●○ 요약 데이터 과제

논문자료 요약 데이터

●○ 개요: 논문 요약 데이터셋이란?

텍스트 요약은 컴퓨터 프로그램을 이용해 문서의 축약된 내용을 자동으로 생성하는 것을 의미하며 딥러닝에 기반한 자연어 이해와 생성 기술이 필요하다. 논문 요약의 경우 지속적으로 연구되어온 추출 요약 방식으로 요약 대상 문단을 선정, 문장에 대한 속성 값을 부여하고, 생성요약을 병행하는 방안을 수립하여 구축했으며 논문 전체에 대한 요약과, 섹션별 요약 대상 원문과 생성 요약문 35만 건을 쌍으로 제작하여 논문의 구조와 특성을 반영하였다. 논문의 수집과 정제, 가공은 ㈜나라지식정보와 ㈜단아코퍼레이션이 공동으로 수행하였고 데이터 모델링과 활용서비스 개발은 ㈜포티투마루에서 담당하였다.

현재까지는 제목이나 초록, 저자 등 키워드만으로 검색해 정형화된 정보를 얻을 수밖에 없고 관심 있는 논문의 연구 주제나 연구 방법, 연구 결과를 파악하려면 일일이 원문을 직접 읽어보고 판단해야 했지만 논문 요약 데이터셋 구성과 활용 서비스를 이용하여 AI가 논문을 대신 읽고 요약해주는 서비스 개발을 목표로 활발한 연구와 다양한 서비스 개발을 진행하고 있다.

기계 독해 기술은 질의응답 서비스에 주로 사용되며, 사람이 맥락을 이해하고 논리적으로 답을 찾는 것처럼 질의에 대한 답을 찾는 기술이며, AI 챗봇 상담, 방대한 전문지식에 대한 질의응답 및 시맨틱 검색 등에 활용하기 위해 활발한 연구 및 상용화를 진행하고 있다.

논문 요약 사례 대해서는 아래를 참고할 수 있다.

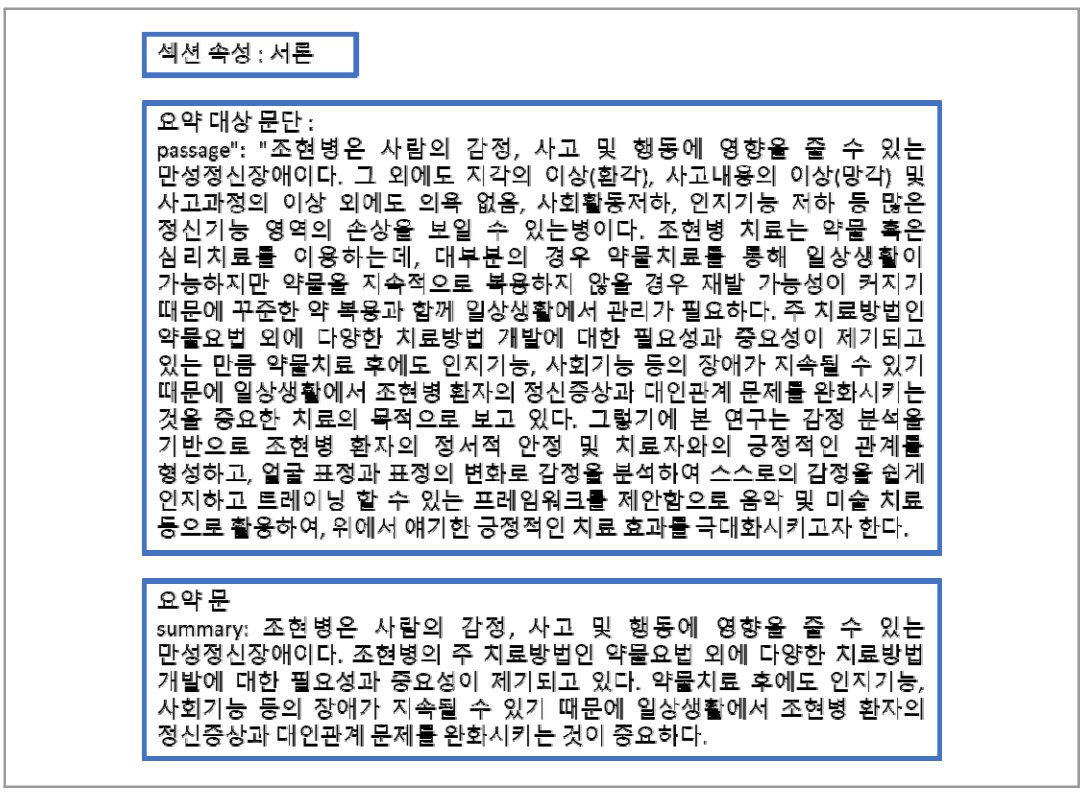

그림1 | 논문 요약의 사례

●○ 데이터셋의 구성

논문요약 데이터셋은 논문 전체에 대한 요약과 섹션별 요약 35만 건을 제작하며 각각의 데이터셋은 메타데이터와 전체 요약문, 섹션별 요약문으로 구성되어 있다. 메타데이터는 학술 논문의 제목, 저자, 발행처, 발행년, 연구재단 학술 분류 등의 기본적인 정보를 포함하고 있으며 전체 요약문 35만 건은 최근 10년간의 다양한 분야의 학술 논문을 포괄하고 연구주제, 배경, 방법, 결론 등의 속성을 부여하여 논문의 검색과 활용이라는 서비스를 제공하기에 충분한 양이며 섹션별 요약 35만 건은 인공지능을 통한 요약문 생성 모델링과 활용 서비스를 제공할 수 있도록 했다.

데이터셋 구성	포함 내용	제공 방식
메타데이터	저자, 제목, 발행처, 발행년, 학술 분류 등 (35만 건)	JSON 포맷 파일
전체 요약	학술 논문 초록을 활용하여 속성 값을 부여(35만 건)	
섹션별 요약	논문의 핵심적 문단과 문단에 대한 생성요약문(35만 건)	

●○ 논문 요약 데이터셋의 분야별 분포와 구성기준

수집 대상 학술 논문의 원천 데이터는 Open Access를 통해, 저작물의 이용허락을 확보한 논문과 과학기술정보통신연구원(KISTI), 국립중앙도서관 등의 학술 논문 자료 서비스 통해 35만 건을 확보하였다.

또한 한국연구재단(nrf.re.kr)의 학술논문의 분류 체계를 바탕으로 특정 주제 혹은 문서 유형, 작성자 및 성향에 편향되지 않도록 다양한 분야의 논문을 수집 대상으로 선정하였으며 수집한 학술논문의 분야별 수량은 아래 표와 같다.

인문학	사회과학	자연과학	공학	의약학	농수해양학	예술체육학	복합학	계
65,000	166,700	8,500	36,000	1,300	5,000	28,000	39,500	350,000

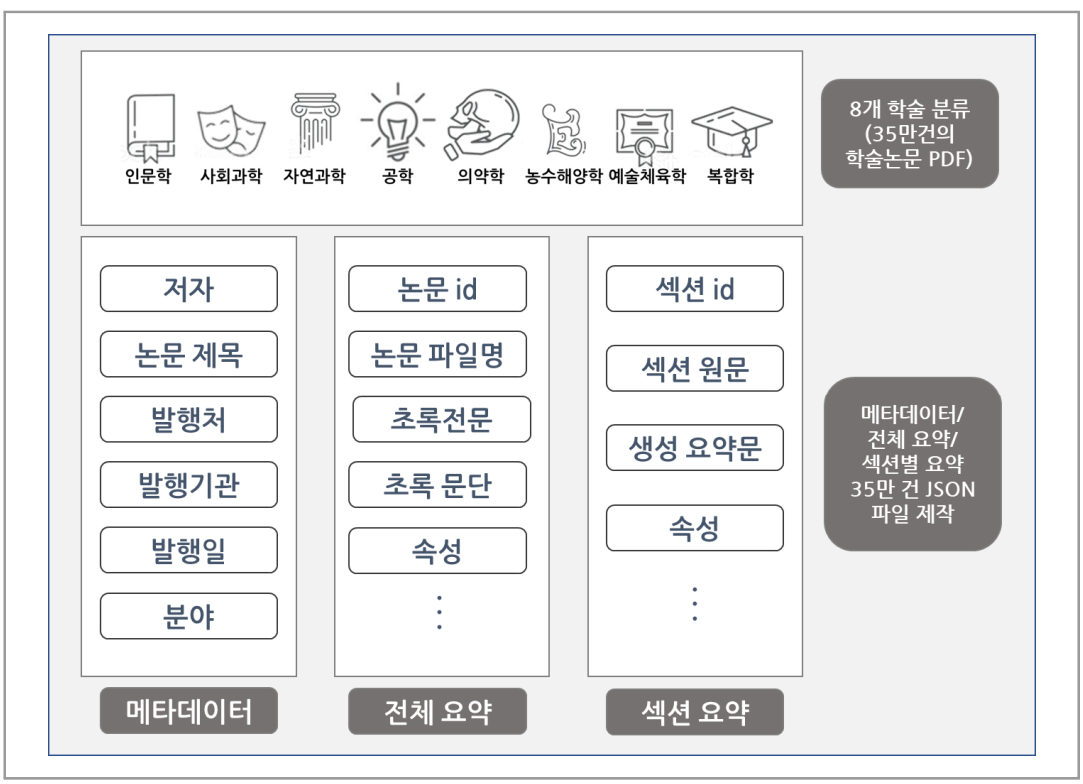

그림2 | 데이터셋 구성 개요

학술논문 데이터셋 구성 원칙과 주요 특징은 다음과 같다.

- 논문의 작성 시 일반적으로 IMRaD 포멧(Introduction, Materials and Methods, Results and Discusion, Conclusion)을 따르는 것이 표준이라고 할 수 있음. 이러한 체계를 따르지 않은 학술 관련 간행물은 학술논문(Academical Paper)가 아닌 에세이(Essay)라고 하며, 이러한 에세이는 학계의 대가(大家)가 발표한 몇몇 예외를 제외하고 일반적으로 학술적 가치 또는 연구업적으로 인정하지 않는다.
- IMRaD 포맷은 서론부에 해당하는 도입(Introduction)에서, 학자의 연구를 촉발시킨 문제 발생의 배경(연구배경)과 연구의 목표(연구목적) 및 연구의 필요성을 기술하며, 본론에 해당하는 이론적 배경(기존연구), 연구방법, 연구결과 및 토론에서는 각각 연결성이 있는 기존의 연구들에 대한 설명과 해당 연구들과 비교했을 때 자신의 연구가 갖는 독창성, 연구목적을 달성하기 위해 기획한 연구의 추진 절차 및 수단(연구방법), 연구방법을 객관적으로 적용하여 얻어진 결과물(연구결과), 그리고 연구결과에 대한 분석(토론)으로 구성된다.
- 결론에 해당하는 연구결론(Conclusion)에서는 연구결과가 연구목적에서 설정한 목표를 달성했다는 근거를 제시할 수 있는지와 향후에 다른 연구자가 자신의 연구성과(논문)을 참고하여 다른 새로운 연구를 수행함에 있어 어떻게 기여할 수 있는지 등을 기술한다.
- 위에서 제시된 논문의 구성요소 중 이론적 배경(기존연구)은 현재 논문의 참고문헌에 기록된 기존 연구들에 대한 설명이 나열되는데, 논문에 따라 연구자의 성명과 해당 논문의 발표연도를 대명사와 같이 사용하거나(예시: '~이상경(1999)에 따르면~'), 제출한 학술저널의 양식에 따른 참고문헌 번호를 사용하는 등 AI 텍스트 학습에 적합하지 않으므로 배제하는 것이 타당하다..
- 서론에 해당하는 도입부는 '서론', '들어가며', '도입', '마중글' 등 다양한 이름으로 챕터명을 작성하기도 하는데, 이러한 사례는 '실험방법', '맺는 글' 등과 같이 본론과 결론에 나타난다.
- 따라서 다양한 논문에 대해 포괄적으로 사용할 수 있는 내용적 속성은 챕터명이 아닌 내용적 섹션으로 판단하는 것이 타당하며, '연구배경', '연구목적', '연구방법', '연구결과', '연구결론'으로 구성하는 것이 적절하다.

논문 주제와 특성에 기반에 전체 요약 및 섹션별 요약

- 학술 논문의 초록은 해당 논문에 가장 이해도가 높은 저자가 연구의 배경, 목적, 방법, 결과 등을 핵심적으로 요약하여 기술한 가장 완성도가 요약문이라 할 수 있다. 특히 연구목적이나 연구방법 등 본문의 기술적 내용을 강하게 함축해야 하는 점에 있어서 해당 학술 논문에 대해 가장 전문적 지식을 보유했다고 가정할 수 있는 저자 요약문인 초록을 학습데이터로 활용하는 것이 적절하다.
- 전문분야 내지 학술연구는 연구주제 설정 → 연구모델(연구방법) 설계 → 선행조사 → 연구수행 → 연구결과의 절차로 진행되는 것이 전형적인 방식이며 따라서 초록의 문장을 연구배경, 연구목적, 연구방법, 연구결과로 구분하여 속성(Role)을 부여하는 것이 논문 전체 요약의 핵심이라 할 수 있다.

●○ 데이터 구조

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

분류	메타데이터	전체 요약	섹션별 요약
내용	저자, 제목, 발행처, 발행기관, 발행일 등	논문ID, 논문 파일명, 초록전문, 속성 등	섹션 ID, 섹션 원문, 생성 요약문, 속성
수량	35만	35만	35만

구분	요소명		예시	유형	길이	필수 여부	설명
	한글	영문명					
메타 데이터	아이디	doc_id	002318564	string		Y	논문 1편을 기준으로 한 id
	논문(연구)명	title	불법 산양삼 검출을 위한 인공지능 기술에서의 산양삼과 인삼 이미지의 분류 기저화 연구	string		Y	
	논문저자	author	박수경, 나호준, 김지혜	string		Y	
	발행년도	publishe_yea	2020년	string		Y	
	한국연구재단 분류	nrf_category	공학	string			학국연구재단 (nrf.re.kr)의 한국학술지인용색인 (KCI)의 연구분야 분류체계 적용
	기타	keyword		string			
본문 및 섹션 요약	파일명	file_name	ART002624634.pdf	string		Y	원문(PDF) 파일명
	아이디	aid	a106981420-a1	string		Y	전체요약, 섹션별 요약의 유형에 따른 id
	초록	abstracts	본 연구는 인삼과 산양삼에 (중략)...전송되는 서비스디자인을 고안했다.(중략) ...이 적은 샘플을 사용해 CNN(VGG16)모델에서 예측 확률 100%를 얻었다	string			논문의 초록 전문
	전체 요약 속성	attr	연구목적				논문 전체 요약문에 대한 속성
	전체 요약문	text	본 연구는 인삼과 산양삼에 (중략)...확립하려했다.				논문 속성이 부여된 요약문 텍스트
	요약문 번호	sentno	1	string		Y	
	섹션별 요약 속성	section_attr	연구목적	string			
	섹션 단락	passage	이와 같은 문제를 해결하기...(중략)---분류 모델의 연구개발을 추진하였다."				
	섹션 요약문	summary	본 연구는 인삼과 산양삼에...(중략)...확립하려했다.				

●○ 데이터 예시

이 데이터는 설명 가능 데이터 기준이며, 표준 데이터셋, 정답 없는 데이터셋은 아래 예시에서 각각 clue, answers가 없는 구조를 가진다.


```
{
    "doc_id": "22450983",
    "meta_date": {
        "title": "감정 트레이닝: 얼굴 표정과 감정 인식 분석을 이용한 이미지 색상 변환",
        "publication": "한국컴퓨터그래픽스학회논문지",
        "publisher": "(사)한국컴퓨터그래픽스학회",
        "nrf_category": "공학",
        "publishe_year": "2018",
        "author": "김종현"
    },
    "total_summary": {
        "file_name": "KCI_FI002381799.pdf",
        "abstracts": "본 논문은 얼굴의 표정 변화를 통해 감정을 분석하는 방법으로 조현병의 초기 증상을 스스로 인지할 수 있는 감정 트레이닝 프레임워크를 제안한다. 먼저, Microsoft의 Emotion API를 이용하여 캡처된 얼굴 표정의 사진으로부터 감정값을 얻고, 피크 분석 기반 표준편차로 시간에 따라 변화하는 얼굴 표정의 미묘한 차이를 인식해 감정상태를 각각 분류한다. 그리하여 Ekman이 제안한 여섯 가지 기본 감정 상태에 반하는 감정들의 정서 및 표현능력이 결핍된 부분에 대해 분석하고, 그 값을 이미지 색상 변환 프레임워크에 통합시켜 사용자 스스로 감정의 변화를 쉽게 인지하고 트레이닝 할 수 있도록 하는 것이 최종 목적이다.",
        "sententce": [
            {
                "sentno": 1,
                "attr": "연구목적",
                "text": "본 논문은 얼굴의 표정 변화를 통해 감정을 분석하는 방법으로 조현병의 초기 증상을 스스로 인지할 수 있는 감정 트레이닝 프레임워크를 제안한다."
            },
            {
                "sentno": 2,
                "attr": "연구방법",
                "text": "먼저, Microsoft의 Emotion API를 이용하여 캡처된 얼굴 표정의 사진으로부터 감정값을 얻고, 피크 분석 기반 표준편차로 시간에 따라 변화하는 얼굴 표정의 미묘한 차이를 인식해 감정상태를 각각 분류한다. 그리하여 Ekman이 제안한 여섯 가지 기본 감정 상태에 반하는 감정들의 정서 및 표현능력이 결핍된 부분에 대해 분석하고, 그 값을 이미지 색상 변환 프레임워크에 통합시켜 사용자 스스로 감정의 변화를 쉽게 인지하고 트레이닝 할 수 있도록 하는 것이 최종 목적이다."
            }
        ]
    },
    "section_summary": [
        {
            "section_attr": "1. 서론",
            "passage": "조현병은 사람의 감정, 사고 및 행동에 영향을 줄 수 있는 만성정신장애이
```

다. 그 외에도 지각의 이상(환각), 사고내용의 이상(망각) 및 사고과정의 이상 외에도 의욕 없음, 사회활동저하, 인지기능 저하 등 많은 정신기능 영역의 손상을 보일 수 있는병이다. 조현병 치료는 약물 혹은 심리치료를 이용하는데, 대부분의 경우 약물치료를 통해 일상생활이 가능하지만 약물을 지속적으로 복용하지 않을 경우 재발 가능성이 커지기 때문에 꾸준한 약 복용과 함께 일상생활에서 관리가 필요하다. 주 치료방법인 약물요법 외에 다양한 치료방법 개발에 대한 필요성과 중요성이 제기되고 있는 만큼 약물치료 후에도 인지기능, 사회기능 등의 장애가 지속될 수 있기 때문에 일상생활에서 조현병 환자의 정신증상과 대인관계 문제를 완화시키는 것을 중요한 치료의 목적으로 보고 있다. 그렇기에 본 연구는 감정 분석을 기반으로 조현병 환자의 정서적 안정 및 치료자와의 긍정적인 관계를 형성하고, 얼굴 표정과 표정의 변화로 감정을 분석하여 스스로의 감정을 쉽게 인지하고 트레이닝 할 수 있는 프레임워크를 제안함으로 음악 및 미술 치료 등으로 활용하여, 위에서 얘기한 긍정적인 치료 효과를 극대화시키고자 한다.",
"summary": "조현병은 사람의 감정, 사고 및 행동에 영향을 줄 수 있는 만성정신장애이다. 조현병의 주 치료방법인 약물요법 외에 다양한 치료방법 개발에 대한 필요성과 중요성이 제기되고 있다. 약물치료 후에도 인지기능, 사회기능 등의 장애가 지속될 수 있기 때문에 일상생활에서 조현병 환자의 정신증상과 대인관계 문제를 완화시키는 것이 중요하다."
}
]
}

●○ 데이터 구축 과정

데이터 구축은 2020 9월부터 12월까지 과학기술정보통신연구원(KISTI), 국립중앙도서관, 한국연구재단 KCI 등재 논문 등을 PDF 다운로드, 데이터 크롤링 등을 통해 수집하고 저작물 이용 정보, 본문의 한글 작성 여부 등을 확인하여 중복이나 유사 본문을 제거하는 필터링을 거쳐 수집대상 35만 건을 기준으로 다양한 학술 분야의 논문을 수집하였다.

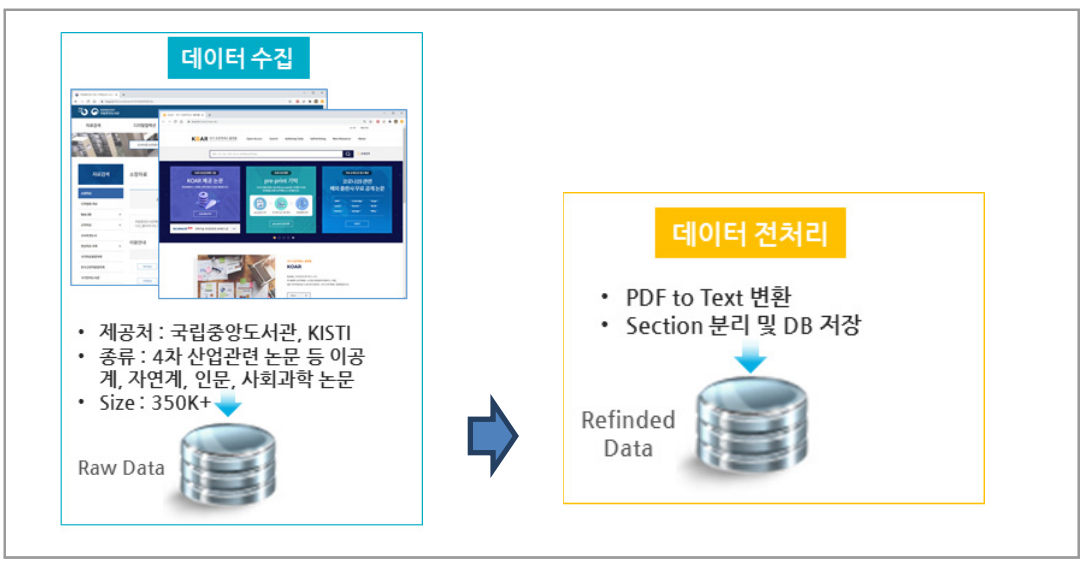

그림3 | 원천 데이터 수집 기준 및 규모

원천 데이터 정제 기준

- 수집한 논문 원천데이터의 메타데이터(타이틀, 저자, 문서종류, 발행처 등)의 필드를 구분하여 본문 텍스트와 초록을 저장하고 데이터 정제를 진행하여 문서 내 불필요한 정보의 삭제 처리하였다.
- 논문 요약문 추출에 부적합한 그림, 서식 등을 삭제 처리하고 PDF에서 추출된 텍스트의 글자 깨짐, 라인 브레이크 등의 각종 오류를 정제 후 데이터 구축을 위한 텍스트를 제작하였다.

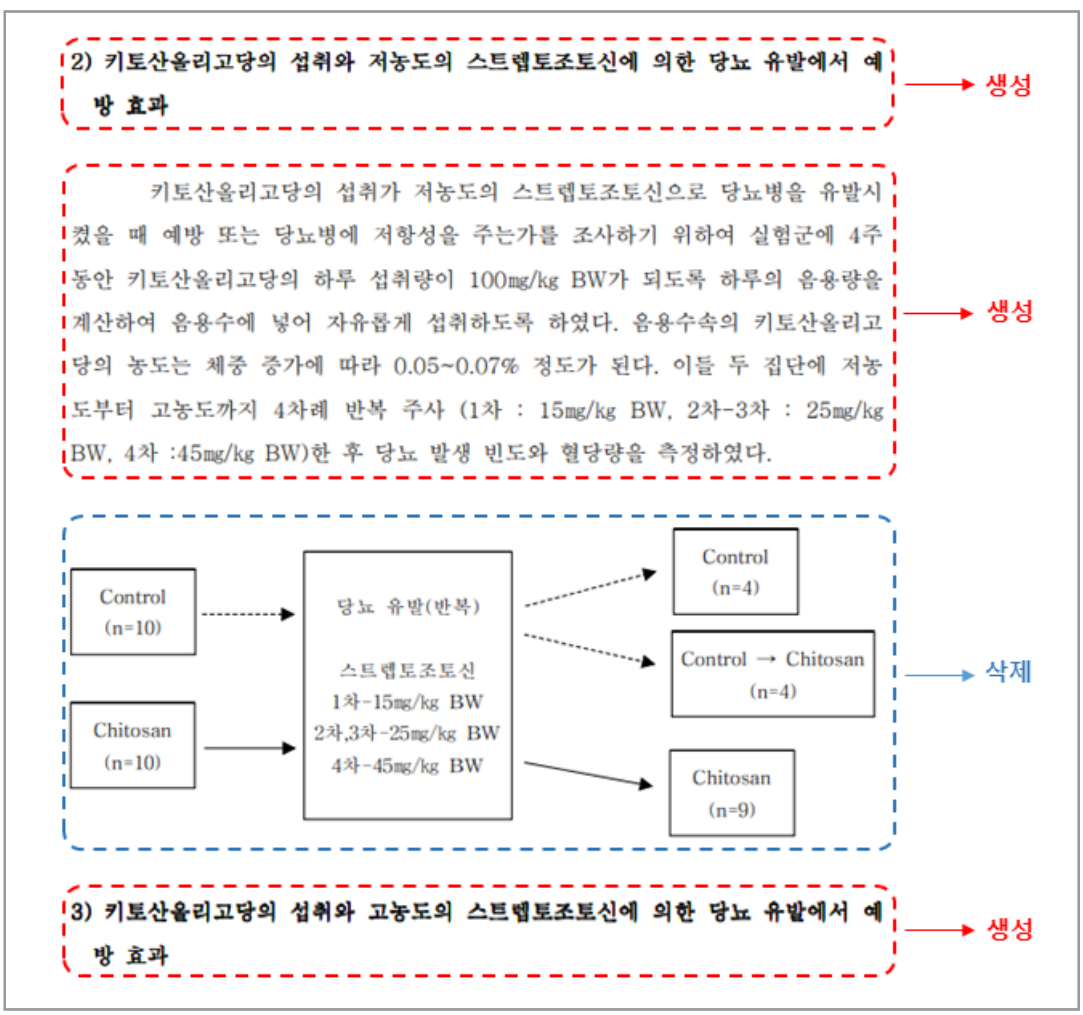

그림4 | 학술 논문의 비텍스트 적 요소 제거 예시

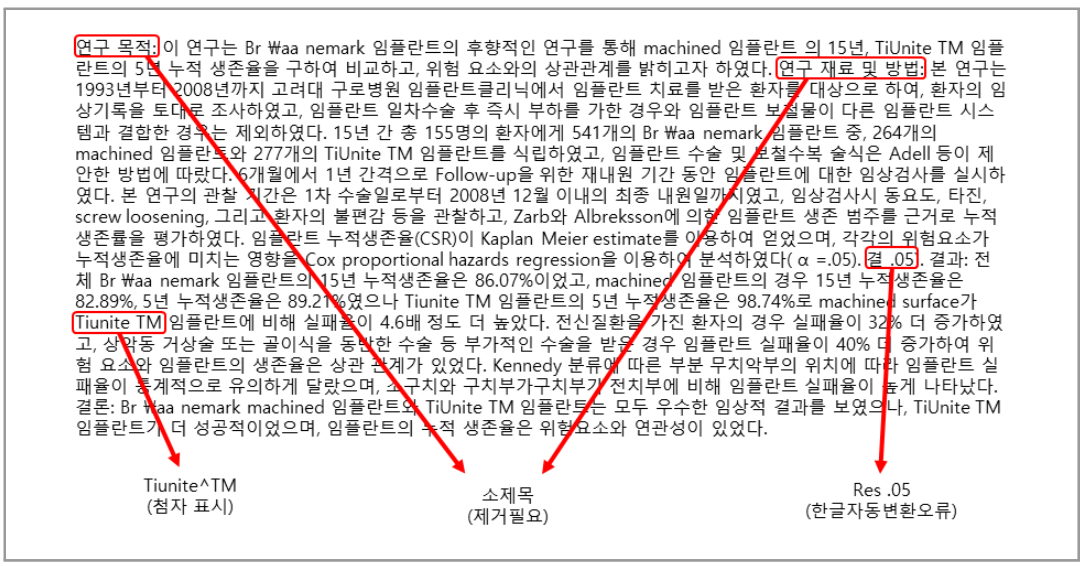

그림5 | 학술 논문의 서식 기호 등 텍스트 정제 예시

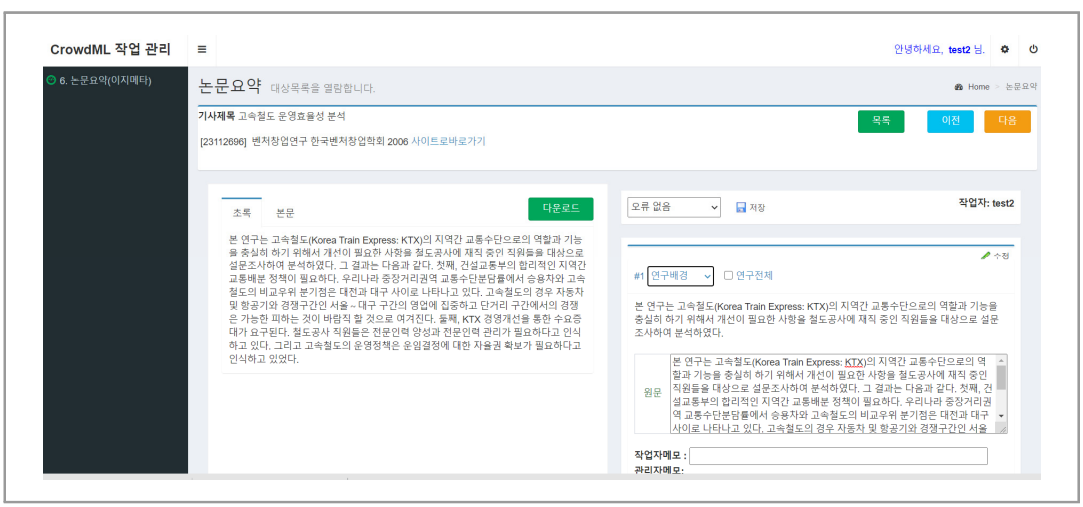

그림6 | 효율적인 데이터 제작을 위한 웹 기반 제작도구

전체 요약문은 초록 또는 본문의 문장 중 태그 규칙에 가장 부합하는 문장을 하나 선택하는 방식으로 가공(추출요약)하고 요약문 마킹(속성 처리)은 연구배경, 연구목적, 연구방법, 연구결과, 연구결론, 연구전체 (최대 6개) 등으로 구성하였다. 따라서 '연구전체'의 문장은 나머지 연구목적, 연구방법, 연구결과, 연구결론 중 하나와 일치할 수 있으며 문장 중 어느 태그도 붙일 수 없는 것들이 존재한다.

각 섹션을 1문장만으로 요약할 수 없는 경우가 발생하면 작업자·검수자의 판단에 의해 가장 적합한 1개의 문장을 선택하며, 부득이한 경우 2개 이상의 문장에 대한 생성 요약 작업을 실시하였다.

●○ 검수와 품질 확보

대량의 데이터를 높은 품질로 생성하기 위하여, 단가는 낮지만 품질 관리가 어려운 크라우드소싱 방식의 데이터생성 작업을 보완하기 위한 검수 프로세스의 정립은 데이터셋 구축에 매우 중요한 의미를 갖는다. 구축 단계에서는 교차 분석, 검수, 전문가 자문, 오류분석, 전문가 자문, 참여기관의 전수 검증을 실시했다. 교차검증 (A작업자 + B작업자 → C검수자)은 동일 자료를 대상으로 복수의 작업자가 분석을 수행하여 그 결과를 통합하고, 2인의 분석팀원이 각지 별도로 작업을 수행한 이후 (동일 팀에 소속된 다른 파트 인원), 양자간에 의견에 차이가 있는 부분은 별도의 상위 수준 검수자가 따로 판단하여 작업 결과를 선택하는 방식을 적용했다.

학술 논문은 학술분야의 다양성과 전문성 때문에 해당 분야의 전공자를 모두 모집하는 것은 불가능하기 때문에 국어학, 언어학, 자연어 처리 전문인력(박사 수료 이상)을 9명 이상 모집하여 품질 검수팀을 구성, 별도 운용하였다. 일치도 분석 및 오류 자동 검출을 위해 자동화 도구(문자수 체크, 허용 문자값 확인 등) 또한 사용하여 기계적, 반복적 오류를 줄일 수 있었다.

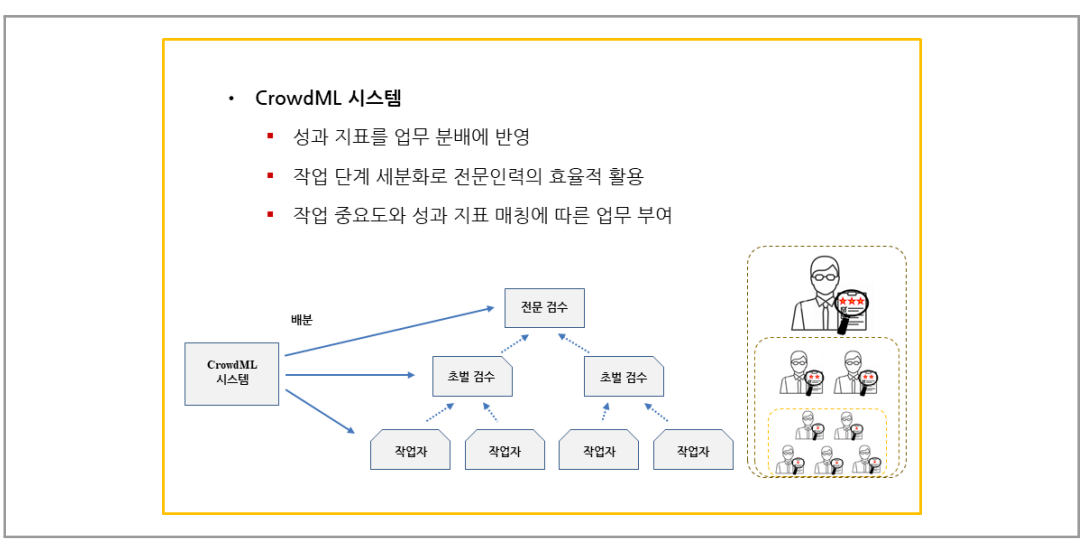

그림7 | 데이터 가공과 검수 기능을 통합(교차 검증 가능)한 효율적 워크벤치 사용

●○ 데이터 구축 담당자

수행기관(주관) : ㈜나라지식정보
(전화: 02-3141-7644, 이메일: nara@narainformation.com)
㈜단아코퍼레이션
(전화: 070-4201-8500, 이메일: dana-corp@naver.com)