The Transactions of the Korea Information Processing Society
TKIPS Vol.14, No.12, pp.1004~1015

eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2025.14.12.1004

MLSQ: A Multimodal-based System for Learning Material Summarization and Question Generation

Geonwoo Yu^† · Sangyoon Lee^† · Jinyoung Ahn^† · Minha Woo^† · Sugyeong Kim†† · Jungoo Lee† · Hyeonwoo Choi† · Yaeran Kim††† · Woonghee Lee††††

ABSTRACT

While the proliferation of digital learning environments has increased the use of diverse multimedia materials, this often leads to passive learning. Existing text-based automatic question generation technologies are insufficient to overcome this limitation. Therefore, this study proposes an AI-based system (MLSQ) that integrates and analyzes video lectures and written materials. This system precisely fuses text data extracted via OCR and STT, along with handwriting information detected from the lecture video as a key emphasis point, based on temporal information and inputs it into a Large Language Model (LLM) to summarize learning content and automatically generate both short-answer and multiple-choice questions. Performance evaluation results showed that the proposed multimodal fusion method demonstrated improved performance over single-modal approaches, with a maximum increase of 3.4%p in BLEU score and 3.1%p in ROUGE-L score. Furthermore, in a 5-point Mean Opinion Score (MOS) user evaluation, the system demonstrated its educational practicality and effectiveness by achieving high scores above 4.0 in all categories, with 'Speech Conversion Reliability' and 'Learning Suitability of Question Generation' receiving an average of 4.33.

Keywords : Multimodal, Automatic Question Generation, Learning Material Summarization, Large Language Model, Optical Character Recognition

MLSQ: 멀티모달 기반 학습 자료 요약 및 문제 생성 시스템

유 건 우^† · 이 상 윤^† · 안 진 영^† · 우 민 하^† · 김 수 경†† · 이 준 구† · 최 현 우† · 김 예 란††† · 이 웅 희††††

요 약

디지털 교육 환경의 보편화에 따라 다양한 멀티미디어 학습 자료가 활용되고 있으나, 이는 종종 학습자의 수동적인 정보 습득으로 이어지는 한계를 가진다. 기존의 텍스트 기반 자동 문제 생성 기술 역시 이러한 한계를 극복하기에는 부족하다. 이에 본 연구는 동영상 강의와 서면 자료를 통합 분석하는 멀티모달 기반 학습 자료 요약 및 문제 생성 시스템(MLSQ)을 제안한다. 이 시스템은 OCR과 STT로 추출된 데이터뿐만 아니라, 강의 영상에서 탐지된 필기 정보를 핵심 강조점으로 활용하여 이 모든 데이터를 시간 정보를 기준으로 정밀하게 융합하고, 이를 대형 언어 모델(LLM)에 입력하여 학습 내용을 요약하며 주관식 및 객관식 문제를 자동 생성한다. 성능 평가 결과, 제안된 멀티모달 융합 방식은 단일 모달 방식 대비 BLEU 점수에서 최대 3.4%p, ROUGE-L 점수에서 최대 3.1%p 향상된 성능을 보였다. 또한, 5점 척도의 사용자 만족도 평가(MOS)에서 음성 변환 신뢰도 및 문제 생성의 학습 적합성 항목이 평균 4.33점을 기록하는 등 모든 항목에서 4.0 이상의 높은 점수를 획득하여 제안 시스템의 교육적 실용성과 효과성을 입증하였다.

키워드 : 멀티모달, 자동 문제 생성, 학습 자료 요약, 대형 언어 모델, 광학 문자 인식

1. 서 론

기존의 학습은 학습자가 동영상 자료와 서면 자료에 의존하여 수동적으로 정보를 습득하고 암기에 의존하는 형태로 주로 이루어진다. 이러한 학습 방식은 학습 후 정보의 빠른 망각을 유발하며, 실제 문제 해결 능력의 향상에 한계가 있다. 그에 따라, 보다 능동적이고 심화된 학습을 유도할 수 있는 새로운 학습 방식의 필요성이 대두되고 있다. 이러한 한계

※ This research was financially supported by Hansung University.
^ These authors contributed equally to this work.
† 비 회 원 : 한성대학교 AI응용학과 학사과정
†† 비 회 원 : 한성대학교 IT융합공학부 학사과정
††† 비 회 원 : 한성대학교 AI응용학과 석사과정
†††† 정 회 원 : 한성대학교 AI응용학과 조교수
Manuscript Received : Semtember 8, 2025
Accepted : October 19, 2025
* Corresponding Author : Woonghee Lee(whlee@hansung.ac.kr)

를 극복하기 위해 문제 생성 기술이 주목받고 있으나, 기존 기술은 주로 단일 모달 데이터를 기반으로 하고 있어 동영상과 문서 자료를 통합적으로 활용하는 데 제약이 있다. 이러한 기술적 한계는 다양한 형태의 학습 자료를 종합적으로 이해하는 데 한계가 되며, 특히 학습자가 다양한 정보를 결합하여 다각적으로 사고하도록 유도하는 데 부족함이 있다. 또한, 기존 시스템은 객관식 문제 생성에 치중하여 학습자의 창의적 및 비판적 사고를 충분히 자극하지 못하는 문제가 있다. 따라서 다양한 형태의 학습 자료를 복합적으로 분석하고 이를 기반으로 심층적 사고를 촉진하는 문제를 생성할 수 있는 시스템이 필요하다.

최근 AI 기술의 발전은 이러한 문제의 해결 가능성을 열어주고 있다. OCR(Optical Character Recognition) 기술을 활용하면 PDF에서 텍스트를 추출할 수 있고, 객체 탐지(Object Detection) 기술은 동영상 내 시각적 정보를 분석하며, STT(Speech-to-Text)[1-4] 기술을 활용하여 강의 음성을 텍스트로 변환할 수 있다. 이러한 멀티모달 데이터를 처리하는 기술들의 융합은 다양한 학습 자료들을 통합적으로 분석하고 활용할 수 있는 시스템의 구현을 가능하게 하며, 궁극적으로 학습자에게 심화된 문제 해결 능력을 제공하는 보다 효과적인 문제 생성 및 학습 지원을 기대할 수 있다. 이를 바탕으로, 본 연구에서는 위와 같은 기존 학습 방식과 문제 생성 기술의 한계를 극복하기 위해 멀티모달 데이터를 활용한 AI 기반 요약 및 문제 생성 시스템인 MLSQ(Multimodal Learning-material Summarization & Question-generation)를 제안한다. 제안된 시스템은 동영상과 서면 자료를 통합적으로 분석하여 학습 내용을 요약하고, 객관식 및 주관식 문제를 생성함으로써 학습자에게 맞춤형 학습 환경을 제공한다. 이를 통해 학습자는 학습 내용을 보다 장기적으로 기억하고, 실제 문제 해결 능력을 강화할 수 있다. 제안된 시스템은 특정 학습 환경에 국한되지 않고 다양한 교육 환경에서 활용될 수 있다. 예를 들어, 대학 강의 자료에 적용하여 강의 내용을 요약하고 관련 문제를 생성함으로써 학생들의 심화 학습을 지원할 수 있다. 또한, 의료 교육과 같은 현장 중심 직업 훈련에 적용되어 복잡한 절차를 요약하고 학습 문제를 생성함으로써 학습 효과를 높일 수 있다. 더 나아가, 청각 장애인을 포함한 학습 취약계층을 위한 학습 보조 시스템으로 활용되어 음성 기반 강의 자료를 텍스트화하고 요약하여 학습 접근성을 향상시킬 수 있다. 이와 같은 다양한 응용 사례는 본 시스템의 유연성과 실용성을 입증한다.

본 논문은 다음과 같이 구성된다. 2장에서는 관련 연구를 검토하여 기존 기술의 한계와 본 논문에서 제안하는 기술과의 차별점을 분석한다. 3장에서는 제안하는 시스템의 설계와 구조를 설명한다. 4장에서는 시스템 구현 및 성능 평가 결과를 제시하며, 마지막으로 5장에서는 결론과 향후 연구 방향을 논의한다.

2. 관련 연구

자동 문제 생성(Automatic Question Generation, AQG)은 교육 기술 분야에서 활발히 연구되어 왔으며, 주로 텍스트 데이터를 기반으로 다양한 형태의 질문을 자동으로 생성하는 방식이 중심을 이루어 왔다[21][19]. 이러한 방식은 사실 기반 질문 생성에는 효과적이나, 고차원적 사고를 요구하거나 학습자의 이해도를 평가할 수 있는 복합적 문제를 생성하는 데 한계가 있다.

이러한 한계를 극복하기 위해 최근에는 이미지, 표, 음성 등의 멀티모달 데이터를 활용한 문제 생성 접근이 주목받고 있다. OCR, 객체 탐지, 음성 인식 기술의 발전은 다양한 정보를 자동 추출할 수 있게 하였으며, 이를 교육용 콘텐츠에 적용하려는 시도가 이루어지고 있다. Zhan 등[22]은 시각 정보와 텍스트 정보를 함께 활용하여 멀티모달 문서로부터 질문-답변 쌍을 자동 생성하는 프레임워크를 제안하였다. 이 연구에서는 문서 내의 이미지와 문장을 연계하여 의미 있는 질의응답 데이터를 구성하고, BERT[9] 기반 모델을 활용하여 문맥에 맞는 질문을 생성하는 구조를 사용하였다. 하지만 해당 모델은 주로 정적인 인쇄 문서 형태의 데이터에 초점을 맞추고 있으며, 동영상이나 음성 정보는 고려되지 않았다. Liu 등[23]은 검색 기반 구조를 도입한 멀티모달 QA 시스템을 제안하였다. 이 시스템은 이미지-텍스트 쌍을 인덱싱하고, 자연어 질문에 대해 관련 이미지와 텍스트를 추출하여 응답하는 구조로 설계되었다. 이를 통해 질의응답 정확도가 향상되었지만, 해당 시스템은 정보 탐색 중심으로 설계되어 학습자 맞춤형 문제 생성을 지원하는 데는 제약이 있다. Wang 등[24]은 예제 기반 문제 생성 시스템에서 distractor(오답 선택지)를 자동 생성하는 알고리즘을 제안하여 객관식 문제 품질을 개선하였다. 특히 유사 의미를 가지되 정답이 아닌 선택지를 생성함으로써 학습자의 이해도를 더 효과적으로 측정할 수 있도록 하였다. 그러나 이 연구 역시 텍스트 기반에서 벗어나지 않으며, 학습자의 실제 학습 맥락을 반영한 문제 생성에는 한계가 있다.

추가적으로, Kim 등[19]이 한국어 BERT 기반의 딥러닝 자동 질문 생성 모델을 제안하였다. 해당 모델은 입력 문장을 바탕으로 WH-질문(What, When, Who 등 'Wh-'로 시작하는 의문사를 사용한 사실 기반 질문)을 생성하는 방식으로, 한국어 문장 구조에 맞춰 성능을 개선한 점에서 의의가 있다. 그러나 이 모델은 단일 문장 또는 문단 수준의 입력만을 처리할 수 있어, 문서 전체나 다양한 모달리티 간 연결을 다루는 데는 한계가 있다. 또한 Park 등[20]은 멀티모달 데이터를 활용한 개인 맞춤형 교육 콘텐츠 생성 방안을 제안하였으며, 이를 통해 학습자의 이해 수준에 따라 다양한 형태의 콘텐츠를 제공하는 구조를 설계하였다. 하지만 해당 연구 역시 정적인 이미지 및 텍스트 데이터에 초점을 두고 있으며, 동영상 강의나 시간 순서가 존재하는 시청각 정보의 통합 분석까지는 도달하지 못하였다.

본 연구는 기존 자동 문제 생성 연구들이 주로 단일 모달 정보에 기반하여 수행되었던 한계를 극복하고자, 문서, 영상, 음성 데이터를 통합적으로 처리할 수 있는 멀티모달 기반 문제 생성 시스템을 제안한다. 제안된 시스템은 문서 정보, 시각 정보, 음성 정보를 자동으로 추출한 후, 서로 다른 모달 정보를 통합 분석하여 학습 내용을 요약하고, 문맥에 부합하는 객관식 및 주관식 문제를 자동 생성할 수 있도록 구성되었다. 구체적으로, 시스템은 PDF 문서에서 OCR 기술을 통해 텍스트를 추출하고, 강의 영상에서 객체 탐지를 통해 주요 시각 요소를 분석하며, 음성 데이터는 Whisper[5] 기반의 음성 인식 모델을 활용하여 텍스트로 변환하였다. 이렇게 수집된 멀티모달 데이터는 대형 언어 모델(Large Language Model, LLM)을 통해 통합 분석되며, 자료 간의 문맥 연계를 고려한 요약 및 문제 생성을 수행한다.

본 연구의 주요 기여는 다음과 같다. 첫째, 기존 연구들이 정적인 단일 모달 기반 문제 생성에 집중한 데 비해, 본 연구는 동영상, 음성, 문서 데이터를 융합하여 처리하는 구조를 실현함으로써 학습 자료의 실제 활용 맥락에 부합하는 문제 생성 방식을 제안하였다. 둘째, 단순한 질의응답 생성이 아닌, 요약과 문제 생성을 연계한 학습 흐름을 구현함으로써 학습자의 맥락 이해와 응용 능력을 고려한 콘텐츠 구성이 가능함을 보였다. 셋째, 제안 시스템을 실제 학습 환경에서 수집한 데이터에 적용하고, 정량적 지표(BLEU, ROUGE 등)와 MOS(Mean Opinion Score) 기반 사용자 평가를 통해 실질적인 유효성을 검증함으로써, 단순 모델 제안 수준을 넘어 교육 현장에서의 실용성을 갖춘 시스템으로 발전 가능성을 확인하였다. 이러한 결과는 제안된 시스템이 기존 단일 모달 기반 문제 생성 방식의 한계를 보완함과 동시에, 멀티모달 AI 기술의 교육 분야 활용 가능성을 구체적으로 실증하였다는 점에서 의의가 있다.

3. MLSQ 처리 시스템

본 절에서는 제안 시스템의 처리 절차를 단계적으로 기술한다. 우선 입력된 멀티모달 데이터(텍스트, 시각, 음성)에 대한 데이터 추출 및 전처리 과정을 설명하고, 이어서 각 모듈의 출력 결과를 시간 정보를 기준으로 통합하는 멀티모달 융합 방식을 다룬다. 마지막으로, 통합된 데이터를 활용하여 대형 언어 모델 기반으로 요약 및 문제를 생성하는 과정을 제시한다.

3.1 제안 시스템의 전체적인 흐름

본 연구에서 제안하는 MLSQ 시스템은 서면 자료와 동영상 강의로부터 문자 및 시각 정보와 음성 정보를 동시에 추출하여 통합 분석하는 구조로 설계되었다. 시스템의 전체적인 처리 흐름은 크게 데이터 추출, 멀티모달 융합, 그리고 콘텐츠 생성의 세 단계로 구성되며, 각 단계는 유기적으로 연결되어 최종적으로 학습자 맞춤형 학습 내용 요약 및 평가 문제를 생성한다.

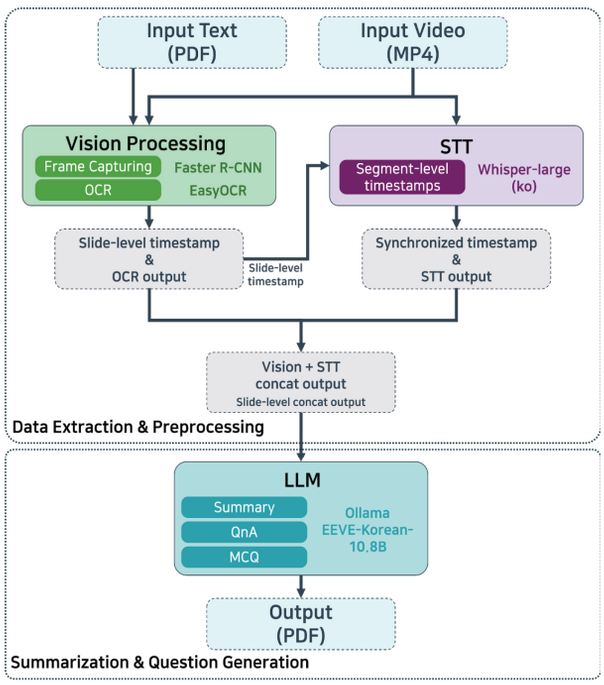


Fig. 1. Overall Flowchart of the Proposed System

Fig. 1과 같이, 입력된 동영상 파일(MP4)은 두 개의 병렬 처리 경로를 통해 분석되며, 입력 PDF는 Vision Processing 모듈에서 추가적으로 활용된다. 첫 번째 경로인 Vision Processing 모듈에서는 프레임 캡처를 통해 입력 동영상의 슬라이드 화면을 추출하고, 자체 학습된 Faster R-CNN[10]을 활용하여 프레임 내 필기를 인식한다. 또한 EasyOCR[11]을 적용하여 입력 PDF의 텍스트 정보를 인식한다. 이후 PDF 슬라이드와 캡처된 프레임을 Perceptual hash, Average hash, Difference hash의 평균값을 기반으로 매칭하며, 각 PDF 슬라이드에서 인식된 텍스트와 필기 정보를 병합한다. 이 과정에서 각 슬라이드의 타임스탬프 정보가 함께 저장되어 STT 모듈 결과와의 동기화 기준으로 활용된다.

두 번째 경로인 STT 모듈에서는 Whisper-large 모델을 사용하여 한국어 음성을 텍스트로 변환하며, 세그먼트 단위의 타임스탬프를 생성한다. 이후 Vision Processing에서 확보한 slide-level timestamp 정보를 기준으로 STT 출력 결과가 슬라이드별로 재정렬되어 각 슬라이드에 해당하는 음성 내용이 매칭되며, 두 처리 경로에서 생성된 결과는 타임스탬프를 기준으로 동기화되어 통합된다. 따라서 슬라이드별 OCR 결과와 해당 시간대의 STT 결과가 결합된 결과인 Vision + STT concat output이 확보되며, 이러한 멀티모달 데이터 융합을 통해 단순히 시각 또는 음성 정보만을 활용할 때보다 더욱 풍부하고 정확한 학습 내용 분석이 가능하다. 최종 단계에서는 통합된 멀티모달 데이터가 Ollama[6]의 EEVE-Korean-10.8B [7] 모델에 입력되어 세 가지 유형의 학습 자료가 생성된다. Summary 모듈을 통해 강의 내용의 핵심을 추출한 요약문이

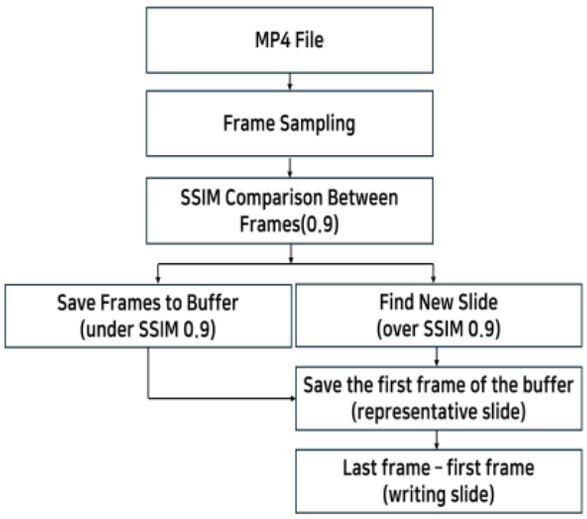


Fig. 2. Overall Flowchart of the Vision Processing

생성되고, QnA(Question and Answer) 모듈에서는 심층적 사고를 유도하는 주관식 문제와 답안이, MCQ(Multiple Choice Questions) 모듈에서는 5지선다 객관식 문제와 정답이 각각 생성된다. 생성된 모든 결과물은 PDF 형태로 출력되어 학습자에게 제공된다.

3.2 데이터 추출 및 전처리

1) 비전 프로세싱

본 연구에서는 필기 기반 강의를 분석하기 위해 비전 프로세싱 기법을 활용하였다. 특히, 강사가 강의 영상 내에서 슬라이드 화면 위에 필기한 부분을 추출하여, 동영상 강의의 핵심 내용을 포함한 주요 부분을 효과적으로 파악하고자 하였다. Fig. 2는 이러한 비전 프로세싱의 구조를 전체적으로 보여준다. 입력 데이터로는 슬라이드를 화면에 띄운 상태에서 그 위에 강사가 필기를 하며 강의한 영상을 사용하였으며, 주요 프레임을 자동으로 식별하는 기술을 적용하였다. 프레임 식별 과정은 크게 슬라이드 화면 추출, 필기 감지 단계로 구성되며, 주요 정보를 JSON 파일로 저장하여 이후 OCR 및 STT 단계에서 활용할 수 있도록 하였다.

먼저, 슬라이드 화면 추출 단계는 전체 강의 동영상에서 슬라이드가 전환되는 시점을 자동으로 탐지하고, 각 슬라이드 구간의 마지막 프레임을 추출하여 저장하는 과정이다. 슬라이드 화면을 추출하기 위해 구조적 유사도 지수(Structural Similarity Index, SSIM)를 활용하여 일정 간격의 프레임 간 차이를 측정한다. 이 과정에서는 먼저 비디오 파일을 프레임 단위로 변환한 후, 2초 간격으로 프레임을 샘플링한다. 이후 각 프레임을 그레이스케일로 변환한 뒤, 이전 슬라이드 프레임과 현재 프레임 간 SSIM을 계산하여 유사도를 측정한다. 측정된 SSIM 값이 0.9 미만일 경우 새로운 슬라이드 화면으로 판단하며, 그렇지 않으면 해당 프레임을 버퍼에 저장한다. 즉, SSIM 값이 0.9 이상인 동안, 새로운 슬라이드가 감지되기 전까지 수집된 프레임들은 하나의 버퍼에 저장된다. 이후 SSIM 값이 0.9 미만으로 변하는 순간, 버퍼에 저장된 프레임 중 첫 번째 프레임을 대표 슬라이드로 지정하고 버퍼를 초기화한다.

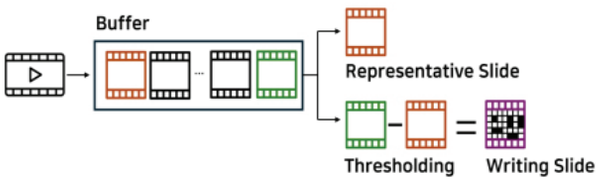


Fig. 3. Process of Extracting Slide Screens

다음으로, 필기 감지는 강의 영상에서 강사가 필기한 구간을 탐지하는 과정으로, 프레임 간 변화율을 분석하는 방식으로 수행된다. Fig. 3에서 보듯, 앞서 버퍼에서 저장된 프레임들 중 첫 번째 프레임과 마지막 프레임을 비교하여 필기된 영역을 추출하고 두 프레임 간 차이를 계산한 후 이를 이진화(Thresholding)하여 필기된 영역만을 강조한다. 이 과정에서 픽셀 값이 특정 임계값(50) 이상이면 흰색(255)으로, 그렇지 않으면 검정색(0)으로 설정하여 필기 영역을 분리한다. 만약 차이 이미지의 전체 픽셀이 흰색이라면 차이가 거의 없는 것이기 때문에 해당 슬라이드에는 필기가 없다고 판단하고, 그렇지 않은 경우엔 배경을 흰색으로 처리한 후 필기된 내용만을 남겨 저장한다. 최종적으로, 각 슬라이드와 필기 프레임은 PNG 형식의 이미지 파일로 저장되며, 해당 프레임이 발생한 시간 정보를 포함한 메타데이터는 JSON 파일로 기록된다. JSON 파일에는 비디오 이름, 슬라이드 번호, 각 슬라이드의 타임스탬프 정보가 포함된다.

Fig. 4는 실제 강의 동영상 중 일부 화면 및 앞서 설명한 과정을 통해 추출된 대표 슬라이드와 필기 슬라이드이다. 이러한 설계를 통해 강의 영상으로부터 슬라이드 화면과 필기 정보를 효과적으로 구분하고 저장할 수 있다. 특히, 슬라이드와 필기 내용이 한 번에 나타나는 프레임을 그대로 저장하는 것이 아니라, 필기 정보만 추출할 수 있도록 하여 이후 OCR 기법을 적용하는데 있어 더욱 효과적으로 필기를 인식할 수 있도록 한다.

Fig. 4와 같이 대표 슬라이드 프레임과 필기 영역 이미지가 성공적으로 분리되면, Vision Processing 모듈은 추출된 시각 정보와 입력된 PDF 문서 간의 정합성을 확보하고 텍스트 데이터를 통합하는 후속 단계로 진입한다. 이 과정은 3.1절에서 개괄한 바와 같이, 원본 PDF의 텍스트 추출, 비디오 프레임과의 매칭, 그리고 필기 정보의 탐지 및 가중치 부여 단계로 세분화된다.

첫째, 강의 자료로 활용된 원본 PDF 문서는 EasyOCR 라이브러리를 통해 모든 슬라이드의 인쇄된 텍스트 정보를 추출하는 전처리를 거친다. 이는 슬라이드의 기본적인 텍스트 콘텐츠를 확보하는 단계이다. 둘째, PDF의 정적 텍스트 정보와 동영상의 동적 프레임 정보를 연결하기 위해 슬라이드 매칭을 수행한다. 이 단계에서는 '슬라이드 화면 추출'을 통해

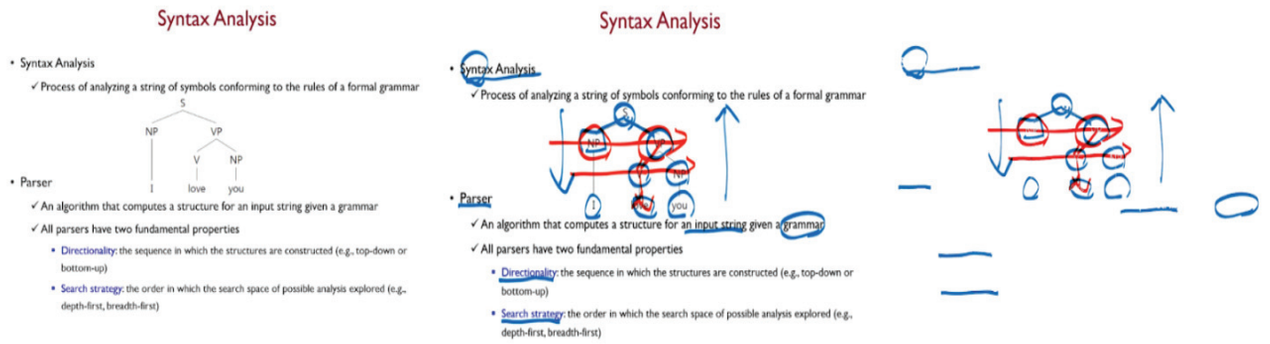


Fig. 4. Actual Results of Vision Processing

확보된 각 대표 슬라이드 프레임과 PDF의 모든 슬라이드 페이지 간의 시각적 유사도를 측정한다. 본 연구에서는 단일 해시 알고리즘의 한계점을 보완하기 위해 Perceptual hash, Average hash, Difference hash의 세 가지 방식을 병렬로 적용하고, 그 평균값을 최종 유사도 점수로 활용한다. 이 점수를 기반으로 가장 유사도가 높은 프레임-페이지 쌍을 식별하여, 영상의 타임라인과 PDF의 페이지 번호 간의 일대일 대응 관계를 확립한다. 셋째, Fig. 4의 오른쪽 이미지와 같이 분리된 필기 슬라이드 이미지는 필기 영역을 탐지(detection)하기 위해 특별히 구성된 Faster R-CNN 모델의 입력으로 사용된다. 이 객체 탐지 모델은 범용 데이터셋이 아닌, 본 연구의 목적에 맞게 자체적으로 구축한 데이터셋을 통해 학습되었다. 해당 데이터셋은 실제 강의 환경에서 빈번하게 발생하는 필기 패턴(예: 원, 별표, 밑줄, 화살표 등)을 단일 혹은 복수로 포함하는 이미지 약 2,000장으로 구성된다. 이를 통해 모델은 슬라이드 배경과 필기 정보를 명확히 구분하여 필기 영역의 바운딩박스(bounding box)를 정확하게 반환하도록 최적화되었다.

마지막으로, 추출된 모든 정보를 유기적으로 통합한다. 앞서 Fig. 4의 예시에서 확인된 바와 같이, 강의 영상 내 필기는 대부분 밑줄, 동그라미, 화살표 등이며, 이는 일반적으로 강사가 해당 지점에서 강조하고자 하는 의도를 나타낸다. 따라서 Faster R-CNN을 통해 탐지된 각 필기 바운딩박스는 단순한 이미지 좌표를 넘어, 이러한 '강조'의 의미를 파악하기 위한 의미론적 연결 과정을 거친다. 즉, 각 필기 바운딩박스의 중점(center point)을 기준으로, EasyOCR로 추출된 PDF 텍스트 블록 중 유클리드 거리가 가장 가까운 텍스트와 매핑된다. 이렇게 필기와 텍스트가 매핑되면, 내용 요약이나 문제 생성 단계에서 대형 언어 모델(LLM)에게 어떤 텍스트가 강조되었는지 명시적으로 전달할 수 있다. 이 매핑 정보는 강사가 의도적으로 강조한 핵심 어휘나 문장을 식별하는 결정적 단서로 활용되며, 해당 텍스트는 이후 요약 및 문제 생성 모듈에서 더 높은 가중치를 부여받게 된다.

이상의 과정을 통해 통합된 데이터(PDF 텍스트, 필기-텍스트 매핑 정보)는 앞서 '슬라이드 화면 추출' 단계에서 JSON 파일로 기록된 각 슬라이드의 타임스탬프 정보와 최종적으로 결합된다. 이 타임스탬프 정보는 STT 모듈에서 생성되는 음성 텍스트 세그먼트와의 동기화를 위한 핵심 정박지(anchor point) 역할을 수행하게 된다.

2) 음성 텍스트 변환

본 시스템에서 음성 처리 절차는 Fig. 1의 STT 블록에 해당하며, 다음 순서로 수행된다. 먼저 동영상에서 추출된 음성 데이터가 STT 모델로 입력되어 세그먼트 단위의 텍스트 변환과 함께 정확한 타임스탬프 정보가 생성된다. 이후 Vision Processing 모듈에서 확보한 슬라이드별 타임스탬프를 기준으로 STT 출력이 슬라이드 구간별로 재정렬되어 각 슬라이드에 해당하는 음성 내용이 정확히 분류된다.

본 연구에서는 동영상 내 음성 데이터를 텍스트로 변환하기 위한 STT 도구로 Whisper[5]를 채택하였다. Whisper는 OpenAI[8]에서 개발한 자동 음성 인식 시스템으로, 다음과 같은 근거를 바탕으로 본 시스템에 적합하다고 판단되었다. 첫째, Whisper는 다양한 언어에 대한 강력한 지원 기능을 제공한다. 특히 본 연구에서 다루는 한국어('ko') 음성 인식에 있어 우수한 성능을 보이며, 언어 식별 기능과 함께 다국어 지원을 통해 향후 시스템의 확장성을 확보할 수 있다. 둘째, Whisper는 다양한 모델 크기(예: 'large', 'medium', 'small' 등)를 제공하여 컴퓨팅 자원과 정확도 사이의 균형을 조절할 수 있다. 본 연구에서는 기본적으로 'large' 모델을 사용하여 높은 정확도를 확보하고자 하였으며, 우리의 제안 시스템에선 'stt_version' 파라미터를 추가하여 시스템 요구사항에 따라 정확도를 조정 가능하도록 함으로써 유연성을 확보하였다. 셋째, Whisper는 타임스탬프 기능을 제공하여 음성의 시작 및 종료 시간을 정확히 추적할 수 있다. 이는 본 연구에서 중요한 요소로, 슬라이드 전환 시점과 음성 내용을 정확히 연결하여 각 슬라이드에 해당하는 음성 내용을 매핑하는 데 필수적이다. 앞서 설명한 바와 같이, 본 연구에서는 각 음성 세그먼트의 시작과 끝 시간을 활용하여 슬라이드별 내용을 정확히 분리한다.

3) 데이터 통합 및 분석

앞서 3.2.1절과 3.2.2절에서 설명한 과정을 통해 추출된 결

과들은 슬라이드별 통합 과정을 통해 단일 데이터 스트림으로 변환된다. 즉, 각 슬라이드의 시작과 종료 시점 사이에 발생한 음성 내용이 해당 슬라이드의 OCR 결과와 결합되어 Vision + STT concatenated output 형태로 생성된다. 이러한 슬라이드 단위 통합을 통해 텍스트 및 시각적 정보(서면 텍스트 + 슬라이드 이미지 정보)와 음성 정보(음성 변환 텍스트)가 시간적 맥락을 유지하면서 하나의 데이터로 융합된다. 결론적으로 최종적으로 생성된 통합 데이터는 각 슬라이드별로 정렬된다.

3.3 요약 및 문제 생성

본 시스템의 핵심 기능인 요약 및 문제 생성 모듈은 Ollama의 EEVE-Korean-10.8B 모델을 활용하였다. 이 모델은 다음과 같은 이유로 본 시스템에 적합하다고 판단되었다. 첫째, EEVE-Korean-10.8B는 한국어에 특화된 대규모 언어 모델로, 한국어 학습 자료의 분석과 요약, 문제 생성에 우수한 성능을 발휘한다. 둘째, Ollama 프레임워크를 통해 로컬 환경에서도 대규모 언어 모델을 효율적으로 실행할 수 있다. 이는 사용자의 개인정보 보호 강화와 네트워크 의존성 최소화 측면에서 주목할 만한 이점을 제공한다. 셋째, EEVE-Korean- 10.8B 모델은 10.8B 파라미터 규모로, 복잡한 멀티모달 데이터의 맥락을 이해하고 의미 있는 요약과 교육적 가치가 높은 문제를 생성하기에 충분한 성능을 갖추고 있다. 특히 본 연구에서 중요시하는 주관식 및 객관식 문제 생성에 있어, 단순한 사실 기반 질문을 넘어 심층적 이해와 비판적 사고를 촉진하는 고품질 문제를 생성할 수 있다. 넷째, 프롬프트 엔지니어링을 통한 세밀한 출력 제어가 가능하다. 요약(summary), 주관식 문제(qna), 객관식 문제(mcq) 등 각 태스크별로 최적화된 프롬프트를 설계하고, 이를 EEVE-Korean-10.8B 모델에 입력함으로써 모델이 일관된 형식과 구조를 갖춘 고품질 출력물을 생성하도록 유도할 수 있다. 다섯째, 시스템 확장성 측면에서, Ollama 프레임워크는 추후 새로운 언어 모델로의 전환이나 모델 업데이트가 용이하다. 이러한 이유로 해당 모델을 활용하였으며, 본 시스템에서는 모델명을 파라미터화하여 처리함으로써 향후 더 발전된 모델로의 교체를 간소화하였다.

이러한 EEVE-Korean-10.8B 모델을 통해 통합된 멀티모달 데이터를 기반으로 효과적인 학습 자료를 생성하는 과정은 다음과 같이 진행된다. 요약 및 문제 생성을 위해 각 프레임의 텍스트 데이터를 기반으로 설계된 맞춤형 프롬프트를 바탕으로, 통합 및 전처리된 텍스트로부터 요약 문단과 주관식 및 객관식 문항을 생성한다. 요약 작업을 위한 프롬프트는 통합된 프레임별 텍스트에서 학습의 핵심 주제와 중요 개념을 추출하여 간결하고 체계적으로 재구성하도록 설계되었다. 주관식 문제 생성을 위한 프롬프트는 학습 내용을 바탕으로 심층적 사고를 유도하는 질문과 그에 대한 명확한 답변을 도출하도록 구성된다. 객관식 문제 생성 프롬프트는 선택형 질문과 함께 정답 및 적절한 난이도의 오답을 포함한 5개의 선택지를 생성하도록 구성되며, 이는 문제지와 답안지로 구분되어 출력된다. 태스크별 프롬프트의 세부 구조는 Table 1과 같으며, 주요 프레임별 텍스트 데이터와 각 태스크별 프롬프트가 대형 언어 모델의 입력으로 사용되어 최종적인 생성 결과를 얻을 수 있다. 최종적으로 생성된 요약과 문제들은 JSON 또는 텍스트 형식으로 저장된 후, 문제지와 답안지 구분, 시각적 강조 등을 포함한 PDF 문서로 변환된다. 이와 같이 본 시스템은 멀티모달 데이터의 효과적인 처리를 통해 학습자 맞춤형 요약 및 문제를 생성할 수 있도록 설계되었다.

4. 구현 및 성능 평가

본 섹션에서는 먼저 문자 인식 모듈, 음성 인식 모듈, 요약 및 문제 생성 모듈의 성능을 독립적으로 검증한다. 그 이후 제안한 MLSQ 시스템을 단일 모달 접근 방식(OCR Ollama, STT Ollama)과 비교하여 BLEU 및 ROUGE-L 점수를 활용하여 성능을 평가한다.

4.1 시스템 구현 환경 및 처리 과정

MLSQ 시스템은 Python을 주요 프로그래밍 언어로 하여 23코어 CPU, 190GB RAM, NVIDIA A100 GPU(1개)가 탑재된 리눅스 서버 환경에서 구현되었다. 슬라이드 화면 캡처는 OpenCV와 SSIM 기반의 프레임 비교 기법을 통해 수행되었고, 추출된 슬라이드 이미지 및 필기 내용은 PNG 형식으로 저장되었다. 각 프레임의 발생 시점은 JSON 형식의 메타데이터로 함께 기록되었다. 문자 인식은 EasyOCR을 사용하여 슬라이드 이미지로부터 텍스트를 추출하였으며, 결과는 슬라이드 단위의 텍스트 파일로 저장되었다. 음성 인식은 Whisper-large 모델을 기반으로 하여 강의 음성을 세그먼트 단위의 텍스트와 타임스탬프로 변환하였고, 이는 CSV 파일로 저장되었다. 이후 슬라이드 타임스탬프를 기준으로 OCR 결과와 STT 결과를 정렬・결합하여, 각 슬라이드에 대응되는 멀티모달 통합 텍스트 데이터를 구성하였다. 요약 및 문제 생성은 Ollama 프레임워크 기반의 EEVE-Korean-10.8B 모델을 활용하였다. 통합된 멀티모달 데이터를 태스크별 프롬프트와 함께 입력으로 구성하여 요약문, 주관식 및 객관식 문제를 생성하였으며, 결과물은 JSON 또는 텍스트 형식으로 저장되었다. 이후 FPDF 라이브러리를 이용해 명확하게 정리된 PDF 문서로 변환되었다. 본 시스템은 각 처리 단계의 중간 결과를 시각적 및 데이터 형태로 확인할 수 있도록 구성되어 있으며, 전체 파이프라인의 신뢰성과 정확도 향상에 기여하였다.

1) 성능 평가 데이터셋

제안하는 MLSQ 시스템은 멀티모달 입력(비디오, PDF, 음성)을 받아 여러 하위 태스크(OCR, STT, 요약, QG)를 순차적 및 통합적으로 처리하는 복합 시스템이다. 이러한 다중 구성

Table 1. Prompt Design and Structure for Each Task

Task	Prompt
Summary	〈전체 내용〉 {text} 〈/전체 내용〉 · 위 〈전체 내용〉은 강의 영상의 캡쳐본을 ocr하여 추출한 내용과 강의 영상의 녹화 목소리를 추출한 내용을 각 슬라이드 별로 합쳐놓은 내용이다. · 이 내용을 바탕으로 다음과 같은 형식으로 강의 요약을 작성해라.: 1. **강의 주제**: 강의의 주요 주제와 핵심 포인트를 간단히 설명한다. 2. **주요 개념**: 강의에서 다룬 주요 개념을 나열하고, 각 개념에 대해 짧게 설명한다. 3. **중요 내용**: 강의에서 강조된 주요 설명이나 예시를 포함한다.
QnA	〈전체 내용〉 {text} 〈/전체 내용〉 · 위 〈전체 내용〉은 강의 영상의 캡쳐본을 ocr하여 추출한 내용과 강의 영상의 녹화 목소리를 추출한 내용을 각 슬라이드 별로 합쳐놓은 내용이다. · 이 내용을 바탕으로 다음과 같은 형식으로 **총 10개**의 질문과 답변을 생성해라. · 모든 질문과 답변은 주관식 문제의 형식으로 작성해라. · 또한 반드시 문제지와 답변지를 따로 구별하여 작성하라. · 절대 문제와 답변이 혼용되지 말고 각각 문제지에 먼저 모든 문제들이 제공된 후, 답변지에 해당 문제들에 대한 답변이 제공되어야 한다. · 그리고 반드시 총 20개여야 한다(문제 10개, 답변 10개).: [문제지] 문제〈번호〉: 강의 내용을 바탕으로 적절한 질문을 생성한다. [답변지] 답변〈번호〉: 각 질문에 대한 간단하고 명확한 답변을 작성한다.
MCQ	〈전체 내용〉 {text} 〈/전체 내용〉 · 위 〈전체 내용〉은 강의 영상의 캡쳐본을 ocr하여 추출한 내용과 강의 영상의 녹화 목소리를 추출한 내용을 각 슬라이드 별로 합쳐놓은 내용이다. · 이 내용을 바탕으로 다음과 같은 형식으로 **총 10개**의 질문과 답변을 생성해라. · 단, 반드시 문제지(문제와 선택지포함)와 답변지를 따로 구별하여 작성하라. · 즉, 문제지에 총 10개의 문제와 그에 따른 선택지를 제공하고, 답변지에 총 10개의 답변을 제공하라. · 절대 문제와 답변이 혼용되지 말고 각각 문제지에 먼저 모든 문제들이 제공된 후, 답변지에 해당 문제들에 대한 답변이 제공되어야 한다. · 반드시 총 20개여야 한다(문제 10개, 답변 10개).: [문제지] 문제〈번호〉: 강의 내용을 바탕으로 객관식 문제를 작성한다. 선택지: 각 문제에 대해 5개의 선택지(**A**부터 **E**)를 제공한다. 정답은 이 중 하나로 설정한다. [답변지] 정답〈번호〉: 문제에 대한 올바른 정답(선택지)을 명시한다.

요소 시스템의 성능을 재현 가능하게 검증하기 위해, 본 평가는 각 하위 모듈(sub-module)과 핵심 태스크(core task)에 대해 학계에서 표준으로 인정받는 공개 벤치마크를 채택하였다. 각 모듈 및 태스크별 성능 검증에 사용된 벤치마크 구성은 다음과 같다.

- 개별 인식 모듈 평가(4.2.1, 4.2.2절): 시스템의 기반이 되는 인식 모듈의 타당성을 검증한다. 문자 인식(OCR) 모듈은 ICDAR 2015[26]를, 음성 텍스트 변환(STT) 모듈은

Table 2. Performance Comparison of OCR Models

Model	Recall (%)	Precision (%)	F1 (%)	CER (%)	Inference Time (s)
EasyOCR	88.5	90.2	89.3	8.5	0.6
TrOCR	91.3	93.0	92.1	7.1	3.2
TesseractOCR	87.6	85.5	86.7	12.4	1.0
PaddleOCR	83.0	85.6	86.2	11.4	0.7

AI Hub의 '한국어 강의 음성' 데이터셋[27]을 표준 벤치마크로 사용하여 성능을 측정하였다.

- 핵심 태스크 모듈 평가(4.2.3절): 본 시스템의 핵심 LLM인 EEVE-Korean-10.8B의 한국어 성능을 검증한다. 요약 모듈은 AI Hub의 '한국어 문서 요약' 데이터셋[28]을, 문제 생성(QG) 모듈은 KorQuAD 1.0 (The Korean Question Answering Dataset)[29] 데이터셋을 사용하여 표준 한국어 벤치마크 성능을 비교 평가한다.
- 통합 시스템 평가 (4.3.1절): 제안하는 MLSQ의 '멀티모달 융합' 방식 자체의 효과성을 검증한다. 이를 위해 슬라이드와 음성이 모두 포함된 강의 벤치마크인 LPM (Lecture Presentations Multimodal) 데이터셋[30]을 활용하여 '단일 모달 접근 방식'과 '융합 방식'의 성능을 직접 비교한다.

4.2 개별 모듈 평가

본 절에서는 MLSQ 시스템의 성능을 모듈별로 독립적으로 평가하며, 문자 인식(OCR), 음성 인식(STT), 요약 및 문제 생성 모듈의 성능을 각각 분석한다.

1) OCR 모듈 평가

본 실험에서는 OCR 성능을 평가하기 위해 대표적인 문자 인식 모델인 EasyOCR[11], TrOCR[12], TesseractOCR[13], PaddleOCR[14]을 비교 분석하였다. 실험에는 4.1.2절에서 명시한 표준 벤치마크인 ICDAR 2015[26] 데이터를 사용하였으며, 평가 기준으로 정확도(Precision), 재현율(Recall), F1-score, 문자 오류율(Character Error Rate, CER), 평균 추론 시간(Inference Time)을 활용하였다. Table 2는 OCR 모델들의 성능 비교 결과를 나타낸다. 실험 결과, TrOCR이 가장 높은 정확도(Precision)와 F1-score를 기록하였으나, 추론 시간이 상대적으로 길어 실시간 적용에는 한계가 있었다. 반면, EasyOCR은 빠른 속도와 균형 잡힌 성능을 보이며, MLSQ 시스템의 OCR 모듈로 최종 채택되었다.

2) STT 모듈 평가

본 실험에서는 음성 인식(STT) 모델의 성능 비교를 위해 Whisper[5], Google Speech-to-Text[15], DeepSpeech[16]를 4.1.2절에서 명시한 AI Hub의 '한국어 강의 음성' 데이터셋[27]에 적용하여 평가를 수행하였다. 실험에서는 단어 오류율(Word Error Rate, WER), 문자 오류율(Character Error Rate, CER), 평균 추론 시간을 주요 평가 지표로 사용하였다. Table 3은 STT 모델들의 성능 비교 결과를 나타낸다. 실험 결과, Whisper 모델이 WER과 CER에서 가장 낮은 오류율을 기록하여 비교 모델 대비 우수한 성능을 보였다. 이에 따라 MLSQ 시스템에서는 Whisper를 STT 모듈로 채택하였다.

Table 3. Performance Comparison of STT Models

STT Model	WER (%)	CER (%)	Inference Time (s)
Whisper	7.8	5.2	1.4
Google Speech-to-Text	8.5	6.0	1.1
DeepSpeech	12.3	9.8	1.7

Table 4. Performance Comparison of LLMs

LLM Model	BLEU (%)	ROUGE-L (%)
EEVE-Korean-10.8B	59.3	68.5
GPT-3.5	58.2	65.7
KoGPT	55.6	62.4

3) 요약 및 문제 생성 모듈 평가

요약 및 문제 생성(QG) 모듈의 핵심인 LLM 성능을 한국어 벤치마크 기준으로 평가하였다. 본 논문에서 채택한 EEVE-Korean-10.8B[7] 모델의 성능을 GPT-3.5[17] 및 KoGPT[18] 모델과 비교하였다. 요약 모듈 평가는 4.1.2절에서 명시한 AI Hub의 '한국어 문서 요약' 데이터셋[28]을 사용하였다. 데이터셋의 원문(기사)을 각 LLM에 입력하여 요약문을 생성하고, GT(사람 작성 요약문)와 비교하여 ROUGE-L 점수를 측정하였다. 문제 생성 모듈 평가는 KorQuAD 1.0[29] 데이터셋을 사용하였다. 데이터셋의 지문(Context)을 LLM에 입력하여 질문을 생성하게 한 후, GT(사람 작성 질문)와 비교하여 BLEU-4 점수를 측정하였다.

Table 4는 이 두 벤치마크에서 측정한 성능 점수들의 평균값을 비교한 결과이다. 실험 결과, 멀티모달 데이터를 활용하는 EEVE-Korean-10.8B 모델이 BLEU 및 ROUGE-L 점수에서 가장 높은 성능을 기록하였다. 이는 EEVE-Korean-10.8B가 텍스트와 함께 다양한 모달의 정보를 종합적으로 활용하여 더욱 정교한 요약 및 문제 생성을 수행할 수 있음을 보여준다.

4.3 통합 시스템 평가

본 절에서는 개별 모듈을 융합한 MLSQ 시스템의 성능 평가를 수행한다. 이를 통해 제안된 멀티모달 융합 시스템이 단일 모달 접근 방식(OCR Ollama, STT Ollama) 대비 성능 개선 효과를 종합적으로 분석한다.

1) 통합 시스템 성능 평가

본 실험에서는 LPM(Lecture Presentations Multimodal) 데이터셋[30]을 활용하여 제안된 MLSQ 시스템과 단일 모달 접근 방식(OCR Ollama, STT Ollama)들과 성능 비교를 수행하였다. 본 시스템의 기반 LLM인 EEVE-Korean-10.8B[7]는 다국어(multilingual) 기반 모델이므로 영어 벤치마크인 LPM 데이터셋의 처리에도 적합하다. 여기서 단일 모달 접근 방식인 'OCR Ollama'는 LPM 데이터셋의 슬라이드 텍스트(OCR GT)만을, 'STT Ollama'는 음성 스크립트(STT GT)만을 EEVE-Korean-10.8B 모델[7]의 입력으로 사용한 경우를 의미한다.

Table 5는 OCR Ollama 및 STT Ollama를 기반으로 한 단일 모달 접근 방식과 멀티모달을 활용한 MLSQ 시스템의 성능을 비교한 결과를 나타낸다. 실험 결과, MLSQ 시스템이 BLEU 점수에서 최대 3.4%p(STT Ollama 대비), ROUGE-L 점수에서 최대 3.1%p(OCR Ollama 대비) 향상된 성능을 기록하였다. 이러한 객관적 지표의 향상은 제안된 멀티모달 접근 방식이 기존 OCR 및 STT 기반 단일 모달 시스템보다 우수한 성능을 제공함을 입증한다. 성능 향상의 주된 요인은 두 가지로 분석된다. 첫째, 텍스트(OCR)와 음성(STT) 정보를 시간적으로 정렬하고 통합함으로써, 단일 모달에서 발생할 수 있는 정보 누락이나 문맥 손실을 상호 보완하였다. 둘째, 3.2.1절에서 기술한 바와 같이, 탐지된 필기 정보(시각 정보)와 매핑된 텍스트를 LLM의 프롬프트에 '강조' 항목으로 명시적으로 전달함으로써, 모델이 강사의 의도가 반영된 핵심 내용에 더욱 중점을 두고(focus) 요약 및 문제를 생성할 수 있었기 때문으로 판단된다. 이는 단순히 개별 데이터를 처리하는 수준을 넘어, 이질적인 정보원을 유기적으로 결합하여 학습 내용의 의미 구조를 보다 정교하게 반영할 수 있었음을 시사한다.

텍스트(OCR)와 음성(STT) 정보를 통합 처리하는 멀티모달 방식은 OCR 또는 STT 텍스트만을 사용하는 단일 모달 접근 방식에 비해 시간 및 공간 복잡도 측면에서 일정 수준의 오버헤드가 수반될 수 있다. 그러나 본 시스템이 지향하는 학습 자료 요약 및 문제 생성 기능은 실시간 응답을 최우선으로 요구하는 서비스가 아니며, 누락된 정보를 상호 보완하여 고품질의 학습 자료를 생성하는 것이 학습 효과를 극대화하는 데 더욱 중요하다. 따라서 본 시스템의 목적성에는 멀티모달 접근 방식의 채택이 더 적합하다고 판단된다. 더욱이, 본 시스템의 융합 방식은 각 모듈에서 전처리된 텍스트 데이터를 타임스탬프 기준으로 결합하는 것이므로, 연산량이 기하급수적으로 증가하지 않으면서도 성능 향상을 효과적으로 달성할 수 있다.

Table 5. Performance Comparison between MLSQ and Single-Modal Approaches for Summarization and Question Generation

System	BLEU (%)	ROUGE-L (%)
OCR Ollama	58.2	67.0
STT Ollama	57.8	68.4
MLSQ	61.2	70.1

2) 사용자 기반 평가

본 실험에서는 제안된 MLSQ 시스템의 성능을 보다 종합적으로 검증하기 위해 주관적 사용자 평가(Subjective Evaluation)를 수행하였다. 사용자 평가 실험은 실제 학습 자료를 일상적으로 접하고 활용하는 서비스 대상군인 대학생 및 대학원생 24명을 대상으로 진행되었다. 남성 14명(58.3%)과 여성 10명(41.7%)이었으며, 연령대는 20대(20-29세)로 구성되었다. 각 참가자는 OCR, STT, 요약 및 문제 생성 결과물의 품질을 MOS(Mean Opinion Score) 방식으로 평가하였다. MOS는 1점(매우 불만족)에서 5점(매우 만족)까지의 척도로 이루어진 평가 방식이다. 사용자 평가는 다음과 같은 기준으로 수행되었다.

- 문자 인식 정확성: 문자 인식 결과가 원본 텍스트와 유사한지 평가
- 음성 변환 신뢰도: 음성 변환된 텍스트가 원본 음성과 일치하는 정도 평가
- 요약의 정보 보존율: 요약된 내용이 원본의 핵심 정보를 포함하는지 평가
- 문제 생성의 학습 적합성: 생성된 문제가 학습자의 이해도 향상에 적절한지 평가

사용자는 각 항목에 대해 점수를 부여하였으며, 평가 결과는 Fig. 5와 Fig. 6에 제시되었다. Fig. 5는 제안된 MLSQ 시스

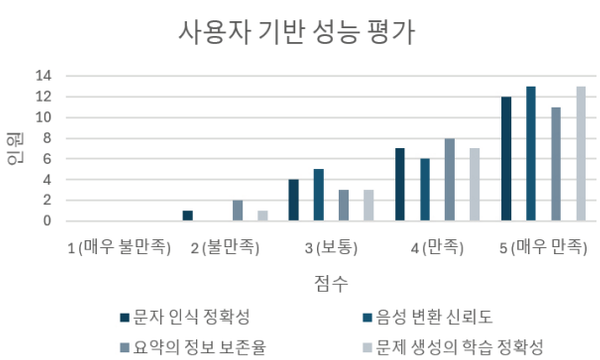


Fig. 5. Result of User-Based Performance Evaluation

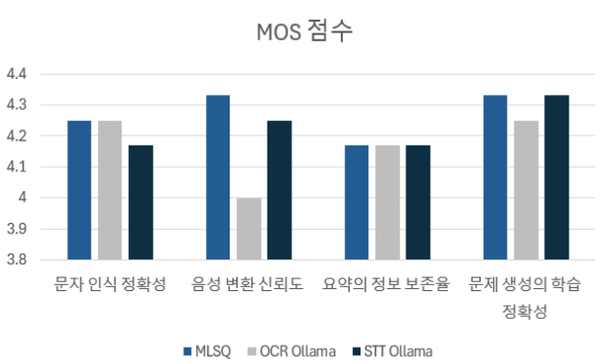


Fig. 6. Comparison of MOS Scores between MLSQ and Single-Modal Approaches

템의 각 항목별 점수 분포를 나타내며, 대부분의 응답자가 4점 이상(만족~매우 만족)으로 평가하였다. 특히 음성 변환 신뢰도와 문제 생성의 학습 적합성 항목에서 각각 평균 4.33점으로 가장 높은 점수를 기록하였다.

Fig. 6은 제안된 MLSQ 시스템과 단일 모달 접근 방식(OCR Ollama, STT Ollama)의 평균 MOS 점수를 비교하여 나타낸다. 비교 결과, 제안된 MLSQ 시스템은 문자 인식 정확성(4.25) 및 음성 변환 신뢰도(4.33)에서 단일 모달 접근 방식과 동일하거나 우수한 성능을 보였으며, 요약의 정보 보존율(4.17)과 문제 생성의 학습 적합성(4.33)에서도 단일 모달 대비 동등하거나 더 높은 평가를 받았다. 이러한 결과는 제안된 MLSQ 시스템이 단일 모달 접근 방식 대비 멀티모달 정보를 효과적으로 통합하여 보다 명확하고 교육적으로 실질적인 콘텐츠를 제공할 수 있음을 시사한다. 또한 모든 평가 항목에서 평균 4.0 이상의 높은 점수를 기록하여 시스템 전반의 사용자 만족도와 교육적 활용 가능성이 뛰어남을 확인할 수 있었다.

5. 결론 및 향후 연구

본 연구에서는 멀티모달 데이터를 활용한 AI 기반 학습 자료 요약 및 문제 생성 시스템을 제안하였다. 기존의 단일 모달 접근 방식이 가지는 한계를 극복하기 위해 OCR, STT, 요약 및 문제 생성 기술을 통합하여 학습 자료를 종합적으로 분석하고 활용할 수 있는 시스템을 구축하였다. 이를 통해 동영상과 문서 자료를 효과적으로 처리하고 학습 내용을 요약하며, 자동으로 문제를 생성하는 시스템을 구현하였다. 성능 검증 결과, 제안 시스템은 단일 모달 방식 대비 요약 및 문제 생성 성능에서 BLEU 점수가 최대 3.4%p, ROUGE-L 점수가 최대 3.1%p 향상되는 등 객관적 지표에서 우수성을 보였다. 또한, 사용자 평가를 통해 제안 시스템의 활용 가능성을 검토한 결과, 제안된 MLSQ 시스템이 단일 모달 접근 방식보다 우수한 성능을 보이며 높은 사용자 만족도를 기록하였다. 특히, 음성 변환 신뢰도(4.33)와 문제 생성 학습 적합성(4.33)에서 기존 기법 보다 높은 점수를 얻으며, 학습 자료 분석 및 문제 생성에서의 실용성을 입증하였다.

향후 연구에서는 대형 언어 모델 기반 요약 및 문제 생성의 정밀도 향상 및 자동 난이도 조절 및 개별 학습자 맞춤형 문제 제공을 통해 본 시스템의 실용성과 성능을 더욱 개선하고자 한다. 이를 통해 보다 정교한 AI 기반 학습 지원 시스템을 구축하고, 다양한 교육 환경에서의 적용 가능성을 높일 것이다.

References

[1] A. Gulati et al., “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in neural information processing systems*, 2020.

[3] Z. Yao et al., “Zipformer: A faster and better encoder for automatic speech recognition,” *arXiv preprint arXiv:2310.11230*, 2023.

[4] Y. Meng et al., “Dolphin: A Large-Scale Automatic Speech Recognition Model for Eastern Languages,” *arXiv preprint arXiv:2503.20212,* 2025.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*, 2023.

[6] Ollama [Internet], https://github.com/ollama/ollama

[7] S. Kim, S. Choi, and M. Jeong, “Efficient and effective vocabulary expansion towards multilingual large language models,” *arXiv preprint arXiv:2402.14714*, 2024.

[8] Whisper [Internet], https://openai.com/research/whisper

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[10] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *arXiv preprint arXiv:1506.01497*, 2015.

[11] EasyOCR [Internet], https://github.com/JaidedAI/EasyOCR

[12] M. Li et al., “TrOCR: Transformer-based optical character recognition with pre-trained models,” *arXiv preprint arXiv:2109.10282*, 2021.

[13] R. Smith, “Tesseract: An open-source optical character recognition engine,” in *Linux Journal,* 2007.

[14] Y. Du et al. “PP-OCR: A practical ultra-lightweight OCR system,” *arXiv preprint arXiv:2009.09941*, 2020.

[15] Google Cloud Speech-to-Text API [Internet], https://cloud.google.com/speech-to-text

[16] A. Hannun et al., “Deep Speech: Scaling up end-to-end speech recognition,” *arXiv preprint arXiv:1412.5567*, 2014.

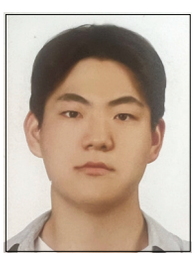
[17] GPT-3.5 Model Documentation [Internet], https://platform.openai.com/docs/models/gpt-3-5

[18] KoGPT [Internet], https://github.com/kakaobrain/kogpt

[19] J. Kim and M. Lee, “딥러닝을 활용한 자동 문제 생성 시스템 설계 및 구현,” in『한국정보과학회논문지: 소프트웨어 및 응용』, 2022.

[20] S. Park, H. Lee, and J. Jeong, “멀티모달 데이터를 활용한 개인

맞춤형 교육 콘텐츠 생성 연구," in『한국교육공학회지』, 2021.
[21] M. Kurdi, M. Zahera, and H. Al-Samarraie, "A systematic review of automatic question generation for educational purposes," in *Computers & Education*, 2021.
[22] X. Zhan, Y. Li, and Y. Xu, "Synthetic multimodal question generation," *arXiv preprint arXiv:2407.02233*, 2024.
[23] H. Liu, Y. Jiang, and M. Sun, "UniRaG: Unification, retrieval, and generation for multimodal question answering," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
[24] Z. Wang, Y. Chen, and C. Lin, "Chain-of-Exemplar: Enhancing distractor generation for multimodal educational question generation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.
[25] Ou, T. S., Huang, Y. H., & Chen, H. H. (2011). "SSIM-based perceptual rate control for video coding," *IEEE transactions on circuits and systems for video technology*, Vol.21, No.5, pp.682-691.
[26] D. Karatzas et al., "ICDAR 2015 robust reading competition," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015.
[27] AI Hub, "한국어 강의 음성," [Internet], https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=115
[28] AI Hub, "한국어 문서 요약 텍스트," [Internet], https://aihub.or.kr/aidata/8054
[29] "KorQuAD 1.0," [Internet], https://korquad.github.io/
[30] D. W. Lee et al., "Lecture Presentations Multimodal Dataset: Towards Understanding Multimodality in Educational Videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.



유 건 우
https://orcid.org/0009-0009-0305-5086
e-mail : yctw57597@hansung.ac.kr
2022년 3월~현 재 한성대학교 AI응용학과 학사과정
관심분야 : AI, IoT, Computer Vision, NLP

이 상 윤
https://orcid.org/0009-0002-5344-7386
e-mail : lsy010404@hansung.ac.kr
2022년 3월~현 재 한성대학교 AI응용학과 학사과정
관심분야 : AI, Internet of Things, Computer Vision

안 진 영
https://orcid.org/0009-0005-1239-1402
e-mail : jinyoung0309@hansung.ac.kr
2022년 3월~현 재 한성대학교 AI응용학과 학사과정
관심분야 : AI, Deep/Machine learning

우 민 하
https://orcid.org/0009-0003-6213-0265
e-mail : dn0108@hansung.ac.kr
2022년 3월~현 재 한성대학교 AI응용학과 학사과정
관심분야 : AI

김 수 경
https://orcid.org/0009-0001-8867-715X
e-mail : 2171175@hansung.ac.kr
2021년 3월~현 재 한성대학교 IT융합공학부 학사과정
관심분야 : AI, IoT, Deep/Machine learning

이 준 구
https://orcid.org/0009-0006-0408-4565
e-mail : dlwnsrn8211@hansung.ac.kr
2021년 3월~현 재 한성대학교 AI응용학과 학사과정
관심분야 : AI, IoT, Computer Vision

최 현 우
https://orcid.org/0009-0004-6826-4149
e-mail : apinball91@hansung.ac.kr
2021년 3월~현 재 한성대학교 AI응용학과 학사과정
관심분야 : AI, Deep/Machine learning, Generative AI, Computer Vision

김 예 란

https://orcid.org/0000-0002-5788-6178
e-mail : yaerankim@hansung.ac.kr
2024년 2월 한성대학교 빅데이터트랙 학사
2024년 3월~현 재 한성대학교 AI응용학과 석사과정
관심분야 : AI, Generative model, LLM

이 웅 희

https://orcid.org/0000-0003-0856-6415
e-mail : whlee@hansung.ac.kr
2013년 2월 고려대학교 전기전자전파공학부 학사
2019년 2월 고려대학교 전기전자공학과 박사
2019년 5월 LG유플러스 선임 연구원
2019년 8월~2021년 2월 삼성전자 Staff Engineer
2021년 3월~현 재 한성대학교 AI응용학과 조교수
관심분야 : Internet of Things, AI, Deep/Machine learning, TCP/IP Protocol, Network analysis/modeling