# 사용자 필기 정보 기반 문서 요약 웹서비스 개발*

김민아 ○, 김민석, 박상근
경희대학교 소프트웨어융합학과
eulneul@khu.ac.kr, kmss0206@khu.ac.kr, sk.park@khu.ac.kr

# Development of Document Summarization Web Service Based on User Handwriting Information

Mina Kim ○, Minseok Kim, Sangkeun Park
Department of Software Convergence, Kyung Hee University

**요 약**

문서에 작성한 필기 정보는 회의나 수업 내용을 효과적으로 회상하는 데 매우 효과적이다. 전자 문서를 활용한 학습이 대중화되면서, 해당 전자문서의 내용을 요약하기 위한 다양한 연구 및 서비스가 등장했다. 기존 문서 요약 연구 및 서비스는 원문 텍스트에서 중요한 텍스트를 자동으로 추출하고 요약하는데, 이때 학습을 위해 사용자가 중요하게 표시한 필기 정보가 반영되지 않는다는 한계가 존재한다. 본 연구는 문서에서 사용자의 필기 객체를 인식하고, 필기가 있는 구문에 가중치를 부여하며 문서를 요약하는 서비스를 개발하고 사용자 스터디를 통해 활용성을 검증하였다.

## 1. 서 론

코로나19 영향으로 교육의 디지털화가 가속화되었다. 이에 따라 종이로 된 학습 자료 보다 PDF 등의 디지털 학습 자료 활용이 대중화되었으며, 디지털 교육자료 활용을 위한 태블릿 PC 판매량도 급격히 늘어났다. 특히, 주된 학생 연령층인 13~29세의 2022년 태블릿 PC 보유율은 전년 대비 약 50% 정도 상승했다[1].

방대한 양의 디지털 자료를 효과적으로 활용해서 학습할 수 있도록, 효과적인 문서 요약을 위한 다양한 연구도 수행되었다 [2, 3, 4]. 하지만 기존의 문서 요약은 전체적인 맥락 파악에만 도움되지만, 사용자가 학습 과정에서 중요하다고 표시하거나 기록한 필기 내용을 전혀 반영하지 않는다는 한계가 존재한다. 필기는 사용자가 학습 자료를 회상하는데 도움을 주는 핵심 요소이다. DeZure et al.[5]의 연구에 따르면, 시험 전에 노트필기와 함께 공부하는 행위는 과거의 기억을 회상하는데 도움을 주고 새로운 지식을 배우는데 매우 유용하다.

본 연구에서는 디지털 문서에서 사용자가 직접 필기한 정보를 추출하고, 이를 문서 요약에 반영하는 사용자 필기 정보 기반 문서를 요약하는 새로운 모델을 제안한다. 그리고 해당 모델을 학습에 효과적으로 사용할 수 있도록 디지털 문서를 업로드하면 사용자의 필기 부분을 인식하고 이를 반영한 요약 텍스트를 제공하는 웹서비스를 개발했다. 해당 연구의 활용성을 검증하기 위해 사용자 스터디를 수행하고, 사용자의 필기 정보가 반영된 문서 요약이 학습에 유용하게 활용될 수 있음을 확인하였다.

## 2. 관련 연구

문서 요약 성능을 개선하기 위한 연구가 활발히 이루어지고 있다. 특히, 문서 요약 성능과 깊은 관련이 있는 요약문과 원문 사이의 불일치 문제 해결, 그리고 원문에서의 중요 문장 인식 및 요약 반영은 문서 요약의 성능을 판단하는 데 있어 매우 중요한 요소이다. 이러한 문제 해결을 위해, 구다훈 et al.[2] 은 원문과 생성 요약문에 불일치가 발생하는 점을 정답 요약문과 생성 요약문의 중복성을 원인으로 보고, 모델 학습 과정에서 해당 중복성을 줄이려는 함수를 제안하였다. 전민규 & 김남규[3]는 문서 내의 중요한 문장을 의미적 유사도를 기준으로 구분하고 문서 요약에 해당

* "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2024년도 SW중심대학사업의 결과로 수행되었음"(2023-0-00042)

중요도를 반영하고자 하였다. 구동준 et al.[4] 은 특허 문서 요약 보조 시스템으로 핵심 정보를 도출하는 서비스를 개발하였다. 구체적으로 특허청에서 제공하는 요약서 가이드라인을 기반으로 특허 문서의 특성을 추출하고 데이터를 가공하였다.

문서 요약 기술을 활용해 사용자의 학습을 돕기 위한 연구도 진행되고 있다. 예를 들어, Wang et al. [6]은 학생들의 노트 요약을 위한 자료 통합 웹서비스를 개발하였다. 해당 웹서비스는 학습자의 인터넷 자료, 강의 노트, 슬라이드 내용을 통합하여 요약문의 필수 문장을 선택하고 결과 페이지에서 해당 문장을 하이라이트로 표시했다. 최용균 et al. [7]은 인쇄자료의 하이라이트를 추출하여 그 문장을 파일로 저장해주는 서비스를 개발하였다.

본 연구에서는 기존의 문서 요약 메커니즘에서 더 나아가, 사용자의 '필기'에 가중치를 부여하는 새로운 문서 요약 알고리즘을 제안한다. 기존 연구에서 문장의 중요도 선정을 다양한 손실함수와 유사도를 통해 진행하여 사용자가 개입할 수 없었던 반면, 사용자마다 중요한 문장이 '필기'에 드러나 있음을 가설로 설정하고 사용자의 학습에 도움이 되도록 해당 필기에 맞추어 문서 요약을 진행하고자 한다. 더불어, 해당 문서 요약 기술이 사용자의 학습에 도움이 되도록 사용자의 필기를 반영한 문서 요약 웹서비스를 개발해 해당 연구의 사용성을 검증한다.

## 3. 필기 및 문서 요약 서비스 WriteNow

본 연구에서는 사용자의 필기가 포함된 PDF 파일에서, 필기 내용이 지워지지 않도록 문서를 요약하는 웹서비스 WriteNow를 개발했다. 해당 서비스는 1) 필기 정보 추출 모듈, 2) 필기에 가중치를 부여한 문서 요약 모듈로 구성되어 있다.

### 3.1 필기 정보 추출

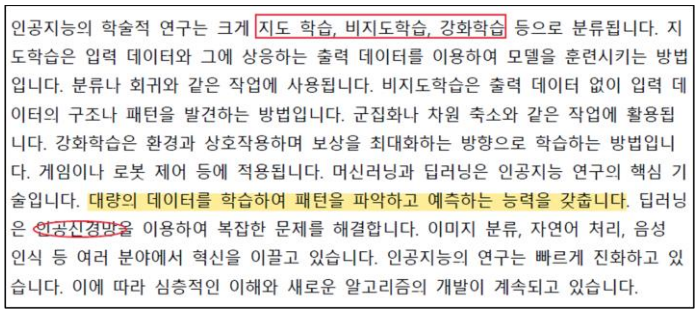
인공지능의 학습적 연구는 크게 지도 학습, 비지도학습, 강화학습 등으로 분류됩니다. 지도학습은 입력 데이터와 그에 상응하는 출력 데이터를 이용하여 모델을 훈련시키는 방법입니다. 분류나 회귀와 같은 작업에 사용됩니다. 비지도학습은 출력 데이터 없이 입력 데이터의 구조나 패턴을 발견하는 방법입니다. 군집화나 차원 축소와 같은 작업에 활용됩니다. 강화학습은 환경과 상호작용하며 보상을 최대화하는 방향으로 학습하는 방법입니다. 게임이나 로봇 제어 등에 적용됩니다. 머신러닝과 딥러닝은 인공지능 연구의 핵심 기술입니다. 대량의 데이터를 학습하여 패턴을 파악하고 예측하는 능력을 갖춥니다. 딥러닝은 인공신경망을 이용하여 복잡한 문제를 해결합니다. 이미지 분류, 자연어 처리, 음성 인식 등 여러 분야에서 혁신을 이끌고 있습니다. 인공지능의 연구는 빠르게 진화하고 있습니다. 이에 따라 심층적인 이해와 새로운 알고리즘의 개발이 계속되고 있습니다.

그림 1. 사용자 필기가 포함된 원본 텍스트

필기 정보 추출 모듈에서는 PDF 파일에서 사용자의 필기 객체를 추출하고, 각 객체가 원문에서 어떤 텍스트와 관련이 있는지 식별한다. PDF 파일 포맷에서는 사용자가 필기 어플리케이션 등을 활용해 필기한 내용(예: 하이라이트, 밑줄, 네모, 동그라미 등)이 벡터 그래픽 객체로 저장된다 [그림 1]. PyMuPDF [2] 라이브러리를 활용해 해당 벡터 그래픽 객체를 추출하고, 해당 객체의 위치 좌표를 기반으로 필기가 적용된 텍스트도 함께 추출한다. 이렇게 추출된 텍스트는 필기가 반영된 중요한 텍스트이므로 문서 요약 모델에서 가중치를 주는 데 활용한다.

### 3.2 문서 요약 모델

문서 요약을 위해 BART 모델을 한국형으로 바꾼 사전 학습된 트랜스포머 기반 언어 모델인 KoBART[3]를 사용했다. 필기 정보 추출 모듈에서 추출된, 필기가 반영된 텍스트를 중요 텍스트로 판단하고 해당 텍스트에 가중치를 부여한 후에 KoBART 모델로 문서 요약을 수행한다.

중요 텍스트에 가중치를 부여하는 방법은 다음과 같다. 원문에서 사용자가 필기한 중요 텍스트가 포함된 문장들을 찾아낸다. 중요 텍스트가 포함된 문장마다, 해당 문장과 동일한 문장을 하나 더 생성하여 해당 문장 뒤에 이어 붙인다. 이제 중요 텍스트가 포함된 문장들은 원문에 총 2개씩 존재하게 되고, 이 텍스트를 KoBART 모델로 요약하면 필기 정보가 포함된 중요 텍스트는 사라지지 않고 요약문에 그대로 남게 된다.

사용자의 실제 필기가 적용된 텍스트를 찾아 강조 표시를 추가하기 위해, 요약된 결과에서 다시 사용자의 필기가 포함된 중요 텍스트를 찾는다. 요약문을 생성할 때 문장이 패러프레이징될 수도 있기 때문에, 요약된 결과의 모든 단어를 순회하면서 중요 텍스트와 일치하는 구문을 찾아 70%이상 일치하면 노란색으로 하이라이팅 효과를 삽입한다.

### 3.3 WriteNow 웹서비스

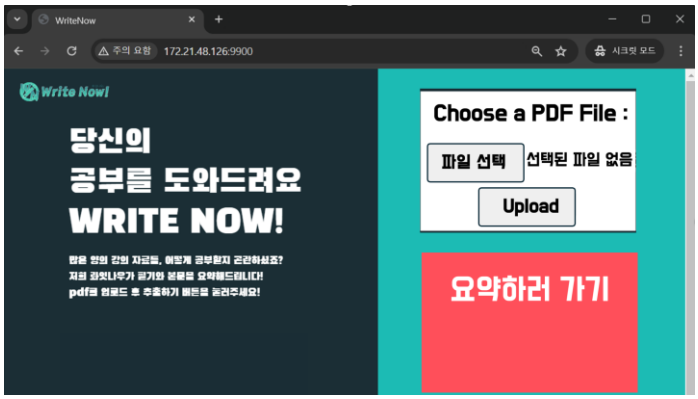


그림 2. WriteNow 화면

[2] https://pymupdf.readthedocs.io/en/latest/
[3] https://github.com/SKT-AI/KoBART

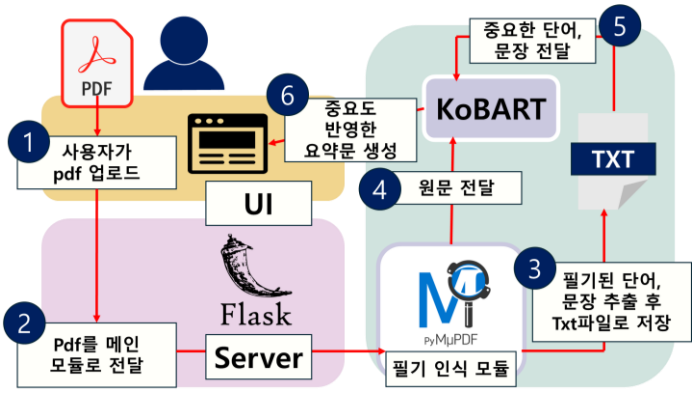


그림 3. WriteNow 아키텍쳐

인공지능의 학술적 연구는 크게 지도 학습, 비지도 학습, 강화 학습 등으로 분류가 되며, 머신 러닝과 딥러닝은 인공지능 연구의 핵심 기 술로 대량의 데이터를 학습하여 패턴을 파악하고 예측하는 능력을 갖추고 있고, 딥러닝은인공신경망을이용하여 복잡한 문제를 해결합하여 인공지능 연구의 핵심 기 술로 대량의 데이터를 학습하여 패턴을 파악하고 예측하는 능력을 갖췄다.

그림 4. 사용자 필기를 반영해 요약된 텍스트

본 논문에서 필기 정보 추출 모듈과 요약 모델을 통합하여 실제로 필기 정보를 담고 있는 문서를 요약해 주는 웹서비스를 개발했다 [그림2]. 해당 서비스의 작동 구조는 [그림 3]과 같다. 사용자가 요약하고 싶은 문서를 업로드하면, 해당 문서가 Flask로 구현한 서버로 전송된다. 서버에서는 해당 문서에서 필기 정보를 추출하고 문서 요약을 수행한다. 예를 들어, [그림 1]과 같은 PDF 파일을 업로드하면 [그림 4]와 같이 요약된 문서가 생성된다. 사용자가 필기한 내용이 요약된 텍스트에 그대로 남아있고, 해당 부분이 노란색으로 강조되어 있음을 확인할 수 있다.

## 4. 사용자 스터디

평소에 태블릿 PC의 필기 어플리케이션을 활용해 학습하는 3명의 20대 참여자(남자 1, 여자 2)를 모집했다. 필기 어플리케이션의 사용 목적에 대한 질문에, “스스로 정리한 요약문에 필기 내용이 빠지진 않았는지 확인하기 위해”, “필기 내용을 바탕으로 수업 중 중요한 내용을 다시 떠올리기 위해”라고 응답했다. 이에 따라, 디지털을 활용한 학습에서도 학습자의 필기는 여전히 중요함을 확인할 수 있었다.

각 참여자는 WriteNow의 사용법에 대한 설명을 듣고, 본인이 실제로 필기하며 학습에 활용한 전자문서(PDF)를 WriteNow를 통해 요약된 결과를 확인했다. WriteNow가 필기 정보를 반영해서 생성한 요약문의 학습 효과에 대한 질문에, 참여자들 모두 필기를 반영하여 요약을 해주는 아이디어에 긍정적인 의견을 내놓았다. 예를 들어, “공부할 때 필기 위주로 공부하므로, 요약문에 필기가 반영된다면 공부할 때 도움이 될 것 같다", “필기가 요약문에도 표시되니 이것을 바탕으로 공부를 하면 좋을것 같다”, “논문처럼 긴 글을 필기와 함께 요약할 때는 유용하다”라는 응답이 있었다. 특히, 한 참여자는, “기존의 필기 없이 그냥 요약되는 서비스는 자기가 중요하다고 생각되는 부분이 요약이 안 될 수 있을 것 같아서 공부하는데 사용하지 않을 것 같다”고 기존의 텍스트 요약 모델 및 서비스에 대한 한계를 언급했다.

## 5. 결론

본 연구에서는 사용자의 필기를 반영해서 문서를 요약해 주는 모듈을 개발하고, 이를 기반으로 실제 문서 요약을 수행할 수 있는 웹서비스를 개발했다. 사용자 스터디를 통해, 필기 정보와 자료를 함께 요약하면 이전 내용을 효과적으로 활용할 수 있음을 확인했다. 또한 기존 연구와 달리 사용자 중심적 접근 방법을 사용하여, 사용자가 결과물을 직접 제어할 수 있다는 점에서 큰 차이가 있다. 사용자는 자신의 필요와 선호에 따라 요약을 조정하고, 개인적인 판단에 따라 중요하다고 생각하는 부분을 요약문에 포함할 수 있다. 향후에 문서 요약의 정확도를 높이고, OCR을 활용한 손글씨, 사진, 표 등의 고급 객체를 인식 후 문서 요약에 반영한다면 사용자의 학습에 더욱 도움이 될 것으로 기대된다.

## 6. 참고문헌

[1]조아라, “노트북·스마트폰에 치이더니 '수요 폭발'…대학생 '필수품' 됐다”, 한국경제, 2022.12.14
[2]구다훈, et al. “생성요약의 사실 불일치 문제 개선을 위한 관련성과 중복성을 고려한 손실 함수 기반의 KoBART 모델”, 한국정보기술학회논문지, 20(12), 25-36, 2022
[3]전민규, 김남규. “문서 요약 품질 향상을 위한 의미 기반 추가 사전학습 방법론”, 한국지능정보시스템학회 학술대회논문집, 2022, 154 - 155, 2022
[4]구동준, et al. “BART 모델 기반의 긴 특허 문서 요약 시스템 개발”, 한국정보과학회 학술발표논문집, 2022, 410 - 412, 2022
[5]DeZure, et al. “Research On Student NoteTaking: Implications For Faculty and Graduate Student Instructors”, CRLT Occasional Papers, 16, 1-8, 2001
[6]Hei-Chia Wang, et al. “NoteSum: An integrated note summarization system by using text mining algorithms”, Information Sciences, 513, 536-552, 2020
[7]최용훈, et al. “Google Vision API를 활용한 문자 추출과 요약노트 시스템”, 한국HCI학회, 2019, 685-687, 2019