르기까지 다양한 형태로 진행되어 왔으며, KoBART는 높은 유연성과 정확성을 보여준다. 그러나 비정형 데이터(Non-standard data)에 대한 연구는 상대적으로 부족하다. 예를 들어, 영화 시나리오와 같은 비정형 데이터에 대한 KoBART의 성능을 평가하거나 활용한 사례는 드물며, 이러한 데이터에서 모델이 얼마나 효과적으로 작동할 수 있는지에 대한 체계적인 검토가 필요하다. 이를 통해 KoBART의 한계를 파악하고, 향후 연구에서 이를 개선하거나 새로운 응용 분야를 개척할 수 있는 기반을 마련할 수 있다.

2.4 영화 시나리오 패러프레이징 연구의 한계

기존의 패러프레이징 연구는 주로 텍스트 데이터의 패러프레이즈 추출, 감지, 생성에 초점이 맞추어져 있었으며, 영화 시나리오와 같은 특정 도메인의 텍스트를 활용한 자연어 처리 모델의 개발은 제한적으로 이루어졌다[16]-[18]. 패러프레이즈 추출 기술은 대규모 패러프레이즈 말뭉치를 구축하는 데 유용하며, 이는 이후 패러프레이즈 감지 및 생성 모델의 데이터 소스로 활용될 수 있다. 특히 패러프레이즈 감지는 텍스트의 의역을 식별함으로써 언어 데이터의 다양성을 강화하며, 이를 통해 다양한 자연어 처리 응용 분야에 기여할 수 있다. 이러한 기술은 자연어 처리 모델의 언어적 다양성에 대한 처리 능력을 평가하는 주요 척도가 되어 왔다.

그러나 한국어 영화 시나리오의 경우, 언어적 특성과 텍스트의 맥락적 복잡성으로 인해 기존 연구에서 잘 다루어지지 않았다. 한국어 영화 시나리오에는 다음과 같은 독특한 특징이 있다. 첫째, 비형식적이고 구어체적인 표현이 자주 등장하며, 둘째, 대화의 문맥 의존성이 높고, 셋째, 대사와 함께 복잡한 감정 전달이 요구된다. 이러한 특성은 일반적인 패러프레이징 모델이 효과적으로 작동하기 어렵게 만든다. 더욱이, 단순히 문장의 구조를 재구성하는 것을 넘어 영화 시나리오의 창의적 특성을 유지하면서도 의미의 일관성을 보장해야 하는 추가적인 과제가 존재한다.

이에 따라, 본 연구는 한국어 영화 시나리오에 특화된 패러프레이징 기술을 KoBART를 활용하여 구현하였다. 특히, 기존 연구에서 다루지 않았던 시나리오의 지문과 화면해설 간의 관계를 학습하도록 모델을 설계하였다. 이를 통해 시나리오의 지문을 화면해설 형태로 변환하는 기술을 개발하였으며, 이는 화면해설 대본 작성 및 영상 콘텐츠 제작 과정에서 활용 가능성을 제시한다. 본 접근은 한국어 영화 대본 패러프레이징의 한계를 극복하고, 해당 도메인에 적합한 자연어 처리 모델을 제안하는 데 활용할 것으로 기대된다.

III. 제안한 지문 패러프레이징 기법

영화의 설계도로 볼 수 있는 시나리오는 화면해설과 여러 측면에서 공통된 특징을 보인다. 예를 들어, 소리 정보인 대사와 화면 정보인 지문이 서로 긴밀히 연관되어 반복적으로 나타나며, 지문에는 등장인물의 행동, 장소, 표정 등이 간결하면서도 명료하게 서술된다. 이러한 지문은 과도한 미사여구 없이 간결하게 작성되고, 화면의 전개 흐름과 일치하도록 기술된다. 또한, 시나리오는 장면별 구분이 명확하여 각 장면의 구조를 명료하게 이해할 수 있도록 돕는다.

본 논문에서는 이러한 특성을 활용하여 시나리오와 영화 영상을 입력으로 받아 화면해설이 포함된 영상을 자동으로 제작할 수 있는 시스템을 제안한다. 특히, 본 연구는 촬영과 편집 과정에서 발생하는 시나리오와 최종 영화 간의 차이를 보정하고, 시나리오에 포함된 소리 정보, 미사여구 등 원형 그대로 사용하기 어려운 요소를 적절히 수정하여 화면해설 형태로 변환(패러프레이징)하는 방법론을 제안한다. 제안된 시스템의 전체 구성은 그림 2와 같다. 구체적으로, 영화와 시나리오의 주요 요소를 패러프레이징 요소로 활용한다. 영화에서는 대화(인물의 발화) 사이의 공백과 대화 내용(인물의 발화 내용)을 주요 정보로 활용하고, 시나리오에서는 대사와 지문을 주요 입력으로 사용한다. 다음으로, 유사도 비교를 통해 영화의 인물 발화 내용과 시나리오의 대사를 대응시킴으로써 영화를 기준으로 시나리오를 필터링하여, 영화화 되지 않은 시나리오의 대사를 제거한다. 이 과정에서 대응된 영화 대사 전후에 공백이 존재하고, 대응된 시나리오 대사 전후에 지문이 존재하는 구간을 화면해설 삽입 가능 구간으로 판단한다. 이후, 이러한 구간의 시나리오 지문을 화면해설에 적합한 형태로 패러프레이징하고, 이를 음성화하여 영상에 오버레이함으로써 화면해설이 포함된 영화를 자동으로 제작한다.

본 논문에서는 필름 메이커스 커뮤니티에 개인 학습용으로 공유된 105편의 시나리오를 분석과 지문 데이터 수집에, 해당 시나리오를 기반으로 제작된 영화 7편을 분석과 시스템 검증에 활용하였다. 이를 통해 영화와 시나리오 간 차이를 분석하고, 지문 패러프레이징 기법의 설계 및 평가를 위한 데이터를 구축하였다.

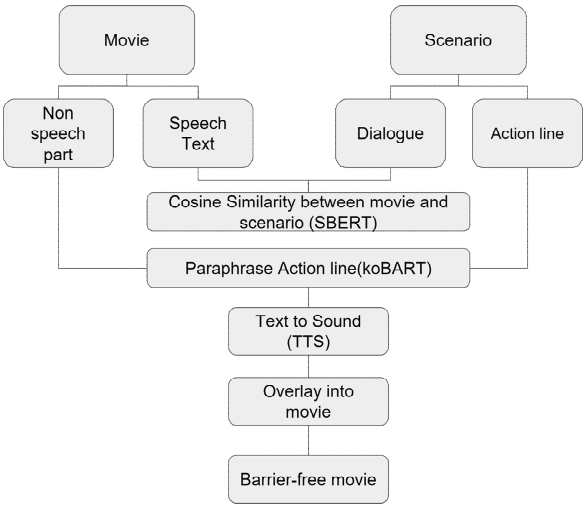


그림 2. 전체 시스템 구조도
Fig. 2. Overall system architecture

3.1 비발화 구간 및 대화 텍스트 추출

시나리오의 대사와 실제 영화의 발화를 활용하여 시나리오와 영화 간의 차이를 보정하고, 최종적으로는 시나리오의 지문이 영화의 어떤 부분에 해당하는 지문인지를 확인하기 때문에, 영화에 등장하는 인물의 발화 내용을 텍스트 형태로 변환하여 저장하는 작업이 수행되어야 한다. 동시에 인물의 음성이 없어 음성 화면해설이 삽입될 수 있는 비발화 구간의 시간 정보(타임스탬프)도 저장되어야 한다.

인물의 음성과 기타 소리가 혼재된 오디오 파일 내에서 인물의 음성이 등장하는 구간을 파악하는 발화감지(Speech activity detection)기술이 있다[19][20]. 본 논문에서는 영화 오디오상에서 인물의 발화가 등장하는 구간을 감지하기 위한 모델로 Pyannote를 사용하였다[21]. 영화 영상이 입력되면, Pyannote의 발화 감지 기술을 이용해 인물의 음성이 등장하는 발화구간과 그렇지 않은 비발화구간으로 구분된다. 이는 화면해설이 삽입될 수 있는 비발화 구간의 시간 정보 저장을 위한 과정이다. 검출된 발화 구간 내 대사의 내용은 정확한 텍스트로 변환되어야 한다.

이를 위해 음성을 텍스트 형태로 변환하는 STT(Speech-to-Text) 기술 중 OpenAI에서 개발한 Whisper 모델을 사용하였다. Whisper는 한국어를 포함한 97개의 언어를 지원하며, 노이즈가 섞인 음성 데이터와 일상 대화 데이터 등 다양한 환경에서 학습되었다. 이러한 특성 덕분에 Whisper는 소음 환경에서도 높은 전사(Transcription) 성능을 보여 영화 오디오와 같은 복잡한 음향 데이터를 처리하는 데 적합하다.

본 연구에서는 Whisper 모델 중 약 7억 6천 9백만 개의 매개변수를 가진 Whisper-medium 모델을 사용하였다. Whisper-medium은 모델 크기와 성능의 균형 면에서 효율적이며, 대규모 데이터 처리에 있어 적합한 정확도를 제공한다. 특히, 본 연구에서는 Whisper 모델이 비발화 구간을 처리할 때 배경음악, 효과음 등의 영향으로 텍스트를 출력하는 오류(예: 아, 어 등을 출력하는 오류)를 최소화하기 위한 추가적인 방법론을 적용하였다. 이 과정은 시스템의 전반적인 정확도를 향상시키는 데 기여하였다.

발화 감지를 통해 분리된 비발화 구간의 시간 정보는 영화 딕셔너리(Dictionary)에 저장된다. 발화 구간으로 분류된 음성 데이터는 Whisper-medium 모델에 입력되며, 이 모델은 대사의 내용을 텍스트로 변환하여 영화 딕셔너리에 저장한다. 딕셔너리는 각 대사의 등장 시간과 텍스트 내용을 포함하며, 이는 대사 분석 및 추가적인 처리 작업을 위한 기반 데이터를 제공한다. 최종적으로 비발화 구간 딕셔너리, 영화 딕셔너리에 저장되는 값은 각각 표 1, 표 2와 같다.

표 1. 비발화 구간 딕셔너리 예시
Table 1. Non speech part dictionary example

[{'speak1': 19.80281, 'speak2': 26.09719}...]

표 2. 영화 딕셔너리 예시 (영화 <너의 결혼식> 中)
Table 2. Movie dictionary example(Movie <Your Wedding>)

[{'start': 8.68219, 'end': 10.84219, 'text': '내 인생은 불가능의 연속들이었다.'}, {'start': 13.17094, 'end': 14.53781, 'text': '하지만...'}, ...]

3.2 시나리오 대사 및 지문 파싱

105편의 시나리오를 분석한 결과, 대부분의 시나리오는 배역, 대사, 지문을 명확히 구분해야 하는 시나리오의 특성을 충실히 반영하고 있음을 확인하였다. 구체적으로, 시나리오 작성에서 배역과 대사를 분리하는 방식으로 뉴라인(`\n`), 탭(`\t`), 콜론(`:`) 중 하나를 주로 사용하는 경향이 나타났다. 이러한 규칙적 패턴은 시나리오 작성 방식이 일관성을 유지하며, 이를 기반으로 시스템화가 가능하다는 점을 시사한다.

사용자가 입력한 docx 형태의 시나리오 파일에서 이러한 패턴을 자동으로 분석하고 처리한다. 입력된 파일 내 문장 구조를 탐색하여, 뉴라인(`\n`), 탭(`\t`), 콜론(`:`) 각각의 타입별 문장 개수를 계산한 후, 가장 많은 문장이 해당하는 타입을 시나리오의 주된 형식으로 판단한다. 이 판단을 바탕으로, 해당 형식에 맞게 장면 헤드라인, 지문, 대사가 자동으로 구분되고 파싱될 수 있도록 하였다.

파싱에 사용된 정규식은 표 3과 같다. 이를 통해 사용자는 다양한 형식의 시나리오 파일을 입력하더라도 일관된 분석 결과를 얻을 수 있으며, 시나리오의 구조적 특성에 기반한 효율적인 데이터 처리가 가능하다.

표 3. 파싱에 사용된 정규식
Table 3. Regular expression used for parsing

Type	Regular expression
\n	r'^([가-힣]{1,5}[0-9]?[A-z]?[a-z]?\s?)(\(V\.O\.\))?(\(V\.O\))?(\(v\.o\.\))?(\(v\.o\))?\s?(\(NA\))?(\(Na\))?(\(na\))?(\(N\))?(\s*)?\n{1,}(\s*)?([가-힣A-Za-z0-9\s\(\)\,\?\!\.\…]*)'
\t	r'^([가-힣]{1,5}[0-9]?[a-z]?\s?)(\(V\.O\.\))?(\(V\.O\))?(\(v\.o\.\))?(\(v\.o\))?\s?(\(NA\))?(\(Na\))?(\(na\))?(\(N\))?(\s*)?\t{1,}(\s*)?([가-힣A-Za-z0-9\s\(\)\,\?\!\.\…]*)'
:	r'^([가-힣]{1,5}[0-9]?[a-z]?\s?)(\(V\.O\.\))?(\(V\.O\))?(\(v\.o\.\))?(\(v\.o\))?\s?(\(NA\))?(\(Na\))?(\(na\))?(\(N\))?(\s*)?\t{1,}(\s*)?([가-힣A-Za-z0-9\s\(\)\,\?\!\.\…]*)',lines)

최종적으로 시나리오 딕셔너리는 text, person, start, end, flag, sn의 값을 가지며, 씬 헤드라인, 지문, 대사에 따라 표 4와 같이 구성된다.

표 4. 씬헤드라인, 지문, 대사별 딕셔너리 구성
Table 4. Configuring dictionaries by scene heading, action lines, and dialogues

Type / Key	Scene heading	Action lines	Dialogue
text	Scene number	Action lines (contents)	Dialogue (contents)
person	none	none	name of character
start	0	0	0 (After comparing similarity: Starting time)
end	0	0	0 (After comparing similarity: Ending time)
flag	1	2	0
sn	Scene number	Scene number	Scene number

이때 text 값이 원문 시나리오에 서술된 순서 그대로, 하나의 시나리오 리스트(List)가 구성된다. 3.1과 3.2의 과정을 거치면 비발화 딕셔너리, 영화 딕셔너리와 시나리오 딕셔너리에는 각각 표 1, 표 2, 표 5와 같은 내용이 저장된다.

표 5. 시나리오 딕셔너리 예시(영화 <너의 결혼식> 中)
Table 5. Scenario dictionary example(From movie <Your Wedding>)

{'text': '1. 타이틀 백', 'person': 'none', 'start': 0, 'end': 0, 'flag': 0, 'sn': '1'}
{'text': '부아앙~ OO중학교 교문을 들어서는 오토바이 뒤 우편물 자루 속 자두색 봉투', 'person': 'none', 'start': 0, 'end': 0, 'flag': 2, 'sn': '1'}
{'text': '가만있는 애들을 왜 자꾸 시비 걸어서 패 재껴? , 'person': '담임', 'start': 0, 'end': 0, 'flag': 1, 'sn': '6'}

3.3 영화 발화와 시나리오 대사간의 유사도 비교

촬영과 편집을 거치며 최종 영화와 시나리오 간의 차이가 발생하기에, 대사를 활용한 유사도 비교를 통해 영화를 기준으로 시나리오를 필터링 하여, 화면해설로 사용될 수 있는 지문을 추출하는데 활용하였다.

자연어 처리 기술에서 문장 간 유사성을 판단하는 작업은 중요한 연구 주제 중 하나이다. 이러한 유사도 비교 작업에는 주로 Google에서 개발한 BERT(Bidirectional Encoder Representations from Transformers) 모델이 활용된다[22]. BERT는 Transformer의 인코더 구조를 기반으로 하며, 양방향 학습을 통해 문맥을 이해하고, 자연어의 의미를 보다 깊이 분석할 수 있다. 특히, 문장 유사도 비교 작업에 최적화된 SBERT(Sentence-BERT)는 BERT 모델을 확장하여 문장 간 유사성을 측정하는 데 높은 성능을 보여준다[23]. 본 연구에서는 SBERT의 한국어 모델인 jhgan/ko-sbert-sts를 사용하여 영화 대사와 시나리오 대사의 유사도를 분석하였다[24]. 해당 모델은 사전 학습된 한국어 데이터와 코사인 유사도를 통해 정밀한 문장 유사도 계산이 가능하다.

유사도 측정을 위해 코사인 유사도(cosine similarity)를 사용하였다. 이는 두 벡터의 각도를 기반으로 유사도를 정량적으로 측정하는 방법으로, 두 벡터가 완전히 동일한 방향을 가질수록 값이 1에 가까워지고, 반대로 완전히 다른 방향일 경우 -1에 가까워진다. 코사인 유사도 식 (1)를 기반으로 영화 대사와 시나리오 대사의 벡터 값을 비교하여 유사도를 평가하였다.

$$\cos(\Theta) = A \cdot B \,/\, \|A\| \cdot \|B\| \qquad (1)$$

이때 '반응어', '응답표현', '인사말' 등은 반복적으로 사용될 가능성이 높아 유사도 분석에서 잡음으로 작용할 수 있다. 예를 들어, "안녕"과 같은 인사말은 시나리오와 영화에서 여러 번 반복될 가능성이 크지만, "내가 오죽하면 장례 치르다 나오겠냐"와 같은 대사는 이야기의 흐름 상 고유한 의미를 가지고 있으며 반복될 가능성이 낮다. 잡음으로 작용될 수 있는 대표적인 반응어, 응답표현, 인사말로는 표 6과 같은 표현들이 있다.

한국어 대화에서 반응어와 응답표현은 대부분 1~5음절의 짧은 형태이다.

표 6. 응답표현 및 인사말 예시
Table 6. Examples of reaction words and greetings

ex) 안녕하세요 / 안녕하십니까 / 안녕 / 그래? / 그래요? / 정말? / 정말로? / 정말이야? / 정말이에요? / 진짜? / 진짜로? / 네 / 예 / 어 / 응 / 그래요 / 그래 / 알겠어 / 알겠어요 / 맞아 / 맞어

영상 콘텐츠 이해를 위한 시각 정보에 대한 청각적 정보 제공이라는 화면해설의 특성상, 부족한 해설보다 잘못된 내용을 동반한 과잉해설이 심각한 오류이기에, 잡음 제거의 기준을 보수적으로 설정해, 공백 제외 5 캐릭터 미만의 대사는 유사도 비교 대상에서 제외하였다. 이와 같은 전처리 과정을 통해 영화 대사와 시나리오 대사의 유사도를 더 정확하게 분석할 수 있다. 그다음 단계에서는 시각 정보에 대한 청각적 정보 제공이라는 화면해설의 특성 등을 고려하여, 장면을 해설하지 않는 오류보다 잘못된 해설을 하는 오류인 과잉 해설을 줄이기 위해 2회에 걸쳐 유사도 비교를 진행하였다.

영화는 흐름이 있는 스토리를 가지고, 대사도 흐름의 한 구성요소이다. 이를 고려하면, 그림 3에서와 같이 연속된 영화의 발화(Speech) ①~⑦중 ①~③가 시나리오 S#1의 대사(Dialogue)와 유사하다고 연속적으로 매칭 되다가, 도중에 대사④만 시나리오 S#4의 대사와 매칭이 되고, 다시 대사⑥, ⑦은 시나리오 S#1의 대사와 매칭이 된다면, 다른 2개의 케이스 중 (1)의 경우가 대다수를 차지한다.

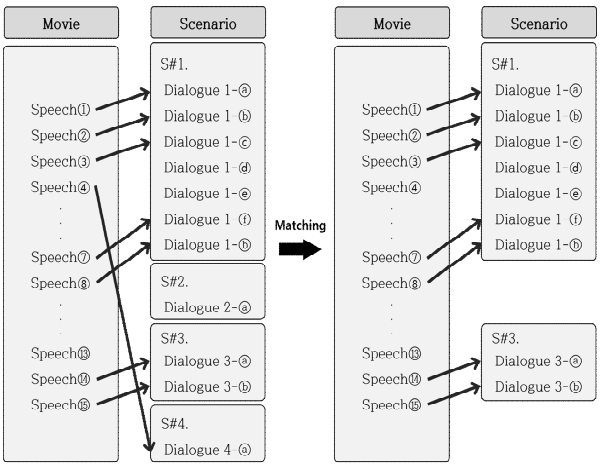


그림 3. 유사도 비교 작업 예시
Fig. 3. Cosine similarity compare task example

(1) 대사④가 시나리오에 존재하지 않는 대사(애드리브 등)인 경우,

(2) 촬영과 편집을 거쳐 시나리오 상에서의 씬(Scene)의 순서가 1-4-1로 변경된 경우

표 7. 음성 화면 해설 가능 판단 원시코드
Table 7. Voice screen commentability determination pseudo code

```
<Input>
# S : Dictionary of Dialogues in Scenario
# D : Dictionary of Speech in Movie
# threshold : a value for threshold
similarities <- dictionary of length len(D)
refined_results <- array

for i <- 1 to len(D):
  for j <- 1 to len(S):
      similarities.append(j, scene number, cosinesim(S,
D), D.sentence, S.sentence)
      best_match = max(similarities, key=lambda x: x[3])
      if best_match[3] > threshold:
          refined_results.append((D[i],best_match[2],
best_match[1], best_match[3]))
          used_scene_indices.add(best_match[0])
          # best_match[0] : index
          # best_match[1] : scene number
          # best_match[2] : selected sentence
          # best_match[3] : cosine similarity value
        current_scene <- None
          count <- 0
          filtered_results <- array
for i <- 1 to len(refined_result):
    if refined_result[i][2] == refined_result[i+1][2]:
            if refined_result[i][2] != current_scene:
                        current_scene =
refined_result[i][2]
                count = 2
            else:
                count += 1
      else:
            if count >= 2:

filtered_results.extend(refined_results[i-count+1:i+1])
              current_scene = None
              count = 0
  if count >= 2:
      filtered_results.extend(refined_results[-count:])
return filtered_results
```

이와 같이 대사가 시나리오에 존재하지 않는 경우를 고려하여 1차 유사도 비교에서는 시나리오 대사 전체와 영화 대사 전체 간의 유사도를 비교하였다. 이때 연속된 영화 대사가 시나리오에서 같은 씬(S#1, S#3)의 대사와 연속적으로 2회 이상 매칭된 경우를 파악하였다. 해당하는 시나리오의 씬만 선별하여 2차 유사도 비교 작업을 진행한다.

유사하다고 판단된 시나리오 대사와 영화 대사가 각각 속한 딕셔너리를 활용해, 화면해설이 진행될 수 있는 케이스를 추출할 수 있다. 표 7에서와 같이 입력에 대하여 2가지 조건을 모두 충족한 경우를 음성 화면해설이 가능한 케이스로 판단한다.

3.4 지문 패러프레이징을 통한 화면해설 텍스트 생성

화면해설이 가능한 경우라 하더라도 시나리오의 특성상 다음과 같은 한계가 존재한다. 첫째, 시나리오 지문이 지나치게 길 경우, 비발화 구간 내에서 해당 지문을 모두 해설하기가 어려울 수 있다. 둘째, 시나리오 지문은 종종 완전하지 않은 문장을 포함하거나, 소리 정보를 직접적으로 서술하는 등 화면해설로 바로 활용하기에 적합하지 않은 형태로 작성되어 있다. 따라서 시나리오 지문을 원형 그대로 화면해설로 사용하기에는 한계가 있으며, 지문을 화면해설 형태로 변환하기 위한 적절한 패러프레이징(Paraphrasing) 과정이 필수적이다. 본 연구에서는 사전 학습된 한국어 기반의 BART 모델인 koBART를 활용하여 시나리오 지문을 화면해설에 적합한 형태로 패러프레이징하는 작업을 진행하였다.

Ⅳ. 실험 결과

4.1 데이터 수집 및 증강

실험에 사용된 데이터는 필름메이커스 커뮤니티에서 개인 학습용으로 공유된 다양한 장르의 시나리오 105편에서 수집된 총 9,744개의 지문으로 구성되었다[25]. 본 연구의 목적은 시나리오와 영화 간의 대사를 기반으로 한 비교를 수행하는 것이므로, 시나리오 내에서 대사와 대사 사이에 위치한 지문 전체를 하나의 독립적인 지문으로 간주하였다.

데이터셋은 시나리오 지문과 화면해설 형태로 패러프레이징된 문장이 쌍을 이루는 방식으로 구축되었다. 이를 위해 넷플릭스 화면해설 규정과 방송통신위원회의 화면해설 가이드라인을 참고하여, 연구자가 지문을 수정하고 이를 기반으로 레이블링을 수행하였다. 이러한 과정을 통해 데이터가 화면해설의 규격에 부합하도록 조정하였다.

모델의 성능 향상을 위해 데이터 증강(Data augmentation) 기법을 적용하였다. 구체적으로, 동의어 교체 방식을 활용하여 오리지널 텍스트와 패러프레이즈 텍스트 간의 데이터 증강을 진행하였다. 이 방식은 텍스트 내 등장하는 '인물'의 이름을 활용하여 데이터를 변형하는 과정으로, 원본 문장과 패러프레이즈된 문장 간의 골드 레이블링을 유지하면서도 추가적인 데이터를 생성할 수 있는 효과적인 방법이다.

동의어 교체 방식의 구현은 개체명 인식(NER, Named Entity Recognition) 기법을 사용하여 텍스트에서 '인물'로 인식된 토큰을 추출하는 것으로 시작한다. 이후, 오리지널 텍스트와 패러프레이즈 텍스트에서 동일하게 추출된 인물명이 존재할 경우, 이를 200개의 임의로 생성된 이름 중 하나로 랜덤하게 대체하였다. 예를 들어, 오리지널 텍스트와 패러프레이즈 텍스트에서 동일하게 '지우'와 '혜인'이라는 이름이 등장하면, 해당 이름을 각각 '지환'과 '태윤'으로 변경하여 데이터를 증강하였다. 이와 같은 데이터 증강 과정을 통해 총 34,006개의 데이터셋으로 확장하였다. 이를 학습 데이터 확보와 일반화 성능 평가의 균형을 맞추기 위하여 8:2의 비율로 분할하여, 27,204개의 훈련 데이터와 6,802개의 검증 데이터를 구성하였다. 데이터 증강 기법은 모델의 성능을 높이는 데 핵심적인 역할을 했으며, 특히 골드 레이블을 유지하며 데이터의 다양성을 확보함으로써 학습 효율을 극대화할 수 있다.

4.2 koBART 모델 학습

BART-base 모델의 인코더 파라미터를 프리징(Freezing)한 상태에서 학습을 진행하였다.

프리징된 파라미터는 사전학습(Pretraining) 단계에서 학습된 가중치를 그대로 유지하며, 디코더 및 기타 학습 가능한 레이어에 대해서만 업데이트를 수행하였다. 이를 통해 학습 효율성을 높이고, 제한된 데이터에서 과적합(Overfitting)을 방지하는 데 초점을 맞췄다.

학습에 사용된 주요 하이퍼파라미터 세팅은 표 8과 같다.

표 8. 하이퍼파라미터 세팅
Table 8. Hyperparameter setting

Parameter	Value	Unit
Epochs	80	count
Learning rate	3e-5	unitless
Batch size	64	samples
Weight decay	0.01	unitless
total_num_steps	34,080	steps

Validation loss가 상승할 때 학습이 자동으로 종료되도록 조기 종료 조건을 설정하여, 불필요한 학습 과정을 줄이고 최적의 모델을 도출하였다. 학습 과정에서 에포크(epoch)에 따른 훈련 손실(Training loss)과 검증 손실(Validation loss)의 변화를 시각화한 결과는 그림 4와 같다.

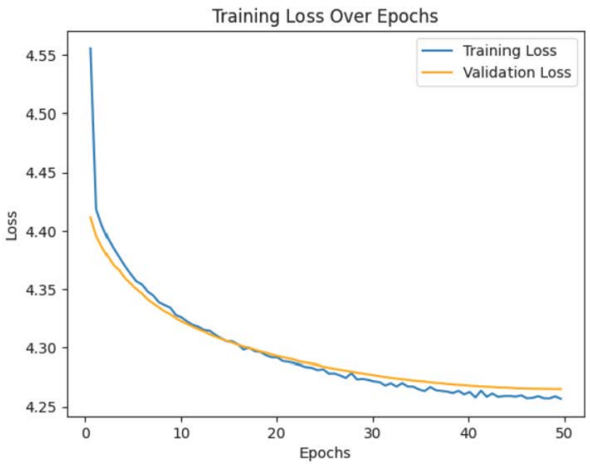


그림 4. 모델 학습 결과
Fig. 4. Result of model training

초기 에포크에서는 훈련 손실과 검증 손실이 모두 높은 값을 보이며, 모델이 데이터 패턴을 학습하는 초기 과정을 반영하였다. 그러나 학습이 진행됨에 따라 두 손실 값은 지속적으로 감소하였고, 이는 모델이 점진적으로 데이터를 효과적으로 학습하며 성능이 향상되고 있음을 나타낸다. 특히, 훈련 손실과 검증 손실 간의 값 차이가 크지 않고, 두 손실이 유사한 패턴으로 감소한 점은 학습 과정에서 과적합이 발생하지 않았음을 시사한다. 이는 프리징 전략과 적절한 하이퍼파라미터 조정이 효과적이었음을 보여준다. 약 70 에포크 이후에는 손실 감소 폭이 완만해지며 수렴하는 경향을 보였고, 이는 모델이 충분히 학습되었음을 나타낸다. 이 시점 이후로는 추가적인 학습이 모델의 성능에 큰 영향을 미치지 않는 것으로 판단된다. 결론적으로, 이러한 학습 결과는 제안된 모델이 안정적으로 학습되었으며, 훈련 데이터와 검증 데이터 모두에서 높은 일반화 성능을 보이는 모델임을 뒷받침한다. 본 실험을 통해 BART-base 모델의 인코더를 프리징한 접근법이 효율적인 학습과 더불어 안정적인 성능을 확보하였음을 알 수 있다.

4.3 지문-화면해설 패러프레이징 모델

koBART를 활용하여 시나리오 지문을 화면해설에 적합한 형태로 패러프레이징한 결과는 표 9와 같다. 중요한 정보의 손실 없이, 간결하지만 완전한 문장 형태의 화면해설로 패러프레이징 된 것을 알 수 있다.

표 9. koBART 기반 지문 패러프레이징한 결과
Table 9. Paraphrased result based on koBART

Action line (Original text)	Audio description text (Paraphrase text)
동구, 일어나 밖으로 나가면서, 밖에는 동구의 친구 기명이 동구에게 반갑게 손을 흔들며 인사하고 있다.	동구가 일어나 밖으로 나간다. 밖에는 동구의 친구 기명이 손을 흔들며 인사하였다.

4.4 화면해설 텍스트의 음성화 및 영상 오버레이

TTS(Text-To-Speech)를 활용하여 음성을 생성한 후, 이를 영상에 오버레이하여 화면해설 음성이 포함된 영상을 제작할 수 있다. 텍스트를 음성으로 변환하는 대표적인 TTS 모델로는 구글 텍스트 음성 변환(Google Text-to-Speech), 아마존 폴리(Amazon Polly), 마이크로소프트 애저 음성 서비스(Microsoft Azure Speech Service), 그리고 MeloTTS 등이 있다.

구글 텍스트 음성 변환은 높은 품질의 음성 출력을 제공하며, 다국어 지원과 함께 다양한 플랫폼에서 사용할 수 있는 유연성이 특징이다.

아마존 폴리는 딥러닝 기술을 활용하여 인간의 목소리와 유사한 자연스러운 음성을 합성하는 데 중점을 두며, 사용자가 음성 스타일을 선택할 수 있는 기능을 제공한다. 마이크로소프트 애저 음성 서비스는 다양한 언어와 음성을 지원하며, 음성 특성을 커스터마이징할 수 있어 특정 시나리오에 맞춘 음성 제작이 가능하다. 이 중에서, MeloTTS는 MIT와 MyShell.ai에서 개발한 고성능 다중 언어 TTS 라이브러리로, 영어, 스페인어, 프랑스어, 중국어, 일본어, 한국어 등 다양한 언어를 지원하며 CPU 기반 실시간 추론이 가능하다. 특히 MeloTTS는 경량화된 설계로 고품질 음성을 생성하면서도 효율적인 자원 활용이 가능한 점에서 차별성을 지닌다.

본 연구에서는 MeloTTS를 활용하여 화면해설 음성을 생성하였다[26]. 화면해설 음성을 영상에 삽입하는 과정에서는 기존의 화면해설 영화 제작 규정을 준수하였다. 화면해설 음성이 삽입되는 구간에서는 영화의 원본 소리를 적절히 감소시켜, 화면해설 음성이 명확하게 전달될 수 있도록 조정하였다. 이를 통해 시각적 정보가 제한된 사용자들에게도 영상의 내용을 효과적으로 전달할 수 있도록 하였다.

4.5 구현결과

시나리오 파일과 영화 영상을 업로드하여 화면해설 영화를 생성할 수 있는 웹 어플리케이션은 그림 5와 같다.

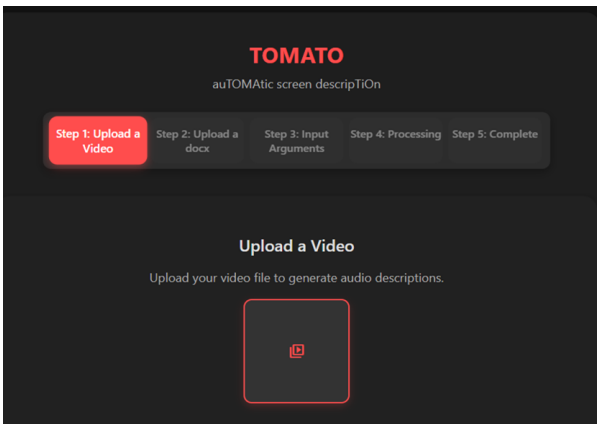


그림 5. 웹 페이지 화면
Fig. 5. Web page screen

사용자가 시나리오 파일과 영화 영상 파일을 업로드 하면, 영화의 알맞은 부분에 지문이 적절히 패러프레이징된 화면해설 음성이 삽입되어, 영상의 시각 정보를 청각적으로 수용할 수 있다. 그림 6은 영화 분홍신의 시나리오로, 학생2가 학생1을 밀치고 신발을 빼앗는 장면이다. 표 10은 시스템을 거쳐 출력된 영상의 화면과 음성을 기재한 표이다. 그림 6의 시나리오에서 지문에 해당하는 부분은 영화에서 대사 없이 인물의 행동으로만 표현된다. 표 10에서 알 수 있듯, 시스템을 거치면, 해당 부분에 음성 화면해설이 포함되어, 청각 정보만으로도 상황을 이해할 수 있게 만들어준다.

여학생의 어깨를 낚아채는 손. 깜짝 놀라 돌아보면, 서있는 친구.

여고
깜짝이야. 떨어질뻔 했잖아!

여학생의 말은 아랑곳 않은채 분홍신에 멍하니 시선주는 친구.

여고
내가 방금 줏은거야. 여기 있길래..

친구
(계속해서 물끄러미 바라보다가) 벗어.

여고
...?

친구
이거..내가 먼저 봤어. 내가 가질래.

여고
(어이없다) 여기서 삼십분 동안 너 기다렸어.

갑자기 확-여학생을 밀치는 친구. 여학생 넘어지면, 신겨진 분홍신 강제로 벗겨내는 친구.

여고
야! 뭐하는거야?

그림 6. 시스템 입력 시나리오 일부 <분홍신> 中
Fig. 6. Example of part of system input from scenario <The red shoes>

이외에도 7편의 영화에 대하여 테스트가 진행되었으며, 구현 결과는 유튜브 링크(https://youtu.be/o9jGf4geg3Y)를 통해 확인할 수 있다.

표 10. 시스템 출력 영상 <분홍신> 中
Table 10. Example of system output video <The red shoes>

Video	
Dialogue	학생1: 떨어질 뻔 했잖아!
Video	 
Dialogue	-
Generated audio description	여학생의 말은 아랑곳하지 않고 분홍신을 멍하니 바라본다.
Video	
Dialogue	학생1: 아 이거? 방금 주운거야 여기 있길래
Video	
Dialogue	학생2: 벗어. 이거 내가 먼저 봤어
Video	
Dialogue	학생1: 야, 여기서 너 30분 동안 기다렸어
Video	 
Dialogue	-
Generated audio description	여학생을 밀친다. 여학생이 넘어지고 친구가 분홍신을 강제로 벗겨낸다.

V. 결론 및 향후 과제

본 연구는 글로벌 OTT 서비스의 활성화와 함께 필요성이 증가하는 장애인의 영상 콘텐츠 접근성을 개선하기 위한 방안을 모색하며, 특히 시각장애인을 위한 배리어프리 화면 해설 제작 자동화 기술 개발에 초점을 맞추었다. 이를 위해 KoBART 기반의 패러프레이징 기법을 활용하여 시나리오의 주요 지문을 효과적으로 텍스트로 변환하고, 이를 기반으로 자연스러운 화면해설을 생성하는 시스템을 제안하였다.

제안된 시스템은 화면해설을 음성 형태로 제공하여 시각장애인의 콘텐츠 접근성을 대폭 향상시키는 동시에, 수작업으로 진행되던 화면해설 제작 과정에 비해 시간과 비용을 절감하는 효과를 기대할 수 있다. 실험 결과, 본 시스템은 높은 정확도와 유용성을 입증하였으며, 배리어프리 콘텐츠 제작의 효율성을 증대시켜 장애인뿐만 아니라 비장애인을 포함한 다양한 사용자에게 혜택을 제공할 가능성을 확인하였다. 화면해설 제작 기술의 자동화는 금전적 및 시간적 제약으로 인해 제작이 어려웠던 기존에 제작된 콘텐츠의 배리어프리 콘텐츠로의 변환을 확대하여, 시각장애인들이 더 많은 배리어프리 콘텐츠에 접근할 수 있는 환경을 조성할 것으로 기대한다. 이는 시각장애인의 콘텐츠 접근권을 확대할 뿐만 아니라, 사회 전반적으로 포용성과 접근성을 증진시키는 데 기여한다.

향후 연구에서는 영상처리 기술을 활용하여 시나리오에 포함되지 않은 시각적 정보를 분석하고 이를 화면해설에 반영함으로써, 더욱 정교한 화면해설 자동 생성 시스템을 구축하고 배리어프리 콘텐츠의 품질과 접근성을 더욱 향상시켜 나갈 계획이다.

Reference

[1] T. Han, M. Bain, A. Nagrani, G. Varol, W. Xie and A. Zisserman, "AutoAD: Movie Description in Context", 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 18930-18940, Jun. 2023. https://doi.org/10.1109/CVPR52729.2023.01815.

[2] Korean Film Council, KOREA Box-office Information System, https://www.kobis.or.kr/kobis/business/stat/them/findShowTypeShareList.do. [accessed: Dec. 28, 2024]

[3] Korea Creative Content Agency, "Strategies to Enhance Content Distribution for Better Media Accessibility of Individuals with Visual and Hearing Impairments", Korea Create Content Agency, Nov. 2023. https://www.kocca.kr/kocca/bbs/view/B0000147/2004436.do?searchCnd=&searchWrd=&cateTp1=&cateTp2=&useYn=&menuNo=204153&categorys=0&subcate=0&cateCode=&type=&instNo=0&questionTp=&ufSetting=&recovery=&option1=&option2=&year=&morePage=&qtp=&domainId=&sortCode=&pageIndex=1. [accessed: Mar. 03, 2025]

[4] Netflix, https://www.netflix.com/. [accessed: Jan. 05, 2025]

[5] S. Gang, "Do You Know 'Audio Description Writers,' the People Who Show the World Through Words?", Hankyoreh, Nov. 2022. https://www.hani.co.kr/arti/culture/book/1065307.html. [accessed: Jan. 08, 2025]

[6] CJ OLIVENETWORKS, "CJ OliveNetworks Provides 'AI Voice Cloning' Technology for Audio Description Broadcasts on tvN", Jul. 2022. https://www.cjolivenetworks.co.kr/news/press_release/detail/595?ca=ALL. [accessed: Jan 5, 2025]

[7] J. Na, "A study of directing way to “commentaring screen” and “barrier-free movie” on broadcast", The Korea Contents Association Review, Vol. 11, No. 1, pp. 54-58, Mar. 2013.

[8] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic, "Finding actors and actions in movies", 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, pp. 2280-2287, Dec. 2013. https://doi.org/10.1109/ICCV.2013.283.

[9] A. Brown, E. Coto, and A. Zisserman, "Automated video labelling: Identifying faces by corroborative evidence", 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), Tokyo, Japan, pp. 77-83, Sep. 2021. https://doi.org/10.1109/MIPR51284.2021.00019.

[10] Q. Huang, W. Liu, and D. Lin, "Person search in videos with one portrait through visual and temporal links", Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11217, pp. 437-454, Jul. 2018. https://doi.org/10.1007/978-3-030-01261-8_26.

[11] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating visual representations from unlabeled video", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 98-106, Jun. 2016. https://doi.org/10.1109/CVPR.2016.18.

[12] A. Kukleva, M. Tapaswi, and I. Laptev, "Learning interactions and relationships between movie characters", 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 9846-9855, Jun. 2020. https://doi.org/10.1109/CVPR42600.2020.00987.

[13] G. Pavlakos., E. Weber, M. Tancik, and A. Kanazawa, "The One Where They Reconstructed 3D Humans and Environments in TV Shows", Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, Vol. 13697, Tel Aviv, Israel, pp 732-749, Oct. 2022. https://doi.org/10.1007/978-3-031-19836-6_41.

[14] D. Zeng, H. Zhang, L. Xiang, J. Wang, and G. Ji, "User-Oriented Paraphrase Generation With Keywords Controlled Network", IEEE Access, Vol. 7, pp. 80542-80551, Jun. 2019. https://doi.org/10.1109/ACCESS.2019.2923057.

[15] H. Yu, D. Son, J. Yang, A. Lee, S. Oh, and J. Kim, "A Study on Performance Analysis of Question Generation based on Korean Pretrained Language Model", 2023 14th International Conference on Information and Communication

Technology Convergence (ICTC), Jeju Island, Korea, pp. 1239-1241, Oct. 2023. https://doi.org/10.1109/ICTC58733.2023.10393015.

[16] J. Ganitkevitch, B. V. Durme, and C. Callison-Burch, "PPDB: The paraphrase database", Proceedings of NAACL-HLT, pp. 758-764, Jun. 2013. https://aclanthology.org/N13-1092/. [accessed: Fab. 25, 2025]

[17] L. Tian, N. Hui, L. Kong, K. Chen, H. Qi, and Z. Han, "Sentence para-phrase detection using classification models", Forum for Information Retrieval Evaluation, Lecture Notes in Computer Science, Vol. 10478, Jan. 2018. https://doi.org/10.1007/978-3-319-73606-8_13.

[18] J. Berant and P. Liang, "Semantic parsing via paraphrasing", Proceedings of the 52nd ACL, Baltimore, MD, USA, Vol. 1, pp. 1415-1425, Jun. 2014. https://doi.org/10.3115/v1/P14-1133.

[19] T. Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe, "End-to-End Automatic Speech Recognition Integrated with CTC-Based Voice Activity Detection", ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 6999-7003, May 2020. https://doi.org/10.1109/ICASSP40776.2020.9054358.

[20] J. Svirsky and O. Lindenbaum, "SG-VAD: Stochastic Gates Based Speech Activity Detection", 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5, Jun. 2023. https://doi.org/10.1109/ICASSP49357.2023.10096938.

[21] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization", Interspeech 2023, Dublin, Ireland, pp. 3222-3226, Aug. 2023. https://doi.org/10.21437/interspeech.2023-205.

[22] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv:1810.04805v2, Oct. 2018. https://doi.org/10.48550/arXiv.1810.04805.

[20] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks", Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, pp. 3982-3992, Nov. 2019. https://doi.org/10.18653/v1/D19-1410.

[21] jhgan, "ko-sbert-sts", https://huggingface.co/jhgan/ko-sbert-sts. [accessed: Fab. 25, 2025]

[22] Filmmakers, https://www.filmmakers.co.kr/ [accessed: Jan. 08, 2025]

[23] H. Choi, J. Bae, J. Lee, S. Mun, J. Lee, and H. Cho, "Mels-Tts : Multi-Emotion Multi-Lingual Multi-Speaker Text-To-Speech System Via Disentangled Style Tokens", ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, pp. 12682-12686, Apr. 2024. https://doi.org/10.1109/ICASSP48485.2024.10446852.

저자소개

이 혜 준 (Haejune Lee)

2024년 6월 : 고려대학교 지능정보 SW아카데미 4기 수료(640H)
2020년 3월 ~ 현재 : 고려대학교 미디어학부 학사과정
관심분야 : 데이터사이언스, 미디어데이터분석

박 준 형 (Junhyeong Park)

2024년 6월 : 고려대학교 지능정보 SW아카데미 4기 수료(640H)
2019년 2월 ~ 현재 : Stony Brook University, Computer Science 학사과정
관심분야 : NLP, 멀티모달

윤 태 원 (Taewon Yoon)

2024년 6월 : 고려대학교 지능정보 SW아카데미 4기 수료(640H)
2022년 2월 ~ 현재 : 중앙대학교 미래교육원 소프트웨어 디자인 학사과정
관심분야 : UX, UI, 디자인, 데이터 분석, 컴퓨터 비전, NLP, 딥러닝

조 재 하 (Jaeha Jo)

2017년 3월 ~ 2024년 2월 : 건국대학교 경제학과(학사)
2024년 6월 : 고려대학교 지능정보 SW아카데미 4기 수료(640H)
관심분야 : 데이터분석, 경제학

허 나 영 (Nayoung Heo)

2024년 6월 : 고려대학교 지능정보 SW아카데미 4기 수료(640H)
2022년 2월 ~ 현재 : 서울과학기술대학교 전기정보공학과 학사과정
관심분야 : 데이터분석, 머신러닝, 딥러닝, 전기정보공학

황 대 은 (Dae-eun Hwang)

2018년 3월 ~ 2024년 2월 : 상명대학교 융합전자공학전공(학사)
2024년 6월 : 고려대학교 지능정보 SW아카데미 4기 수료(640H)
관심분야 : 컴퓨터 비전, 딥러닝, NLP

유 길 상 (Gil Sang Yoo)

2010년 2월 : 중앙대학교 영상공학과(박사)
2010년 3월 ~ 현재 : (사)한국컴퓨터게임학회 부회장
2011년 3월 ~ 현재 : 고려대학교 정보대학 정보창의교육연구소/지능정보 SW아카데미 교수
2023년 3월 ~ 현재 : (사)한국미디어 아트산업협회 수석부회장
관심분야 : 데이터사이언스, 데이터 시각화, 빅데이터 분석, 3D영상 콘텐츠, 머신러닝, 딥러닝, 컴퓨터교육