●○ 요약 데이터 과제

도서자료 요약 데이터

●○ 개요: 도서요약 데이터셋이란?

자연어처리(NLP, Natural Language Processing) 분야에서 가장 도전적인 분야인 생성요약(Abstractive Summarization) 기술 개발에 활용할 수 있는 학습 데이터셋으로 ㈜딥네츄럴에서 구축했으며, 20만건의 한국어 도서자료 원문과 요약문으로 구성되어 있다.

생성요약 기술은 사람이 본문에서 핵심 내용을 파악하고 이를 요약하는 것처럼 주어진 텍스트를 요약하는 기술이며, 본문의 내용을 그대로 사용하는 추출요약과 달리 완성된 하나의 새로운 요약문을 생성한다는 점에서 난이도가 높은 과제로 손꼽힌다. 텍스트 정보가 시각매체에서 스마트스피커, 커넥티드카 등 음성 및 영상 매체로 확장됨에 따라 많은 양의 텍스트에서 빠르게 정보를 추출하여 전달할 수 있는 요약 서비스에 대한 수요가 증가하고 있어 상용화에 대한 연구가 활발히 진행되고 있다.

생성요약의 사례 대해서는 아래를 참고할 수 있다.

원문	생명공학기술, 특히 유전자 재조합 기술에 의하여 창출되는 LMOs 및 그 제품에 의한 인간의 건강 및 자연생태계의 비가역적 파괴를 사전에 방지할 수 있는 추진 방향으로 실험실에서의 안전한 생명공학기술의 적용과 LMOs 및 그 제품의 안전한 환경도입단계로 구분하여 설정할 수 있다. 실험실에서 안전한 생명공학기술의 적용을 위해서는 유전자재조합 기술에 의하여 LMOs가 창출되고 연구되는 실험실에서의 안전성이 유지되어야한다. LMOs의 환경도입단계는 창출된 LMOs 및 그 제품의 연구과정으로서 LMOs의 환경적응능력과 연구자가 원하는 LMOs의 능력발현을 검정할 수 있는 환경도입실험단계, LMOs 및 그 제품의 국내·외간 교역단계와 운송단계로 구분할 수 있는데, 생명공학기술의 안전성은 각 단계별로 구분하여 안전한 생명공학기술의 실험, LMOs 및 그 제품의교역 및 운송체계를 유지함으로써 달성할 수 있다. 본 연구의 목적은 생명공학기술의 응용에 대한 자연생태계의 안전성 유지체계에서 LMOs 및 그 제품에 의한 자연생태계의 비가역적 파괴를 사전에 방지할 수 있도록 LMOs 및 그 제품의 환경도입 이전에 LMOs 및 그 제품이 우리나라의 자연 생태계에미칠 수 있는 영향을 사전에 파악할 수 있는 LMOs 및 그 제품의 환경 영향 평가방법 및 평가(기술)지침, 이를 체계적으로 시행할 수 있는 평가제도를 제안하는 데 있다. 특히, LMOs 및 그 제품의 환경영향 평가방법은 UN기관 및 선진외국에서 연구된 생명공학안정성에 대한 평가방법이 체계적으로 분석되어 우리나라의 현실에 적용할 수 있도록 제시하는 데 있다. 즉, LMOs 및 그 제품에 의한 우리나라 생태계의 비가역적인 훼손방지 및 이에 따른 국민들의 불안감을 해소하고 대외적으로 우리나라 생명공학산업의 국제경쟁력 확보와 세계시장으로 진출할 수 있도록 양면을 고려하여 현실적으로 적용할 수있는 LMOs 및 그 제품 환경영향 평가방법과 평가지침을 제안하는 데 있다. 또한, 제시된 LMOs 및 그 제품의 환경영향 평가방법과 평가지침을 체계적으로 시행할 수 있는LMOs 및 그 제품의 환경영향 평가제도를 제안하는 데 있다.

요약문	유전자 재조합 기술의 건전한 추진 방향은 실험실에서의 안전한 기술 적용과 제품의 안전한 환경도입단계로 구분할 수 있으며, 환경도입단계는 다시 환경도입실험단계, 국내외간 교역단계, 운송단계로 나누어진다. 본 연구는 자연생태계 안정성 유지체계를 위한 제품의 환경 영향 평가방법, 평가지침, 평가제도 제안을 목적으로 한다. 특히 환경영향 평가방법은 외부 평가방법 분석 후 우리나라 현실에 적용하는 데 목적이 있다.

그림1 | 기계 독해의 사례

●○ 데이터셋의 구성

본 데이터셋은 도서 자료로부터 추출한 20만개의 문단과 각 문단에 대한 요약문 20만건으로 구성되어 있으며, 각 문단과 요약문이 쌍으로 이루어져 있다.

데이터 종류	포함 내용	제공 방식
생성요약 데이터셋	원문과 요약문(20만 건)	JSON 포맷 파일

●○ 데이터셋의 설계 기준과 분포

데이터셋을 설계 시에는 보건사회, 생명, 조세, 환경, 지역사회 개발, 무역, 경제, 노동 등 다양한 분야의 자료가 포함되도록 하였으며, 저작권 이슈가 해결된 도서자료를 대상으로 구축하였다.

[데이터셋 구축 완료 후 작성 예정]

그림2 | 데이터셋 구성 개요

본문, 질문, 답변의 구성 원칙과 주요 특징은 다음과 같다.

- 원문: 전체 약 3만건의 도서자료 중 플레인 텍스트로 추출이 가능한 자료를 대상으로 개조식 문서, 이미지나 표가 대부분인 문서, 국한문 혼용 문서 등을 제외하여 요약이 가능한 문서를 활용해 데이터를 구축했다. 또한 엔진 학습이 가능하도록 각 문단의 길이는 300-1000자로 한정했다.

[데이터셋 구축 완료 후 작성 예정]

그림3 | **본문 주제별 분류 분포**

- 요약문: 각 원문 문단에서 주요한 문장을 추출하고 이를 자연스럽게 연결하고 요약하는 방법으로 단순히 본문을 복제하는 추출요약과는 다른 데이터를 구축했다. 또한 엔진이 주요 내용을 찾아낼 수 있도록

●○ 데이터 구조

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

No.	속성명	영문명	길이	타입	필수여부	설명
1	원문ID	passage_id	100	String	Y	원문에 부여되는 고유번호 '문서ID_분리순서' 포맷
2	메타데이터	metadata	-	Object	Y	서지 정보에서 추출한 메타데이터
2-1	문서ID	doc_id	100	String	Y	
2-2	문서유형	doc_type	100	String	Y	'도서' 및 '논문'으로 문서유형 구분
2-3	문서명	doc_name	100	String	Y	
2-4	발행자	author	100	String	N	
2-5	발행처	publisher	100	String	N	
2-6	발행일	published_date	8	Date	N	
2-7	주제분류	kdc_label	100	String	Y	해당 원문의 KDC 분류명
2-8	분류기호	kdc_code	3	String	Y	해당 원문의 KDC 분류코드
3	챕터	chapter	100	String	N	해당 원문이 소속된 챕터명
4	원문	passage	1000	String	Y	구축 대상 원문 문단
5	요약문	summary	300	String	Y	원문 문단에 대한 생성요약

●○ 데이터 예시


```
{
    "passage_id":"123456_0001"
    "metadata":
        {
            "doc_id":"123456",
            "doc_type":"도서",
            "doc_name":"북미정상회담:창조적 블랙홀이 될 것인가?",
            "author":"정성윤",
            "publisher":"통일연구원",
            "published_date":"2018-03-21",
            "kdc_label":"사회과학",
            "kdc_code":"300"
        },
    "chapter":null,
    "passage":"최근 정세 변동의 의미와 평가 북핵 문제는 지난 25 년 동안 다양한 철학과 접근법 그리고 전략의 동원에도
    물리지 않았던 난제 중 난제이다. 그 결과 북핵 문제는 한반도와 동북아의 모든 이슈를 삼키고 가두어 버리는 블랙홀이 되어
    버렸다. 그러나 북핵 문제가 새로운 국면에 진입하고 있다. 결정적 계기는 향후 개최될 남북 정상회담과 북미 정상회 담이 될
    것이다. 특히 북핵 위기 25 년 만에 처음으로 북미 정상이 직접 담판을 하게 됨으로써, 북핵 문제의 획기적 진전에 대한
    기대가 높아지고 있다. 두 차례 정상회담이 합의된 가장 큰 배경은 북한의 태도 변화, 미국의 대화 호응, 우리 정부의 강력한
    남북관계 진전 의지와 외교력이다. 이 중 김정은 스스로가 비핵화 의지를 밝힌 것이 정세 변화의 핵심이자 촉발 동인이다. 이러한
    북핵 정세 변화 가 추동하고 있는 구조적 변화 양상, 이를 가능하게 만든 북한의 전략전환 이유, 북미 정상회담 전후의 정세
    방향, 그리고 한국의 정책적 고려사항을 제시한다.",
    "summary":"북핵 문제는 지난 25 년 동안 다양한 노력에도 해결되지 않은 난제로 한반도와 동북아를 모두 아우르는 주요
    이슈가 되었다. 그러나 북핵 문제가 남북 정상회담과 북미 정상회담을 계기로 새로운 양상에 돌입한다. 이는 북한, 미국, 우리
    정부의 태도 변화 및 대화 의지가 바탕이 되었고, 이 중 특히 김정은의 비핵화 의지 표명이 가장 결정적이다. 이러한 구조적
    변화와 숨은 의도, 예상 정세 및 한국의 정책 고려사항을 제시한다."
}
```


[데이터 수령 후 실제 데이터 샘플로 대체 예정]

●○ 데이터 구축 과정

데이터 구축은 2020년 10월부터 2020년 2월까지 국립중앙도서관으로부터 약 5만권의 저작권 해결된 도서를 제공받아 이 중 요약이 불가능한 원문을 제거하는 정제 작업과 정제작업을 거친 원문의 요약 작업을 거쳐 20만개의 요약데이터를 만들어 냈다.

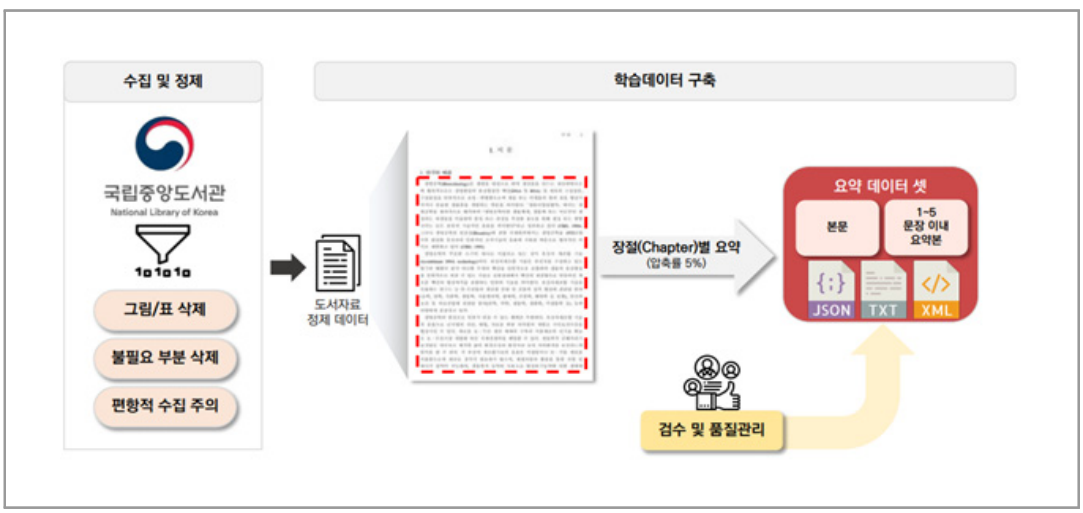

그림4 | 데이터 정제와 분포 균등화

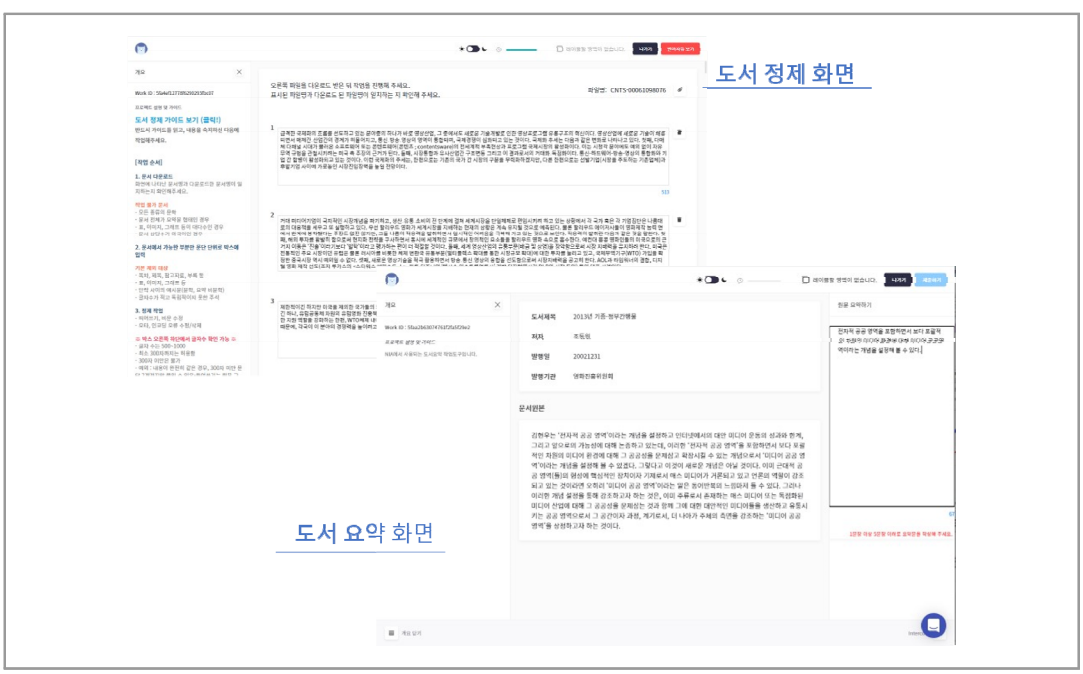

그림5 | 효율적인 데이터 제작을 위한 웹 기반 툴

작업 화면에서 입력시의 오류를 줄이기 위하여 작업 화면을 주기적으로 개선하였으며 일관화 된 가이드를 제공하여 작업의 효율성을 높이는데 노력하였다.

도서 정제 시에도 가이드라인을 만들어 작업을 진행하였으며, 요약 작업에서는42MARU 주관으로 연세대학교, 경북대학교에서 나온 요약 가이드라인을 통하여 요약문의 완성도를 높이는 작업을 진행하였다.

요약 작업의 기준은 ①원문, ②주요문장 선별, ③불필요한 수식어구 제거, ④간략한 표현으로 축약 및 유사 표현으로 변경, ⑤자연스럽게 연결되도록 수정, ⑥요약문 완성의 순으로 방법을 설정한 가이드 문서를 만들고, 각각의 영역에서의 예시문과 함께 제공하여 이해를 높여 작업을 진행하였다.

●○ 검수와 품질 확보

대량의 데이터를 높은 품질로 생성하기 위하여, 프로세스를 몇 단계로 분리하여 작업을 진행하였다. 첫번째로는 품질 관리가 어려운 크라우드의 품질 관리를 위해서 “튜토리얼 프로젝트”를 운영하여 우수 작업자를 확보하는데 활용하였다. 여기서 확보된 우수작업자를 통하여 “도서 정제” 프로젝트를 진행하고, 이를 기반으로 빠른 시간에 양질의 데이터를 다수 확보할 수 있었다. 요약 작업을 진행할 때는, 별도로 국문학과 출신의 크라우드 작업자를 별도로 선별하여 요약작업의 품질을 높일 수 있었다.

검수는 총 3단계로 진행하여 진행하였으며, 각 단계별로 내부 작업자와 외부 크라우드 소싱 작업자를 투입하여 검수 작업의 속도를 높일 수 있었다. 이렇게 만들어진 데이터셋의 총괄 담당자는 국문학과 출신으로 구성하여 가이드셋의 적정성 확인, 원문 문서의 선별, 크라우드 작업 가이드라인 생성, 크라우드를 통한 데이터의 관리의 전문성을 확보할 수 있었다.

다수의 데이터를 가공하는 데에는 웹기반 플랫폼도 상당한 기여를 하였다. 전체적인 작업 일정 확인, 우수 가입자 확보 및 불량 가입자 제외, 총 데이터 셋의 관리등을 플랫폼 기반으로 진행하여, 좀더 빠르고 정확하게 프로젝트를 수행할 수 있었다.

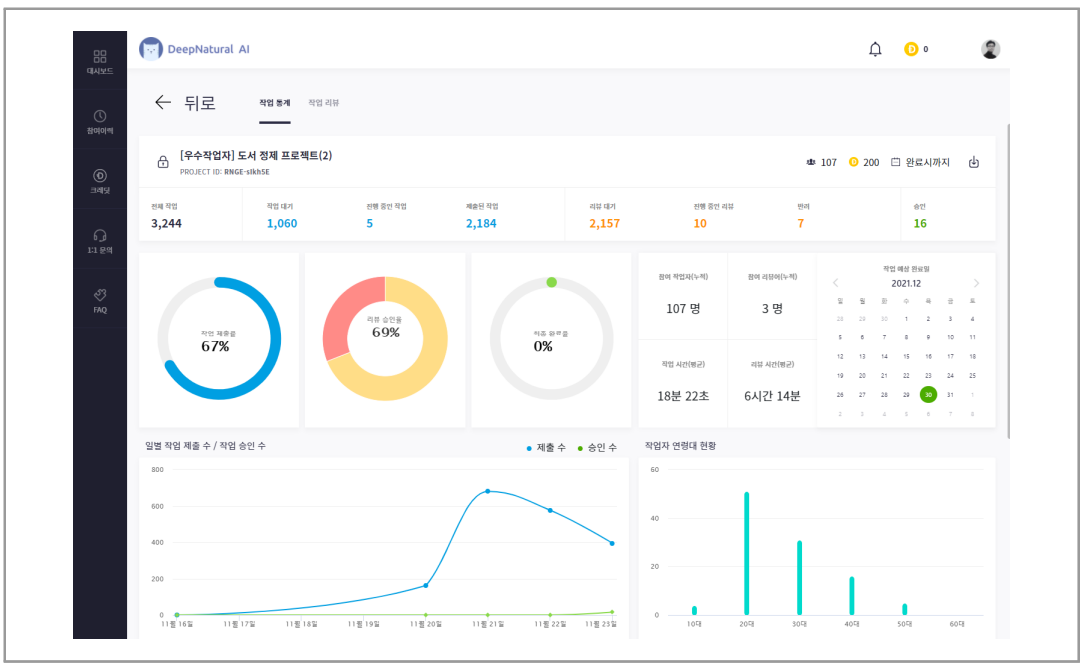

그림6 | 플랫폼 기반의 작업 도구

●○ 데이터 구축 담당자

수행기관(주관) : ㈜딥네츄럴
(전화: 6952-0588, 이메일: contact@deepnatural.ai)