Journal of KIIT. Vol. 24, No. 7, pp. 265-273, Jul. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.7.265

문맥 인식 기반 PDF 키워드 마스킹을 위한 학습 시스템 설계 및 구현

김민준*[1], 이가은*[2], 김지원**, 강창구*[3]

Design and Implementation of a Context-Aware Keyword Masking Learning System for PDF Documents

Minjun Kim*[1], Ga Eun Lee*[2], Jiwon Kim**, and Changgu Kang*[3]

요 약

본 논문에서는 디지털 PDF 환경에서의 학습 효율 향상을 위해 문맥 인식 기반 키워드 마스킹 학습 시스템 PDFMask를 제안한다. 제안 시스템은 PDF 내부의 텍스트 및 위치 정보를 활용하여 문서의 논리적 읽기 순서를 복원하고, 형태소 및 형태-구문 분석을 통해 핵심 키워드를 자동으로 식별한다. 이후 비율 기반 샘플링과 시각적 어노테이션을 적용하여 원본 문서의 레이아웃을 유지하면서도 학습에 적합한 마스킹 PDF를 생성한다. 또한 웹 서버와 처리 엔진을 분리한 비동기 메시지 큐 기반 아키텍처를 적용하여 대용량 PDF 처리 환경에서도 효율적인 시스템 운영이 가능하도록 설계하였다. 제안 시스템은 기존 플래시카드 기반 학습에서 발생하는 문맥 정보 손실 문제와 PDF 어노테이션 기반 학습의 반복적 수작업 부담을 완화한다.

Abstract

This paper proposes PDFMask, a context-aware keyword masking learning system designed to improve learning efficiency in digital PDF environments. The proposed system utilizes textual and positional information embedded in PDF documents to reconstruct the logical reading order and automatically identify key terms using morphological and syntactic analysis. A masking PDF suitable for learning is then generated by applying ratio-based sampling and visual annotation while preserving the original document layout. In addition, an asynchronous message queue-based architecture that separates the web server from the processing engine is adopted to enable efficient handling of large-scale PDF processing requests. The proposed system alleviates the loss of contextual information observed in flashcard-based learning and reduces the manual effort required in PDF annotation-based approaches.

Keywords

PDF document processing, keyword masking, context-aware learning, morphological analysis

* 경상국립대학교 컴퓨터공학부(*[3] 교신저자)
- ORCID[1]: https://orcid.org/0009-0000-4060-582X
- ORCID[2]: https://orcid.org/0009-0001-1259-8527
- ORCID[3]: https://orcid.org/0000-0003-4060-6835
** ㈜페쎄이상
- ORCID: https://orcid.org/0009-0009-5266-5008

· Received: Apr. 08, 2026, Revised: Jun. 02, 2026, Accepted: Jun. 05, 2026
· Corresponding Author: Changgu Kang
Dept. of Computer Science and Engineering, Gyeongsang National University, Korea
Tel.: +82-55-772-3321, Email: cgk@gnu.ac.kr

Ⅰ. 서 론

IT 기술 발전과 함께 PDF 형식의 디지털 문서는 학술 논문, 전공 교재, 기술 보고서 등 지식 습득을 위한 중요한 매체로 이용되고 있다. 특히 대학 교육 환경에서는 출판사가 제공하는 디지털 네이티브 PDF 교재의 비중이 급격히 증가하고 있으며, 학습자들은 태블릿이나 웹 브라우저를 통해 언제 어디서든 원본 문서에 접근하여 학습하는 것이 보편화되었다. 이러한 변화는 클라우드 기반 문서 저장·검색·분석 시스템의 발전과 함께, 문서 처리 기술이 단순한 오프라인 도구를 넘어 확장성과 접근성을 갖춘 정보 인프라로 자리 잡고 있다[1][2].

이러한 문서 환경의 변화에 발맞추어, 광학 문자 인식(OCR, Optical Character Recognition)에서 출발한 문서 처리 기술은 딥러닝 기반의 레이아웃 분석 모델로 급속히 고도화되어 왔으며, 디지털 아카이빙이나 비즈니스 프로세스 자동화 등 다양한 정보시스템 분야에서 활용되고 있다[3]-[6]. 그러나 이들 기술은 주로 스캔 문서의 구조 해석이나 대규모 데이터 추출에 최적화되어 있어, 디지털 네이티브 PDF 교재를 활용한 학습자 중심의 교육적 활용요구와는 차이를 보인다.

교육 분야에서는 학습자가 핵심 내용을 가린 상태에서 스스로 답을 떠올려 보는 능동적 인출기반 학습 전략이 인지과학적으로 높은 효과를 인정받고 있다[7]. 이 이론에 착안하여 디지털 플래시카드, 빈칸 문항 자동 생성 시스템, PDF 어노테이션 앱 등 다양한 학습 보조 도구들이 개발·활용되어 왔다. 그러나 이러한 도구들은 원본 문서의 시각적 문맥 상실, 반복적 수작업의 비효율, 그리고 불필요한 고비용 연산 구조 등의 문제로 인해 전자문서 기반 PDF 환경에서 키워드 단위 학습을 체계적으로 지원하는 데에는 제한적이다.

본 논문에서는 이러한 한계를 해결하기 위해 PDF 내부의 텍스트 및 위치 정보를 활용하여 원본 문서의 논리적 구조와 레이아웃을 유지한 채 핵심 키워드 마스킹을 수행하는 문맥 인식 기반 학습 시스템 PDFMask를 제안한다. 제안 시스템은 형태소 및 형태-구문 분석을 통해 핵심 키워드를 자동으로 식별하고, 해당 위치에 마스킹을 적용하여 학습용 PDF를 생성한다. 이를 통해 기존 플래시카드 방식의 문맥 정보 손실 문제와 PDF 어노테이션 기반 학습의 반복적 수작업 부담을 완화할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 지능형 문서 처리 기술과 디지털 학습 보조 방식의 동향 및 한계점을 고찰한다. 3장에서는 PDFMask의 전체 시스템 구조 및 설계를 기술한다. 4장에서는 실제 배포된 시스템의 시연 및 마스킹 처리 결과를 보이고, 5장에서는 결론 및 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

본 절에서는 지능형 문서 처리 기술의 발전 동향과 기존 디지털 학습 보조 시스템의 현황을 각각 살펴보고, 각 접근 방식의 한계를 종합적으로 분석하여 본 연구의 필요성을 제시한다.

2.1 지능형 문서 처리 기술의 고도화

초기 문서 분석 연구는 주로 스캔된 이미지 문서를 텍스트로 변환하고 색인 및 검색이 가능하도록 하는 데 초점을 맞추었다[1]. 특히 Tesseract[2]와 같은 광학 문자 인식(OCR) 엔진의 등장은 텍스트 기반 문서 처리 기술의 발전을 촉진하였다. 이후 디지털 학술 문서의 시각적 복잡성이 증가함에 따라, 문서 내 텍스트뿐 아니라 문단, 표, 다이아그램 등 다양한 레이아웃 요소를 공간적으로 식별하는 기술이 중요한 과제로 자리잡았다[8].

최근 문서 이해 연구에서는 PubLayNet, DocBank와 같은 대규모 레이아웃 데이터셋이 구축되고[4][5], LayoutLM 계열과 같은 트랜스포머 기반 멀티모달 모델이 도입되면서 높은 수준의 레이아웃 추론 성능이 보고되고 있다[3]. 이러한 모델은 문서의 구조 및 문맥 정보를 효과적으로 반영할 수 있어, 자동화된 문서 아카이빙과 같은 응용 분야에서 활용되고 있다.

2.2 능동적 인출 학습과 디지털 학습 도구

학습자가 문맥 속에서 중요한 키워드를 스스로

가리고, 정답을 기억에서 인출하여 확인하는 능동적 인출 기반 학습은 학업 성취도 향상에 효과적인 것으로 보고되고 있다[7]. 이러한 개념을 디지털 교육에 적용하여, 자연어 처리 기술을 활용한 자동 빈칸 문항 생성 연구도 이루어지고 있다[9]. 이와 같은 접근을 기반으로, 기존 디지털 학습 도구는 크게 두 가지 유형으로 구분할 수 있다.

첫째는 Anki, Quizlet 등 간격 반복 시스템 기반의 디지털 플래시카드이다. 이러한 방식은 높은 암기 효율을 제공하지만[10], 학습자가 원본 문서의 내용을 단위별로 분할하여 수동으로 입력해야 하므로 문서의 문맥 정보가 유지되기 어렵다는 한계가 있다. 둘째는 GoodNotes, Notability, Adobe Acrobat과 같은 PDF 기반 필기 및 주석 도구이다. 이들 도구는 원본 문서의 시각적 구조를 유지할 수 있으나, 마스킹 학습을 위해서는 대상 단어를 사용자가 직접 선택하고 가리는 반복적인 수작업이 필요하다.

2.3 기존 접근 방식의 한계와 제안 방법

앞서 언급된 지능형 문서 처리 기술과 디지털 학습 보조 방식을 디지털 기반 PDF 환경에 적용할 경우, 기술적 효율성과 사용자 상호작용 측면에서 몇 가지 한계가 존재한다.

첫째, 딥러닝 기반 문서 이해 모델은 일반적으로 픽셀 데이터를 입력으로 하는 고비용 연산을 요구하며, 이는 대규모 사용자의 실시간 요청을 처리해야 하는 웹 환경에서 지연 시간과 운영 비용 증가로 이어질 수 있다[11]. 최근 경량화 연구가 진행되고 있으나, 본 연구의 대상인 전공 교재나 학술 논문과 같은 PDF 문서는 이미 단어의 위치 정보와 텍스트 메타데이터를 포함하고 있다. 따라서 이러한 정보를 활용하지 않고 별도의 추론 과정을 수행하는 것은 추가적인 연산 부담을 유발할 수 있다.

둘째, 디지털 플래시카드 기반 학습 방식은 원본 문서에서 텍스트를 분리하여 제시하기 때문에, 문서가 갖는 문맥 및 레이아웃 정보가 유지되기 어렵다는 한계가 있다.

셋째, 상용 PDF 도구를 활용하는 경우 원본 레이아웃은 유지할 수 있으나, 학습자가 직접 단어를 선택하고 가리는 반복적인 수작업이 요구되어 학습 효율을 저하시킬 수 있다.

이에 본 연구에서는 PDF 내부의 텍스트 및 위치 정보를 직접 활용하여, 원본 레이아웃을 유지하면서 자동으로 마스킹을 생성하는 학습 시스템 PDFMask를 제안한다.

III. 시스템 설계

본 장에서는 제안하는 PDF 마스킹 시스템의 구조와 문맥 인식 기반 마스킹 파이프라인의 알고리즘 설계를 기술한다.

3.1 대용량 문서 처리를 위한 비동기 아키텍처

그림 1은 비동기 방식으로 설계된 시스템 아키텍처와 처리 흐름을 보여준다. 시스템은 웹 계층(Django)과 엔진 계층(마스킹 엔진)으로 구성되며, 두 계층 간 태스크 메시지는 메시지 큐(Celery, redis)를 통해 전달된다. 본 시스템은 다수의 사용자가 동시에 대용량 PDF 문서의 마스킹을 요청하는 웹 환경을 전제로 설계되었다. PDF 마스킹 연산은 페이지 수와 텍스트 밀도에 따라 상당한 계산 자원과 처리 시간을 요구하므로, 웹 서버가 연산을 직접 수행하는 동기적 요청-응답 모델에서는 네트워크 타임아웃 및 서버 병목 현상이 발생할 수 있다.

이를 해결하기 위해 본 연구에서는 두 계층을 물리적으로 분리하는 비동기 메시지 큐 기반 아키텍처를 채택하였다. Request Handler는 사용자로부터 PDF 파일과 마스킹 설정을 수신하면, 업로드된 파일을 Shared Storage에 저장한 뒤 고유한 작업 식별자를 생성한다. 마스킹 설정은 시각화 방식, 마스킹 대상 전략, 그리고 전체 마스킹 비율로 구성된다. 이후 대용량 데이터 전송으로 인한 대기열 과부하를 방지하기 위해, Task Producer는 파일 경로, 작업 식별자, 그리고 마스킹 파라미터만을 포함하는 경량 태스크 메시지를 Message Queue에 등록하고 즉시 응답을 반환한다.

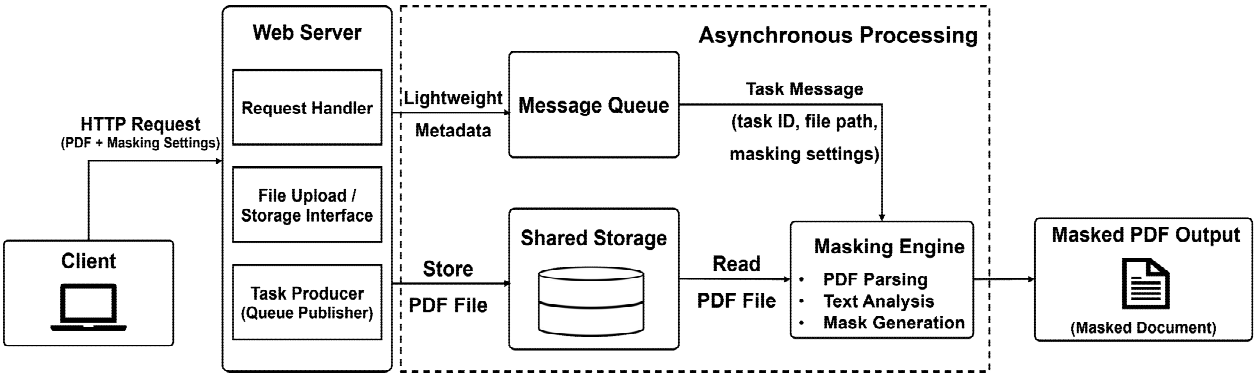


그림 1. 비동기 메시지 큐 기반 PDF 마스킹 시스템의 전체 아키텍처 및 처리 흐름

Fig. 1. Overall architecture and processing workflow of the asynchronous message queue-based PDF masking system

백그라운드에서 동작하는 Masking Engine은 Message Queue를 모니터링하여 새로운 태스크를 수신하며, 메시지에 포함된 경로를 이용해 Shared Storage에서 원본 PDF를 읽어 텍스트 파싱 및 형태소 분석 등의 연산을 수행한다. 이러한 구조는 파일 저장과 연산 처리를 분리함으로써 시스템 부하를 효과적으로 관리하고, 웹 서버와 연산 엔진의 역할을 명확히 구분한다. 또한 사용자 요청 증가에 따라 워커 엔진을 확장할 수 있어 수평적 확장성을 제공한다. 최종 마스킹 결과는 Shared Storage에 저장되며, 클라이언트는 상태 폴링을 통해 처리 완료 시점을 확인하고 마스킹된 PDF를 수신한다.

3.2 문맥 인식 기반 마스킹 생성 파이프라인

본 절에서는 텍스트 레이어가 존재하는 디지털 PDF 환경에서의 마스킹 생성 방법을 설명한다. 제안하는 파이프라인은 PDF의 기하학적 정보와 언어처리 기반 형태-구문 정보를 결합하여, 저지연 환경에서 학습에 적합한 마스킹 영역을 생성하는 것을 목표로 한다. 전체 과정은 텍스트 선형화 및 좌표 매핑, 형태-구문 기반 마스킹 후보 생성, 그리고 마스킹 샘플링 및 시각적 어노테이션 생성의 세 단계로 구성된다.

3.2.1 텍스트 선형화 및 좌표 매핑

디지털 PDF 문서는 시각적 레이아웃의 일관성을 유지하기 위해 텍스트가 의미 단위가 아닌 좌표 기반의 개별 객체로 저장된다. 이러한 특성으로 인해 자연어 처리 모델에 직접 활용하기 위해서는 논리적 읽기 순서를 복원하는 과정이 필요하다.

이를 위해 본 연구에서는 문서 레이아웃 정보를 분석하여 분절된 문자들을 문맥에 부합하는 선형 시퀀스로 재구성하는 텍스트 선형화 과정을 수행한다. i번째 페이지의 j번째 텍스트 라인 L_{ij}는 다음과 같이 정의된다.

$$L_{ij} = \{(c_k, b_k) | 1 \le k \le N\} \tag{1}$$

여기서 c_k는 k번째 문자의 유니코드 값이며, b_k는 해당 문자의 2차원 위치 좌표 (x_0, y_0, x_1, y_1)이고, N은 해당 텍스트 라인을 구성하는 전체 문자 수이다. 전체 PDF 문서 D는 다음과 같이 표현된다.

$$D = \{L_{ij} | 1 \le i \le P, 1 \le j \le M_i\} \tag{2}$$

여기서 P는 전체 페이지 수, M_i는 i번째 페이지의 텍스트 라인 수를 나타낸다.

선형화된 텍스트 시퀀스는 이후 형태소 분석의 입력으로 사용되며, 동시에 각 토큰의 위치 정보를 유지함으로써 분석 결과를 원본 PDF의 물리적 좌표 영역에 정확히 매핑할 수 있도록 한다. 이를 통해 텍스트 기반 처리 과정에서 발생할 수 있는 위치 정보 손실 문제를 보완하고, 실제 마스킹 단계에서 정밀한 영역 지정이 가능하도록 한다.

3.2.2 형태-구문 기반 마스킹 후보 생성

선형화된 텍스트 시퀀스는 형태소 분석기를 통해 분절 및 품사 태깅이 수행된다. 이후 문맥 정보와 구조적 특성을 반영한 다중 전략 기반의 마스킹 후보 생성 과정이 적용된다.

본 시스템은 사용자의 학습 목적에 따라 조사 기반 전략, 명사 기반 전략, 그리고 두 방식을 결합한 혼합 전략을 지원한다. 세 전략은 공통적으로 명사류 품사 토큰을 마스킹 대상으로 삼으며, 본 연구에서는 마스킹 대상 품사 집합 *Next*를 다음과 같이 정의한다.

$$Next = \{NNG, NNP, SL, SN, NR, XPN, XSN\} \quad (3)$$

여기서 NNG는 일반 명사, NNP는 고유 명사, SL은 외래어, SN은 숫자, NR은 수사, XPN은 접두사, XSN은 명사 파생 접미사를 의미한다.

조사 기반 전략은 한국어 문장의 구조적 기능어인 조사를 기준으로 핵심 명사구를 탐지한다. 특정 조사가 검출되면 해당 위치에서 역방향으로 탐색하여 Next에 속하는 태그가 연속되는 구간(Span)을 하나의 후보 단위로 확정한다. 이를 통해 문법적 역할이 명확한 핵심 개념 단위를 추출할 수 있다.

명사 기반 전략은 형태소 분석 과정에서 발생하는 과분절 문제를 완화하기 위해 토큰 시퀀스를 순방향으로 탐색하면서 Next에 속하는 태그가 연속되는 구간을 하나의 명사구 단위로 병합하여 복합 명사 및 도메인 특화 용어의 의미적 일관성을 유지한다.

혼합 전략에서는 두 전략의 구간 집합을 합산한 뒤 중복 및 포함 관계에 있는 구간을 제거(Dedup)하여 최종 후보 집합 M_{final}을 구성한다.

$$M_{final} = S_{context} \cup S_{structure} \quad (4)$$

여기서 $S_{context}$와 $S_{structure}$는 각각 조사 기반 전략과 명사 기반 전략을 통해 추출된 마스킹 후보 집합이다.

3.2.3 마스킹 샘플링 및 시각적 어노테이션 생성

추출된 마스킹 후보 집합에 대해 전체를 일괄적으로 마스킹할 경우 문서의 이해에 필요한 맥락 정보까지 손실될 수 있다. 이를 방지하기 위해 본 시스템은 사용자 설정에 따른 비율 기반 샘플링을 수행한다. 샘플링 과정에서는 지정된 마스킹 비율을 기준으로 후보 집합에서 일부를 선택하며, 매 반복 학습 시 서로 다른 마스킹 패턴이 생성되도록 하여 단순 암기 의존을 줄이고 학습 효과를 높인다.

최종 선택된 마스킹 대상은 앞선 좌표 매핑 정보를 기반으로 원본 PDF 상의 물리적 위치에 시각적으로 적용된다. 본 시스템은 두 가지 방식의 어노테이션을 지원한다. 하나는 해당 영역을 완전히 덮어 내용을 확인할 수 없도록 하는 차단 모드(Redact)이며, 다른 하나는 경계선만 표시하여 위치를 안내하는 강조 모드(Highlight)이다. 인접한 마스킹 영역은 시각적 일관성을 유지하기 위해 자동으로 병합되어 렌더링된다.

Ⅳ. 구현 결과

본 장에서는 구현된 PDFMask 시스템의 화면과 마스킹 처리 결과를 시연을 통해 보여준다. 표 1은 본 시스템의 구현 환경을 나타낸다.

표 1. 구현 환경
Table 1. Implementation environment

Item	Value
OS	Ubuntu 24.04.4 LTS
CPU	Intel Core i9-9900K @ 3.60GHz (8 cores / 16 threads)
RAM	54 GB
Python	3.11
Django	4.2.16
Celery	5.3.6
Redis	6.2.21
kiwipiepy	0.21.0
PyMuPDF	1.26.4

4.1 시스템 시연

그림 2는 PDFMask의 메인 페이지를 나타낸다. 상단 네비게이션 바에는 'PPT → PDF', 'DOCX →

PDF', '빠른 마스킹'의 세 가지 기능 메뉴가 배치되어 있으며, 사용자는 원하는 기능을 선택하여 해당 페이지로 이동할 수 있다.

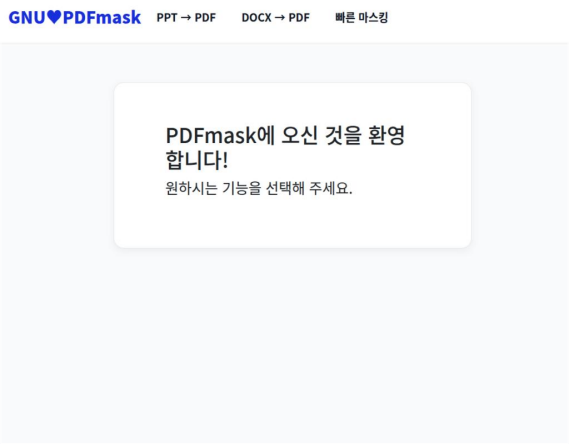


그림 2. PDFMask 시스템의 메인 페이지 인터페이스
Fig. 2. Main interface of the PDFMask system

그림 3은 마스킹 페이지의 입력 폼을 나타낸다. 사용자는 PDF 파일을 업로드한 뒤, 시각화 모드(Redact/Highlight), 타겟 모드(조사 기준/명사 기준/전체 사용), 마스킹 비율(기본값 0.95) 등의 옵션을 설정할 수 있다. 업로드 및 마스킹 실행 버튼을 누르면 파일이 서버로 전송되어 비동기 처리가 시작된다.

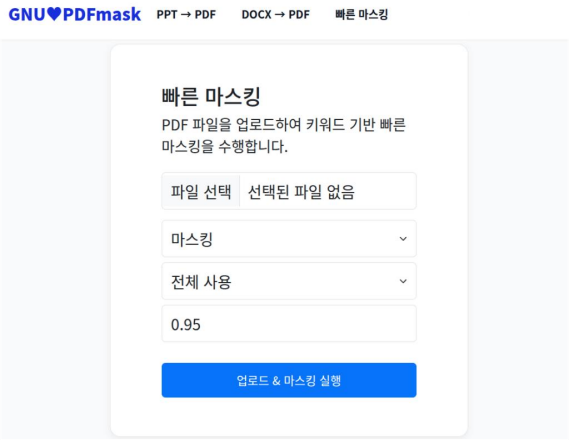
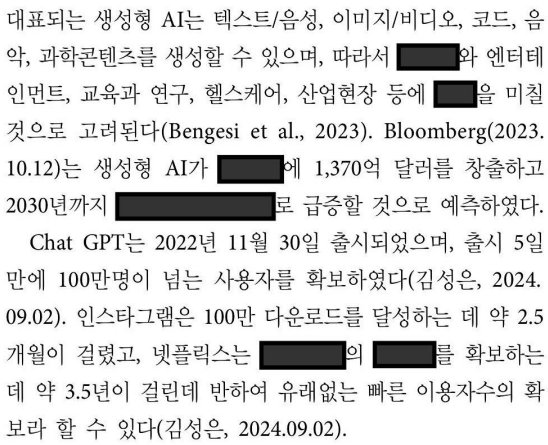


그림 3. 마스킹 옵션 설정
Fig. 3. Masking Options Settings

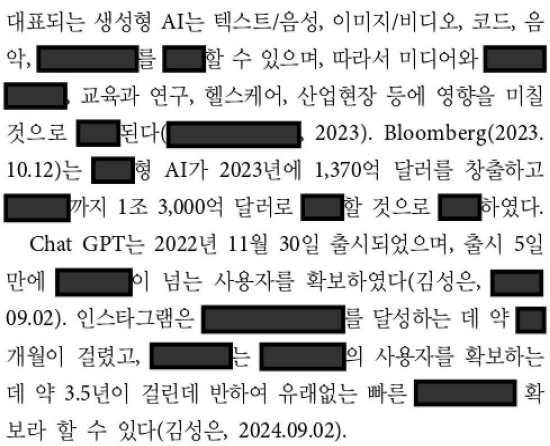
그림 4~7은 동일한 학술 논문 본문(마스킹 비율 0.3)에 타겟 모드별로 마스킹을 적용한 결과의 일부를 확대한 것이다. 그림 4~6은 Redact 모드로, 각각 조사 기준, 명사 기준, 전체 사용 타겟 모드에 해당하며, 그림 7은 Highlight 모드이다.

조사 기준(그림 4)은 핵심 체언만 선택적으로 은폐하여 마스킹 수가 적고 문장 구조 파악이 가능하므로 학습 초기 단계에 적합하다. 명사 기준(그림 5)은 조사 유무와 무관하게 모든 명사 연속 구간을 대상으로 하여 마스킹 범위가 넓으며, 전체 사용(그림 6)은 두 전략의 합집합으로 가장 포괄적인 마스킹을 제공한다. Highlight 모드(그림 7)는 추출된 마스킹 후보를 검정 박스 대신 빨간색 테두리로 강조 표시하여 원본 텍스트를 유지한 채 그 분포를 시각적으로 확인할 수 있다.


대표되는 생성형 AI는 텍스트/음성, 이미지/비디오, 코드, 음악, 과학콘텐츠를 생성할 수 있으며, 따라서 [redacted]와 엔터테인먼트, 교육과 연구, 헬스케어, 산업현장 등에 [redacted]을 미칠 것으로 고려된다(Bengesi et al., 2023). Bloomberg(2023. 10.12)는 생성형 AI가 [redacted]에 1,370억 달러를 창출하고 2030년까지 [redacted]로 급증할 것으로 예측하였다.

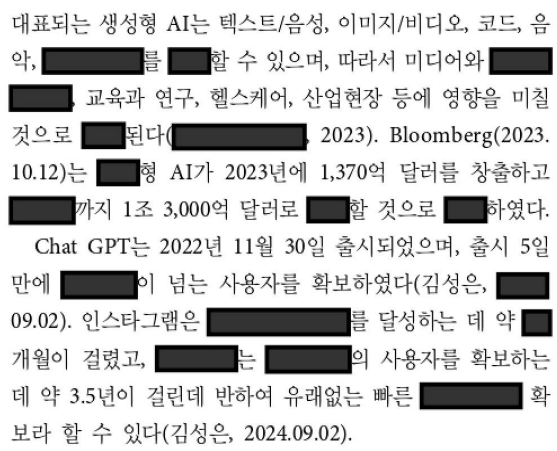
Chat GPT는 2022년 11월 30일 출시되었으며, 출시 5일 만에 100만명이 넘는 사용자를 확보하였다(김성은, 2024. 09.02). 인스타그램은 100만 다운로드를 달성하는 데 약 2.5 개월이 걸렸고, 넷플릭스는 [redacted]의 [redacted]를 확보하는 데 약 3.5년이 걸린데 반하여 유래없는 빠른 이용자수의 확보라 할 수 있다(김성은, 2024.09.02).

그림 4. 조사 기준 마스킹 결과 – Redact 모드
Fig. 4. Masking results using josa-based mode – redact mode


대표되는 생성형 AI는 텍스트/음성, 이미지/비디오, 코드, 음악, [redacted]를 [redacted]할 수 있으며, 따라서 미디어와 [redacted], 교육과 연구, 헬스케어, 산업현장 등에 영향을 미칠 것으로 [redacted]된다([redacted], 2023). Bloomberg(2023. 10.12)는 [redacted]형 AI가 2023년에 1,370억 달러를 창출하고 [redacted]까지 1조 3,000억 달러로 [redacted]할 것으로 [redacted]하였다.

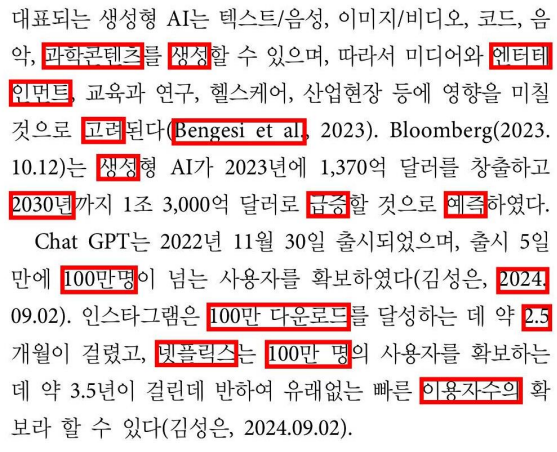
Chat GPT는 2022년 11월 30일 출시되었으며, 출시 5일 만에 [redacted]이 넘는 사용자를 확보하였다(김성은, [redacted] 09.02). 인스타그램은 [redacted]를 달성하는 데 약 [redacted] 개월이 걸렸고, [redacted]는 [redacted]의 사용자를 확보하는 데 약 3.5년이 걸린데 반하여 유래없는 빠른 [redacted] 확보라 할 수 있다(김성은, 2024.09.02).

그림 5. 명사 기준 마스킹 결과 – Redact 모드
Fig. 5. Masking results using noun-based mode – redact mode


대표되는 생성형 AI는 텍스트/음성, 이미지/비디오, 코드, 음악, [redacted]를 [redacted]할 수 있으며, 따라서 미디어와 [redacted], 교육과 연구, 헬스케어, 산업현장 등에 영향을 미칠 것으로 [redacted]된다([redacted], 2023). Bloomberg(2023. 10.12)는 [redacted]형 AI가 2023년에 1,370억 달러를 창출하고 [redacted]까지 1조 3,000억 달러로 [redacted]할 것으로 [redacted]하였다.

Chat GPT는 2022년 11월 30일 출시되었으며, 출시 5일 만에 [redacted]이 넘는 사용자를 확보하였다(김성은, [redacted] 09.02). 인스타그램은 [redacted]를 달성하는 데 약 [redacted] 개월이 걸렸고, [redacted]는 [redacted]의 사용자를 확보하는 데 약 3.5년이 걸린데 반하여 유래없는 빠른 [redacted] 확보라 할 수 있다(김성은, 2024.09.02).

그림 6. 전체 사용 마스킹 결과 – Redact 모드
Fig. 6. Masking results using combined mode – redact mode


대표되는 생성형 AI는 텍스트/음성, 이미지/비디오, 코드, 음악, 과학콘텐츠를 생성할 수 있으며, 따라서 미디어와 엔터테인먼트, 교육과 연구, 헬스케어, 산업현장 등에 영향을 미칠 것으로 고려된다(Bengesi et al., 2023). Bloomberg(2023. 10.12)는 생성형 AI가 2023년에 1,370억 달러를 창출하고 2030년까지 1조 3,000억 달러로 급증할 것으로 예측하였다.

Chat GPT는 2022년 11월 30일 출시되었으며, 출시 5일 만에 100만명이 넘는 사용자를 확보하였다(김성은, 2024. 09.02). 인스타그램은 100만 다운로드를 달성하는 데 약 2.5 개월이 걸렸고, 넷플릭스는 100만 명의 사용자를 확보하는 데 약 3.5년이 걸린데 반하여 유래없는 빠른 이용자수의 확보라 할 수 있다(김성은, 2024.09.02).

그림 7. 마스킹 결과 – Highlight 모드
Fig. 7. Masking results using the highlight mode

4.2 성능 평가

본 시스템의 처리 성능을 평가하기 위해 페이지 수 및 타겟 모드별로 5회 반복 측정을 수행하였다. 본 시스템은 딥러닝 추론 없이 규칙 기반 형태소 분석을 사용하므로 GPU(Graphics Processing Unit) 환경이 불필요하며, 실제 배포 환경과 동일한 CPU (Central Processing Unit)단독 환경에서 실험을 수행하였다.

표 2는 동일 조건에서 5회 반복 측정한 결과이다. 전 구간에서 표준편차가 0.14초 이하로 나타나 측정 결과의 일관성이 확인되었다. 명사 기준 및 전체 사용 모드는 페이지당 약 0.6초, 조사 기준 모드는 페이지당 약 0.25초의 일관된 처리 속도를 보였다. 이는 페이지 이미지 렌더링이나 OCR 과정 없이 PDF에 내장된 텍스트 레이어를 PyMuPDF로 직접 파싱하고, 경량 규칙 기반 형태소 분석기(Kiwipiepy)만을 사용하는 설계에 기인한다. 또한 조사 기준 모드가 명사 기준·전체 사용 모드 대비 처리 시간이 짧은 것은, 명사 연속 구간 탐색 없이 조사 역방향 탐색만 수행하기 때문이다. 마스킹된 키워드의 학습 효과성에 대한 정성적 평가는 추후 연구에서 피험자 실험을 통해 수행할 예정이다.

표 2. 처리 시간 측정 결과 (단위: 초)
Table 2. Processing time measurement results (unit: sec)

Pages	Particle-based	Noun-based	Combined
1	0.33 ± 0.01	0.61 ± 0.00	0.68 ± 0.14
5	1.23 ± 0.01	3.06 ± 0.01	3.06 ± 0.01
10	2.15 ± 0.01	6.19 ± 0.13	6.19 ± 0.14
20	4.00 ± 0.02	12.29 ± 0.14	12.29 ± 0.13

처리 성능 외에, 구현 방식 및 적용 특성 측면에서 기존 연구와 비교하면 표 3과 같다.

표 3. 기존 연구 비교
Table 3. Comparison with related work

Criterion	Matsumori et al. [9]	PDFMask
Candidate selection	MLM-based	Rule-based morphological analysis
Input format	Text corpus	PDF
Layout preservation	X	O
GPU required	O	X
Target language	English	English & Korean
Visual annotation	X	O
Learning purpose	Vocabulary learning	Textbook retrieval learning

V. 결론 및 향후 과제

본 논문에서는 디지털 PDF 환경에서의 학습 효율 향상을 위해 문맥 인식 기반 키워드 마스킹 학습 시스템 PDFMask를 제안하였다. 제안 시스템은 PDF의 레이아웃 및 좌표 정보를 활용하여 텍스트의 논리적 읽기 순서를 복원하고, 형태소 및 형태-

구문 분석을 기반으로 핵심 키워드를 자동으로 식별한다. 이후 비율 기반 샘플링과 시각적 어노테이션을 통해 원본 문서의 레이아웃을 유지하면서도 학습에 적합한 마스킹 PDF를 생성한다.

또한 웹 서버와 처리 엔진을 분리한 비동기 메시지 큐 기반 아키텍처를 적용함으로써, 대용량 PDF 처리 환경에서도 효율적인 시스템 운영이 가능하도록 설계하였다. 이를 통해 기존 플래시카드 기반 학습에서 발생하는 문맥 정보 손실 문제와 PDF 어노테이션 기반 학습의 수작업 부담을 완화할 수 있다.

향후 연구에서는 OCR 기반 텍스트 추출 기술을 도입하여 스캔 문서 및 이미지 기반 PDF까지 처리 범위를 확장하고, 단어의 문맥적 중요도를 반영하는 AI 기반 마스킹 전략을 적용함으로써 보다 효과적인 학습 지원이 가능하도록 발전시킬 예정이다.

References

[1] D. Doermann, "The indexing and retrieval of document images: A survey", Computer Vision and Image Understanding, Vol. 70, No. 3, pp. 287-298, Jun. 1998. https://doi.org/10.1006/cviu.1998.0692.

[2] R. Smith, "An overview of the Tesseract OCR engine", Proc. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, Vol. 2, pp. 629-633, Sep. 2007. https://doi.org/10.1109/ICDAR.2007.4376991.

[3] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of text and layout for document understanding", Proc. 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, pp. 1192-1200, Aug. 2020. https://doi.org/10.1145/3394486.3403172.

[4] X. Zhong, J. Tang, and A. J. Jimeno-Yepes, "PubLayNet: largest dataset ever for document layout analysis", Proc. International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, pp. 1015-1022, Sep. 2019. https://doi.org/10.1109/ICDAR.2019.00166.

[5] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou, "DocBank: A benchmark dataset for document layout analysis", Proc. 28th International Conference on Computational Linguistics (COLING 2020), Barcelona, Spain, pp. 949-960, Dec. 2020. https://doi.org/10.18653/v1/2020.coling-main.82.

[6] Z. Shen, R. Zhang, M. Dell, B. C. G. Lee, J. Carlson, and W. Li, "LayoutParser: A unified toolkit for deep learning based document image analysis", Proc. International Conference on Document Analysis and Recognition, Lausanne, Switzerland, pp. 131-146, Sep. 2021. https://doi.org/10.1007/978-3-030-86549-8_9.

[7] Z. L. Buchin and N. W. Mulligan, "Retrieval-based learning and prior knowledge", Journal of Educational Psychology, Vol. 115, No. 1, pp. 22-35, Jan. 2023. https://doi.org/10.1037/edu0000735.

[8] A. Antonacopoulos, S. Pletschacher, D. Bridson, and C. Papadopoulos, "ICDAR 2009 page segmentation competition", Proc. 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, pp. 1370-1374, Jul. 2009. https://doi.org/10.1109/ICDAR.2009.275.

[9] S. Matsumori, K. Okuoka, R. Shibata, M. Inoue, Y. Fukuchi, and M. Imai, "Mask and cloze: Automatic open cloze question generation using a masked language model", IEEE Access, Vol. 11, pp. 9835-9850, Jan. 2023. https://doi.org/10.1109/ACCESS.2023.3239005.

[10] B. Tabibian, U. Upadhyay, A. De, A. Zarezade, B. Schölkopf, and M. Gomez-Rodriguez, "Enhancing human learning via spaced repetition optimization", Proc. of the National Academy of Sciences (PNAS), Vol. 116, No. 10, pp. 3988-3993, Mar. 2019. https://doi.org/10.1073/pnas.1815156116.

[11] I. Grigorik, "High Performance Browser Networking: What every web developer should know about networking and web performance", O'Reilly Media, Inc., 2013.

저자소개

김 민 준 (Minjun Kim)

2021년 3월 ~ 현재 :
경상국립대학교 컴퓨터공학과
학사과정
관심분야 : 인공지능, 컴퓨터비전

이 가 은 (Ga Eun Lee)

2023년 3월 ~ 현재 :
경상국립대학교 컴퓨터공학과
학사과정
관심분야 : 확장현실, 증강현실

김 지 원 (Jiwon Kim)

2026년 2월 : 경상국립대학교
컴퓨터공학과(공학사)
2026년 3월 ~ 현재 : ㈜메쎄이상
매니저
관심분야 : 데이터 분석, 가상현실,
클라우드 컴퓨팅

강 창 구 (Changgu Kang)

2010년 2월 : 광주과학기술원
정보기전공학부(공학석사)
2017년 8월 : 광주과학기술원
전기전자컴퓨터공학부(공학박사)
2018년 3월 ~ 현재 :
경상국립대학교 컴퓨터공학부
부교수
관심분야 : 컴퓨터 그래픽스, 증강현실, 인공지능