인공지능 학습용 데이터 구축·활용 가이드라인

< 의료, 법률 전문 서적 말뭉치 >

인공지능 학습용 데이터 구축·활용 가이드라인	사업 총괄	고려대학교산학협력단
	데이터 획득(수집)	군자출판사(주), ㈜박영사, ㈜범문에듀케이션, 법문사
	데이터 정제	고려대학교산학협력단
	데이터 가공	고려대학교산학협력단
	데이터 검사	㈜나라지식정보
	크라우드 소싱	고려대학교산학협력단
	저작도구 개발	고려대학교산학협력단
	AI모델 개발	한양대학교산학협력단
데이터 구축·활용 가이드라인 작성		고려대학교산학협력단
데이터 구축·활용 가이드라인 버전	Ver 1.0 (2023.06.23)	

제·개정 이력

개정번호	개정일자	개정내용
V.1.0	2023.06.23	최초 작성

목 차

제1장 데이터 명세 정보

1. 데이터 정보 요약

데이터 명	의료, 법률 전문 서적 말뭉치	
활용 분야	의료, 법률, 말뭉치 데이터, 자연어처리, 인공지능	
데이터 요약	저작재산권 이용허락을 확보한 의료 및 법률 분야 핵심 전문 도서와 분야 전문성과 최신성을 갖춘 특허(의료), 판결문(법률)을 기반으로 1억 어절의 데이터셋 구축	
데이터 출처	의료, 법률 전문 서적	
데이터 통계	데이터 구축 규모	총 102,300,332어절
	데이터 분포 (충분성, 균등성, 편향성 여부 확인)	* 다양성(통계) - 문장 어절 수 분포 : 10어절 미만(0.00%), 10 ~ 20어절(0.01%), 21 ~ 30어절(0.02%), 31 ~ 40어절(0.03%), 41 ~ 50어절(0.05%), 51어절 이상(99.90%) - 도서 출판 기간 분포 : 2010년 미만(6.09%), 2010년 ~ 2011년(4.76%), 2012년 ~ 2013년(7.16%), 2014년 ~ 2015년(10.44%), 2016년 ~ 2017년(11.90%), 2018년 ~ 2019년(18.68%), 2020년 ~ (40.97%) * 다양성(요건) - 문서 표준 분류 분포(분야 구성비) : 의료(48.11%), 법률(51.89%)
데이터 이력	배포 버전	v1.0
	개정 이력	최종 어절 구축 적용
	작성자/배포자	송유연

2. 데이터 포맷

○ 원시데이터 포맷은 출판사별로 보유한 다양한 형태의 텍스트 파일이며, 이를 마스터 테이블 양식으로 통합 관리함

○ 도서명, 저자명, 출판일 등 도서 정보를 원시데이터 마스터 테이블 DB에 보관하고 이를 기반으로 데이터의 다양성 통계 자료에 활용함

○ 전 단계에 걸쳐 데이터 형태는 텍스트 기반의 디지털 파일이며, 구축 단계별로 포맷은 상이함

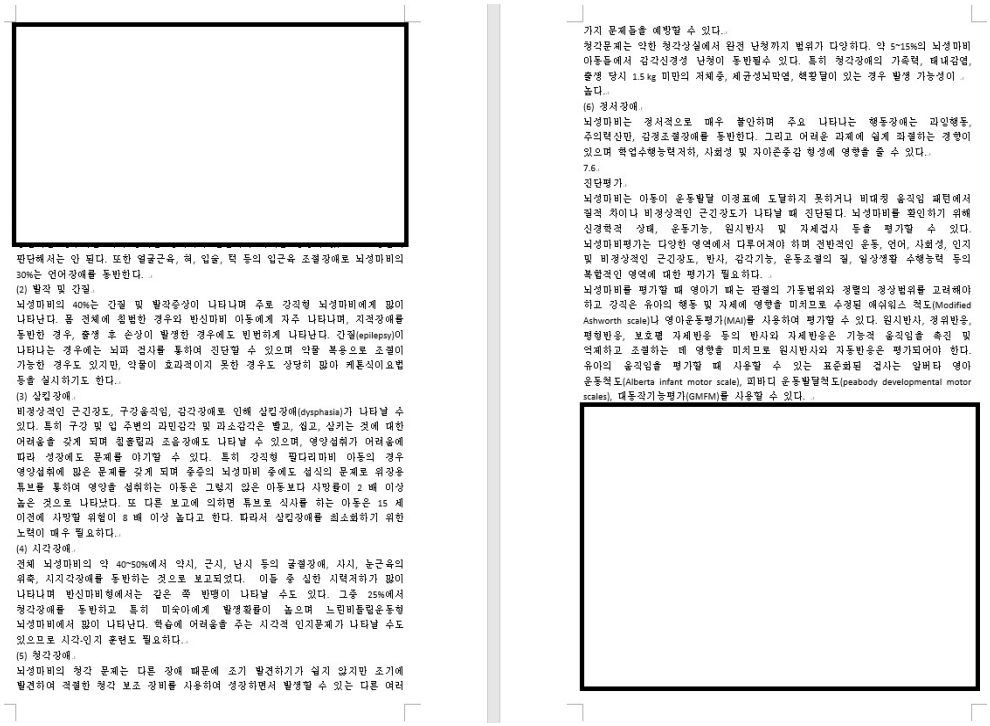
□ 원시데이터 마스터 테이블 정보

No.	속성명	필수여부	속성 및 내용
1	BOOK ID	필수	도서 ID
2	Book.name	필수	도서명
3	Authors	필수	저자명
4	Translators	선택	번역자명
5	Year	필수	도서 출판일
6	Page	필수	총 페이지수
7	Publisher	필수	출판사명
8	Category	필수	도서 카테고리

□ 데이터 구축 단계별 데이터 정보

구분	획득(수집) 단계	정제 단계	가공(라벨링) 단계
데이터 구분	원시데이터	원천데이터	최종데이터
데이터 형태	디지털 파일	디지털 파일	디지털 파일
데이터 포맷	pdf, hwp, MS-Word	txt	JSON

□ 원시데이터 예시 – MS-Word



판단해서는 안 된다. 또한 얼굴근육, 혀, 입술, 턱 등의 입근육 조절장애로 뇌성마비의 30%는 언어장애를 동반한다. ..
(2) 발작 및 간질..
뇌성마비의 40%는 간질 및 발작증상이 나타나며 주로 강직형 뇌성마비에게 많이 나타난다. 몸 전체에 침범한 경우와 반신마비 아동에게 자주 나타나며, 지적장애를 동반한 경우, 출생 후 손상이 발생한 경우에도 빈번하게 나타난다. 간질(epilepsy)이 나타나는 경우에는 뇌파 검사를 통하여 진단할 수 있으며 약물 복용으로 조절이 가능한 경우도 있지만, 약물이 효과적이지 못한 경우도 상당히 많아 케톤식이요법 등을 실시하기도 한다..
(3) 삼킴장애..
비정상적인 근긴장도, 구강움직임, 감각장애로 인해 삼킴장애(dysphasia)가 나타날 수 있다. 특히 구강 및 입 주변의 과민감각 및 과소감각은 빨고, 씹고, 삼키는 것에 대한 어려움을 갖게 되며 침흘림과 조음장애도 나타날 수 있으며, 영양섭취가 어려움에 따라 성장에도 문제를 야기할 수 있다. 특히 강직형 활다리마비 아동의 경우 영양섭취에 많은 문제를 갖게 되며 중증의 뇌성마비 중에도 섭식의 문제로 위장용 튜브를 통하여 영양을 섭취하는 아동은 그렇지 않은 아동보다 사망률이 2 배 이상 높은 것으로 나타났다. 또 다른 보고에 의하면 튜브로 식사를 하는 아동은 15 세 이전에 사망할 위험이 8 배 이상 높다고 한다. 따라서 삼킴장애를 최소화하기 위한 노력이 매우 필요하다..
(4) 시각장애..
전체 뇌성마비의 약 40~50%에서 약시, 근시, 난시 등의 굴절장애, 사시, 눈근육의 위축, 시지각장애를 동반하는 것으로 보고되었다. 이들 중 심한 시력저하가 많이 나타나며 반신마비형에서는 같은 쪽 반맹이 나타날 수도 있다. 그중 25%에서 청각장애를 동반하고 특히 미숙아에게 발생확률이 높으며 느린비틀림운동형 뇌성마비에서 많이 나타난다. 학습에 어려움을 주는 시각적 인지문제가 나타날 수도 있으므로 시각-인지 훈련도 필요하다..
(5) 청각장애..
뇌성마비의 청각 문제는 다른 장애 때문에 조기 발견하기가 쉽지 않지만 조기에 발견하여 적절한 청각 보조 장비를 사용하여 성장하면서 발생할 수 있는 다른 여러 가지 문제들을 예방할 수 있다..
청각문제는 약한 청각상실에서 완전 난청까지 범위가 다양하다. 약 5~15%의 뇌성마비 아동들에서 감각신경성 난청이 동반될수 있다. 특히 청각장애의 가족력, 태내감염, 출생 당시 1.5 kg 미만의 저체중, 세균성뇌막염, 핵황달이 있는 경우 발생 가능성이 높다..
(6) 정서장애..
뇌성마비는 정서적으로 매우 불안하며 주요 나타나는 행동장애는 과잉행동, 주의력산만, 감정조절장애를 동반한다. 그리고 어려운 과제에 쉽게 좌절하는 경향이 있으며 학업수행능력저하, 사회성 및 자아존중감 형성에 영향을 줄 수 있다..
7.6..
진단평가..
뇌성마비는 아동이 운동발달 이정표에 도달하지 못하거나 비대칭 움직임 패턴에서 질적 차이나 비정상적인 근긴장도가 나타날 때 진단된다. 뇌성마비를 확인하기 위해 신경학적 상태, 운동기능, 원시반사 및 자세검사 등을 평가할 수 있다. 뇌성마비평가는 다양한 영역에서 다루어져야 하며 전반적인 운동, 언어, 사회성, 인지 및 비정상적인 근긴장도, 반사, 감각기능, 운동조절의 질, 일상생활 수행능력 등의 복합적인 영역에 대한 평가가 필요하다. ..
뇌성마비를 평가할 때 영아기 때는 관절의 가동범위와 정렬의 정상범위를 고려해야 하고 강직은 유아의 행동 및 자세에 영향을 미치므로 수정된 애쉬워스 척도(Modified Ashworth scale)나 영아운동평가(MAI)를 사용하여 평가할 수 있다. 원시반사, 정위반응, 평형반응, 보호뻗 자세반응 등의 반사와 자세반응은 기능적 움직임을 촉진 및 억제하고 조절하는 데 영향을 미치므로 원시반사와 자동반응은 평가되어야 한다. 유아의 움직임을 평가할 때 사용할 수 있는 표준화된 검사는 알버타 영아 운동척도(Alberta infant motor scale), 피바디 운동발달척도(peabody developmental motor scales), 대동작기능평가(GMFM)를 사용할 수 있다. ..

□ 원천데이터 예시 – txt

[[BODY 8]]

또한 얼굴근육, 혀, 입술, 턱 등의 입근육 조절장애로 뇌성마비의 30%는 언어장애를 동반한다.
(2) 발작 및 간질
뇌성마비의 40%는 간질 및 발작증상이 나타나며 주로 강직형 뇌성마비에게 많이 나타난다. 몸 전체에 침범한 경우와 반신마비 아동에게 자주 나타나며, 지적장애를 동반한 경우, 출생 후 손상이 발생한 경우에도 빈번하게 나타난다. 간질(epilepsy)이 나타나는 경우에는 뇌파 검사를 통하여 진단할 수 있으며 약물 복용으로 조절이 가능한 경우도 있지만, 약물이 효과적이지 못한 경우도 상당히 많아 케톤식이요법 등을 실시하기도 한다.
(3) 삼킴장애
비정상적인 근긴장도, 구강움직임, 감각장애로 인해 삼킴장애(dysphasia)가 나타날 수 있다. 특히 구강 및 입 주변의 과민감각 및 과소감각은 빨고, 씹고, 삼키는 것에 대한 어려움을 갖게 되며 침흘림과 조음장애도 나타날 수 있으며, 영양섭취가 어려움에 따라 성장에도 문제를 야기할 수 있다. 특히 강직형 활다리마비 아동의 경우 영양섭취에 많은 문제를 갖게 되며 중증의 뇌성마비 중에도 섭식의 문제로 위장용 튜브를 통하여 영양을 섭취하는 아동은 그렇지 않은 아동보다 사망률이 2배 이상 높은 것으로 나타났다. 또 다른 보고에 의하면 튜브로 식사를 하는 아동은 15세 이전에 사망할 위험이 8배 이상 높다고 한다. 따라서 삼킴장애를 최소화하기 위한 노력이 매우 필요하다.
(4) 시각장애
전체 뇌성마비의 약 40~50%에서 약시, 근시, 난시 등의 굴절장애, 사시, 눈근육의 위축, 시지각장애를 동반하는 것으로 보고되었다. 이들 중 심한 시력저하가 많이 나타나며 반신마비형에서는 같은 쪽 반맹이 나타날 수도 있다. 그중 25%에서 청각장애를 동반하고 특히 미숙아에게 발생확률이 높으며 느린비틀림운동형 뇌성마비에서 많이 나타난다. 학습에 어려움을 주는 시각적 인지문제가 나타날 수도 있으므로 시각-인지 훈련도 필요하다.
(5) 청각장애
뇌성마비의 청각 문제는 다른 장애 때문에 조기 발견하기가 쉽지 않지만 조기에 발견하여 적절한 청각 보조 장비를 사용하여 성장하면서 발생할 수 있는 다른 여러 가지 문제들을 예방할 수 있다.
청각문제는 약한 청각상실에서 완전 난청)까지 범위가 다양하다. 약 5~15%의 뇌성마비 아동들에서 감각신경성 난청이 동반될수 있다. 특히 청각장애의 가족력, 태내감염, 출생 당시 1.5 kg 미만의 저체중, 세균성뇌막염, 핵황달이 있는 경우 발생 가능성이
높다.
(6) 정서장애
뇌성마비는 정서적으로 매우 불안하며 주요 나타나는 행동장애는 과잉행동, 주의력산만, 감정조절장애를 동반한다. 그리고 어려운 과제에 쉽게 좌절하는 경향이 있으며 학업수행능력저하, 사회성 및 자아존중감 형성에 영향을 줄 수 있다.
7.6
진단평가
뇌성마비는 아동이 운동발달 이정표에 도달하지 못하거나 비대칭 움직임 패턴에서 질적 차이나 비정상적인 근긴장도가 나타날 때 진단된다. 뇌성마비를 확인하기 위해 신경학적 상태, 운동기능, 원시반사 및 자세검사 등을 평가할 수 있다. 뇌성마비평가는 다양한 영역에서 다루어져야 하며 전반적인 운동, 언어, 사회성, 인지 및 비정상적인 근긴장도, 반사, 감각기능, 운동조절의 질, 일상생활 수행능력 등의 복합적인 영역에 대한 평가가 필요하다.
뇌성마비를 평가할 때 영아기 때는 관절의 가동범위와 정렬의 정상범위를 고려해야 하고 강직은 유아의 행동 및 자세에 영향을 미치므로 수정된 애쉬워스 척도(Modified Ashworth scale)나 영아운동평가(MAI)를 사용하여 평가할 수 있다. 원시반사, 정위반응, 평형반응, 보호뻗 자세반응 등의 반사와 자세반응은 기능적 움직임을 촉진 및 억제하고 조절하는 데 영향을 미치므로 원시반사와 자동반응은 평가되어야 한다. 유아의 움직임을 평가할 때 사용할 수 있는 표준화된 검사는 알버타 영아 운동척도(Alberta infant motor scale), 피바디 운동발달척도(peabody developmental motor scales), 대동작기능평가(GMFM)를 사용할 수 있다.

□ 최종데이터 예시 – JSON


```
{
    "book_id": "MTB013183",
    "category": "신경과학",
    "popularity": 2,
    "keyword": [
        "시각장애",
        "정서장애",
        "삼킴장애",
        "청각장애",
```

```
            "뇌성마비"
        ],
        "text": " 또한 얼굴근육, 혀, 입술, 턱 등의 입근육 조절장애로 뇌성마비의 30%는 언어장애를 동반한다. (2) 발작 및 간질 뇌성마비의 40%는 간질 및 발작증상이 나타나며 주로 강직형 뇌성마비에게 많이 나타난다. 몸 전체에 침범한 경우와 반신마비 아동에게 자주 나타나며, 지적장애를 동반한 경우, 출생 후 손상이 발생한 경우에도 빈번하게 나타난다. 간질(epilepsy)이 나타나는 경우에는 뇌파 검사를 통하여 진단할 수 있으며 약물 복용으로 조절이 가능한 경우도 있지만, 약물이 효과적이지 못한 경우도 상당히 많아 케톤식이요법 등을 실시하기도 한다. (3) 삼킴장애 비정상적인 근긴장도, 구강움직임, 감각장애로 인해 삼킴장애(dysphasia)가 나타날 수 있다. 특히 구강 및 입 주변의 과민감각 및 과소감각은 빨고, 씹고, 삼키는 것에 대한 어려움을 갖게 되며 침흘림과 조음장애도 나타날 수 있으며, 영양섭취가 어려움에 따라 성장에도 문제를 야기할 수 있다. 특히 강직형 팔다리마비 아동의 경우 영양섭취에 많은 문제를 갖게 되며 중증의 뇌성마비 중에도 섭식의 문제로 위장용 튜브를 통하여 영양을 섭취하는 아동은 그렇지 않은 아동보다 사망률이 2배 이상 높은 것으로 나타났다. {중략}대동작기능평가(GMFM)를 사용할 수 있다.",
        "word_segment": 451,
        "publication_ymd": "20180901",
        "NE": [
            {
                "id": 1,
                "entity": "얼굴",
                "type": "AM_PART",
                "begin": 4,
                "end": 5
            },
            {
                "id": 2,
                "entity": "혀",
                "type": "AM_PART",
                "begin": 10,
                "end": 10
            },
            {
                "id": 3,
                "entity": "입술",
                "type": "AM_PART",
                "begin": 13,
                "end": 14
            },...
            {
                "id": 6,
                "entity": "근육 조절장애",
                "type": "TMM_DISEASE",
                "begin": 23,
                "end": 29
            },
            {
                "id": 7,
                "entity": "뇌성마비",
                "type": "TMM_DISEASE",
                "begin": 32,
                "end": 35
            },...(중략)
```

3. 어노테이션 포맷

No.			속성명	필수여부	Type	항목 설명	예시
1			totalcount	Y	number	카운트	
2			data	Y	object	데이터셋	
	2-1		data[].book_id	Y	string	도서식별 ID	MTB000001 LTB000001
	2-2		data[].category	Y	string	카테고리	외과, 헌법등 도서의 카테고리
	2-3		data[].popularity	Y	number	라벨링 난이도	1(하), 2(중), 3(상)
	2-4		data[].keyword	N	array	키워드	[“포도당”, “포도당신합성”, “단백질 이화작용”]
	2-5		data[].text	Y	string	텍스트 본문	한국의 동물 보호법은 1991년 단 12개의..
	2-6		data[].word_segment	Y	number	텍스트 어절 수	
	2-7		data[].publication_ymd	Y	string	도서 출판일 특허 출원일 판결문 선고일	예시) 20201020
	2-8		data[].NE	N	object	라벨링 오브젝트	
		2-8-1	data[].NE[].id	N	number	라벨링 순번	
		2-8-2	data[].NE[].entity	N	string	라벨링 대상 텍스트	동물 보호법
		2-8-3	data[].NE[].type	N	string	카테고리 타입	CV_LAW,AM_PART..
		2-8-4	data[].NE[].begin	N	number	Target 시작 위치	
		2-8-5	data[].NE[].end	N	number	Target 끝 위치	

4. 데이터 구성

○ 데이터는 의료와 법률로 분야별로 구분되고, 원천데이터는 txt 파일, 라벨링 데이터는 Json 파일 형태로 제공

○ 원천데이터 텍스트 파일은 도서 및 특허, 판결문에 식별 ID를 부여한 형태임

No.	속성명	항목 설명
1	분야	의료/법률
2	파일명	도서ID(의료 : MTB / 법률 : LTB)+식별ID 특허ID(MPA)+식별ID 판결문ID(LJU)+식별ID

○ 라벨링데이터는 의료분야 json 파일, 법률분야 json 파일로 구성

No.	속성명	항목 설명
1	분야	의료/법률
2	파일명	의료 : medical_corpus.json 법률 : legal_corpus.json

5. 데이터 통계

5.1 데이터 통계

5.1.1 데이터 구축 규모

○ 원천데이터 : 저작재산권 이용허락을 확보한 의료 및 법률 전문 도서와 특허(의료), 판결문(법률)을 기반으로 126,512건 원천데이터 구축

○ 라벨링데이터 : 확보된 원천데이터로 최총 102,300,332어절 구축

과제번호	과제명	구분	원천데이터	라벨링데이터
2-032	의료, 법률 전문 서적 말뭉치 데이터	의료	46,883건	49,217,568어절
		법률	79,630건	53,082,764어절
		합	126,513건	102,300,332어절

5.1.2 데이터 분포

○ 문장 어절 수 분포

어절 구간	원천데이터 수	비율
10 어절 미만	1	0.00%
10 어절 ~ 20 어절	10	0.01%
21 어절 ~ 30 어절	27	0.02%
31 어절 ~ 40 어절	33	0.03%
41 어절 ~ 50 어절	61	0.05%
51 어절 ~	126,381	99.90%
합계	126,513	100%

○ 도서 출판 기간 분포

출판 기간	원천데이터 수	비율
2010년 미만	7,710	6.09%
2010년 ~ 2011년	6,017	4.76%
2012년 ~ 2013년	9,059	7.16%
2014년 ~ 2015년	13,202	10.44%
2016년 ~ 2017년	15,050	11.90%
2018년 ~ 2019년	23,637	18.68%
2020년 ~	51,838	40.97%
합계	126,513	100%

○ 전문 용어 분포

대분류	중분류	전문 용어	건수	비율
의료	STUDY_FIELD	FD_MEDICINE	44,230	0.65%
	THEORY	TR_MED_PROCEDURE	398,906	5.86%
		TR_MED_MEASURE	169,416	2.49%
		TR_MED_OBSERVATION	12,193	0.18%
	ARTIFACTS	AFW_MED_DEVICE	180,709	2.66%
	ORGANIZATION	OGG_MEDICINE	11,397	0.17%
	ANIMAL	AM_PART	268,171	3.94%
	TERM	TMM_DISEASE	685,799	10.08%
		TMM_SYMPTOM	305,514	4.49%
		TMM_DRUG	238,190	3.50%
		TM_CELL_TISSUE_ORGAN	1,018,956	14.98%
법률	ORGANIZATION	OGG_LAW	374,884	5.51%
	CIVILIZATION	CV_LAW	1,352,497	19.88%
		CV_TAX	32,807	0.48%
	EVENT	EV_OTHERS	12,405	0.18%
	TERM	TML_PARTY	470,796	6.92%
		TML_PROCEDURAL_ACTS	208,849	3.07%
		TML_LITIGATION	165,950	2.44%
		TML_PROVISION	508,108	7.47%
		TML_CASE	187,276	2.75%
		TML_JUDGMENT	155,177	2.28%
합계			6,802,230	100%

○ 분류(독해 난이도) 분포

난이도 수준	원천데이터 수	비율
하(일반인이 쉽게 이해)	18,911	14.95%
중(일반인이 어느 정도 이해)	55,643	43.98%
상(일반인이 이해하기 어려움)	51,959	41.07%
합계	126,513	100%

○ 문서 표준 분류 분포 – 분야 구성비

분야	원천데이터 수	어절 수	비율(어절 수 기준)
의료	46,883	49,217,568	48.11%
법률	79,630	53,082,764	51.89%
합계	126,513	102,300,332	100%

ㅇ 문서 표준 분류 분포 – 분야별 카테고리 구성비

분야	카테고리	원천데이터 수	비율
의료	가정의학	431	0.34%
	간호학	7	0.01%
	내과학	6,963	5.50%
	마취과학	434	0.34%
	면역학/병리학/임상병리학	2,461	1.95%
	미생물학/기생충학	1,370	1.08%
	방사선과학	3,864	3.05%
	비뇨기과학	1,110	0.88%
	산부인과학	1,630	1.29%
	생리학	391	0.31%
	생화학	679	0.54%
	성형외과학	640	0.51%
	소아과학	936	0.74%
	신경과학	763	0.60%
	신경외과학	1,658	1.31%
	안과학/임상안광학	821	0.65%
	약학/약리학	3,817	3.02%
	예방의학/직업환경의학	1,453	1.15%
	응급의학	2,299	1.82%
	이비인후과학	2,100	1.66%
	일반외과학	1,573	1.24%
	재활의학/물리치료학/작업치료학	5,767	4.56%
	정신과학	496	0.39%
	정형외과학	1,873	1.48%
	치의학	2,003	1.58%
	피부과학	298	0.24%
	해부학	620	0.49%
	흉부외과학	426	0.34%
법률	경제법	1,547	1.08%
	교육법	150	0.12%
	국제법(공법)	305	0.24%
	국제법(사법)	1,252	0.99%
	노동법등	704	0.56%
	민법일반	12,941	10.23%
	민사소송법	4,548	3.59%
	법학일반	4,208	3.33%
	상사법등	11,102	8.78%
	소비자/보호법	555	0.44%
	의료/보건법	1,231	0.97%
	인권/국제인권법	10	0.01%
	조세/세법	1,006	0.80%
	중재법	6	0.00%
	지방자치법	395	0.31%

	지적재산권법	3,223	2.55%
	친족/상속법	695	0.55%
	토지/부동산/주택/농지법	1,328	1.05%
	해사/해상/해상운송/해상보험법	690	0.55%
	행정법	9,473	7.49%
	헌법	12,196	9.64%
	형법/형사소송법등	11,676	9.23%
	환경/교통법	389	0.31%
합계		126,513	100%

6. 원시데이터 특성

6.1.1 대상분류

No.	분야	항목 설명
1	의료	군자출판사, 범문에듀케이션에서 출판된 도서와 의료 분야 특허 중 수집 선정 기준을 통과한 원시데이터
2	법률	박영사, 법문사에서 출판된 도서와 판결문 중 수집 선정 기준을 통과한 원시데이터

6.1.2 제약조건

No.	분야	수집 조건	내용
1	의료	일부 제약있음(semi-constrained)	- 의료 분야 전문가들이 MOS 정성평가를 수행하여 핵심도서 선정 - 출판사별 저작권 확보 완료
2	법률	일부 제약있음(semi-constrained)	- 법률 분야 전문가들이 MOS 정성평가를 수행하여 핵심도서 선정 - 출판사별 저작권 확보 완료

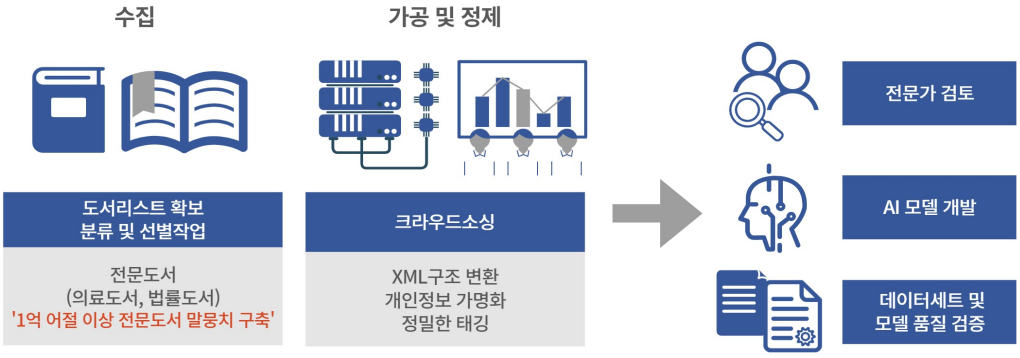
6.1.3 속성

ㅇ pdf, hwp, MS-Word로 된 디지털 형태의 원시데이터는 도서별로 정제되어 txt 기반의 원천데이터로 변환

제2장 데이터 구축

1. 데이터 구축 개요

○ 의료 및 법률 전문도서 말뭉치 구축 사업은 4차 산업혁명에 대비한 우리말 빅데이터(말뭉치) 구축하는 사업의 일환으로, 향후 4차 산업혁명 시대에 의료 및 법률과 같은 고도의 전문성을 필요로 하는 분야의 인공지능 서비스 개발 및 기술 혁신을 위한 중요 자료를 수집하여 활용하기 위한 기반 사업으로 의미가 매우 큰 사업임. 이 사업을 통해 구축된 말뭉치를 의료 및 법률 분야 인공지능 개발 산업과 연구 등에 원활하게 활용될 수 있도록 하는 데 그 목적이 있음.



구축 단계	수행 주체	수행 업무	내용
데이터 수집	참여기관 출판사	원시데이터 확보	- 전문가가 검토한 핵심 전문도서의 디지털 파일 수집 - 저작자와 저작재산권 이용허락 계약을 통한 데이터의 공개·배포·활용 등 법적 권리 확보
데이터 정제	고려대학교 산학협력단	원시데이터 정제	- 저작권 확보된 도서의 최신 원시 파일을 텍스트 형태로 추출 - 크라우드워커가 크라우드워커 플랫폼을 활용하여 텍스트 추출된 데이터 정제 작업 진행
데이터 가공	고려대학교 산학협력단	원천데이터 라벨링	- 크라우드워커가 크라우드워커 플랫폼을 활용하여 라벨링 가이드라인 따라 의료, 법률 분야 카테고리에 해당하는 용어에 라벨링 수행
데이터 검수	나라지식정보	데이터 품질 점검	- 크라우드워커가 크라우드워커 플랫폼을 활용하여 품질 관리 기준 체크리스트에 따라 검수 수행
AI모델 개발	한양대학교 산학협력단	학습모델 구축 및 점검	- 의료, 법률 분야 학습모델 선정 및 모델 구축 - AI 학습모델 유효성 품질 검수

2. 임무 정의

2.1 임무 정의

○ 말뭉치를 이용한 자연어처리 기술은 컴퓨터와 인간 언어 사이의 상호 작용하는 기술로 인공지능의 핵심 기능 중 하나임. 자연어처리 기술은 기계번역, 대화체 질의응답 시스템, 대화시스템, 정보검색 분야 뿐만 아니라 인간의 언어정보처리 원리와 이해를 위한 언어학과 뇌인지 언어정보처리 분야까지 핵심적인 요소로 작용됨.

○ 본 과제에서 구축한 대용량 의료/법률 말뭉치는 전문서적 및 특허, 판결문을 기반으로 확보됨에 따라 의료/법률 분야의 자연어처리 기술에 높은 활용성을 갖음. 특히. 총 1억 어절 이상의 의료/법률 분야의 말뭉치를 확보함으로써 대용량의 데이터를 확보함과 동시에 정밀한 구조로 정제된 말뭉치 체계를 구축함으로써 향후 다양한 분야의 한국어 말뭉치 구축의 표본으로 활용 가능함. 또한, 의료/법률 분야의 특수성으로 인해 말뭉치 데이터 세트의 구축 역량이 부족한 중소·벤처 기업의 수요를 충족시켜 줌으로써 선진국과의 기술력 차를 좁힐 수 있는 인공지능 개발 환경을 조성할 수 있음.

2.2 데이터 구축 유의사항

○ 본 과제에서는 도서 내용을 데이터셋으로 제공하는 만큼 저작권이 중요한데, 데이터 구축 시 저작자와의 저작재산권 이용허락 계약을 확보하였으므로 법적인 문제 발생 소지는 적음
그럼에도, 도서 원문, 도서 목록 등 도서에 대한 정보가 유출되지 않도록 주의하고 해당 정보를 직접적으로 다루는 인력에 대해서는 비밀유지계약서로 동의를 얻어 정보 유출을 방지

3. 획득(수집)

3.1 원시데이터 선정

○ 의료 및 법률 분야 전문 도서와 의료 분야는 특허, 법률 분야는 판결문으로 원시데이터를 수집하고 구축할 의료 및 법률 분야 말뭉치의 품질을 높이기 위해 원시데이터 선정에 있어서 다음과 같은 점을 고려함.

다양성/균일성 /공평성	분야별 다양한 전문 도서가 포함되어 다양한 자연어처리 테스크(task)에 활용할 수 있도록 도서를 연구재단에서 제시하는 국가표준 연구분류 기준의 하위 분류 10가지 이상이 골고루 포함되도록 조정
신뢰성/사실성	현대의 최신 정보를 학습할 수 있는 말뭉치로서 도서의 경우 2010년이후에 출판된 도서 중에서 선정하고, 선정된 도서의 최종 출판된 마지막 버전의 원고를 수집
충분성	각 도서의 내용이 충분히 반영될 수 있도록 최소한 100페이지 이상의 도서를 선정함으로써 인공지능 학습에 필요한 최소한의 말뭉치 어절수를 보장

3.2 획득(수집) 절차

1. 출판사 도서 목록 작성 (도서제목, 출간연도, 저자목록, 페이지, 정가 정보 포함)
2. 전문도서 분류 (중분류, 소분류 및 세분류명)
3. 핵심 전문도서 1차 선정 (분야별 전문가 검토, 분야별 200종 이상 선정)
4. 핵심 전문도서 최종 선정 (사업총괄기관 내부검토, 제외 도서 및 분야별 비율 조정)
5. 저작권 협의 및 검토
6. 원본자료(원시데이터) 확보 (데이터 정제팀에 제공)

3.3 획득(수집) 기준

○ 핵심 전문도서 선정에 있어서 먼저 정리된 도서 목록을 기준으로 분야별 전문가가 분야별로 200종의 핵심 전문도서를 선정하고, 핵심도서의 페이지 수를 통해 예상한 도서별 어절 수를 근거로 한 분야가 전체 15%가 넘지 않도록 조정함.

[표] 의료 및 법률분야 분류안

번호	의학분야	법률분야
1	해부학	법학일반
2	생리학	헌법
3	생화학	행정법
4	면역학/병리학/임상병리학	형법/형사소송법등
5	약학/약리학	민사소송법
6	미생물학/기생충학	국제법(공법)
7	예방의학/직업환경의학	민법일반
8	내과학	친족/상속법
9	일반외과학	상사법등

10	소아과학	해사/해상/해상운송/해상보험법
11	산부인과학	국제법(사법)
12	정신과학	경제법
13	정형외과학	조세/세법
14	신경외과학	노동법등
15	흉부외과학	교육법
16	성형외과학	환경/교통법
17	안과학/임상안광학	의료/보건법
18	이비인후과학	중재법
19	피부과학	지적재산권법
20	비뇨기과학	토지/부동산/주택/농지법
21	방사선과학	지방자치법
22	마취과학	소비자/보호법
23	재활의학/물리치료학/작업치료학	통일관계법
24	신경과학	인권/국제인권법
25	가정의학	
26	응급의학	
27	치의학	
28	간호학	

3.4 획득(수집) 조직

- 본 과제의 수집 대상은 출판물로, 우리나라 의료 및 법률의 대표적인 출판사인 군자출판사(주), ㈜범문에듀케이션, ㈜박영사, 법문사를 중심으로 말뭉치 구축 및 배포를 위한 저작물 활용 동의 의사를 타진하고 참여 의사가 있는 저작권자의 저작물 목록을 확보하여 진행

4. 정제

4.1 원천데이터 규모

구분	데이터 규모
의료	46,883건
법률	79,630건
합	126,513건

4.2 정제 절차

4.2.1 원시데이터 구조 확인

- ㅇ 도서별로 상이한 파일 구조를 확인하고 저작물의 특성을 파악함.

- ㅇ 전자파일의 경우, doc, docx, hwp, hwpx, txt 등 다양한 파일형태로 이루어져 있으며, 그림이나 사진이 많은 도서의 경우, pdf로 되어 있는 경우도 있음.

4.2.2 원문 추출

- ㅇ 챕터별로 표지를 달아 TXT 파일로 변환함. 이 과정에서 원문과 다르게 깨지는 부분을 수정함. 전문 도서의 경우 내부 구조가 저작자에 따라 다양한 형태로 되어 있으므로 일괄적인 원문 추출은 불가능함. 원문의 내용에는 본문을 제외하고 표지, 목차, 머리말, 부록, 색인, 도표 및 그림 및 글에 딸린 캡션, 수식, 각주, 참고문헌, 요약문(Abstract), 질의응답(Q&A) 등이 있음.

- ㅇ 다양한 형식의 원문을 TXT 파일도 변환하는 과정에서 특수문자 등 원문의 글자가 훼손되는 경우도 많고, 원문에 가독성이나 미관을 위하여 고의로 맞춤법을 지키지 않은 사례도 있기 때문에 고품질의 말뭉치를 구축하기 위해서는 수작업 및 크라우드 소싱을 통해 작업 완료함

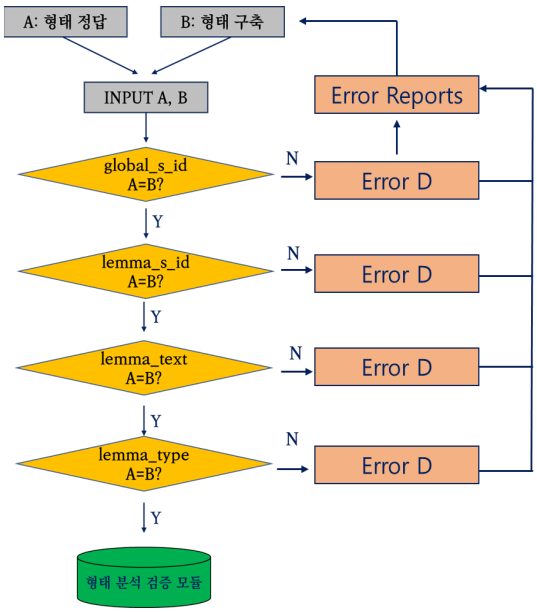
- ㅇ 데이터 제거, 데이터 자르기, 비식별화 등의 과정을 거쳐 원문 추출 작업 진행

4.3 정제 기준

- ㅇ 원천데이터 검증 체크리스트

검증 지표	검증 내용
키-값 쌍(key-value pair) 적합성(relevance) 검사	각 키(key)에 대해 허용되지 않는 값(value)이 있는가?
중복값 검사	이야기 ID 및 문장에서 중복이 발견되는가?
결측값(missing value) 검사	주제, 제목, 문장 등의 항목 중 누락된 정보가 있는가?

ㅇ 원천데이터 오류 검증 방안
- 원천데이터 공정 프로세스에 적합하도록 분석 층위별 개별 모듈을 설계
- 형태, 일관성, 내용, 통합 오류를 순차적으로 검증할 수 있도록 설계
- 형태 분석 검증 모듈은 모든 말뭉치가 공유하는 내용을 검증



검증 항목명	검증 내용
형태 정답	형태 분석 정답 세트
형태 구축	형태 분석 말뭉치
global_s_id	원시 말뭉치부터 공유하는 문장 번호
lemma_id	문장 내 형태소 분절 ID
lemma_text	분절된 형태소
lemma_type	형태소 주석

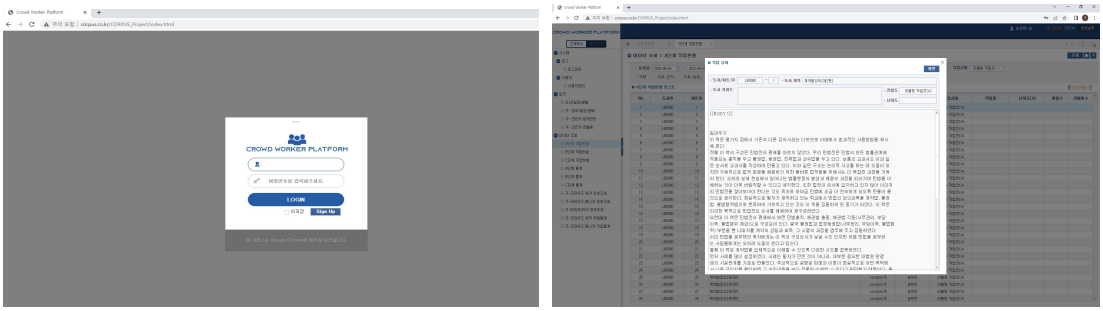
4.4 정제 조직

ㅇ 도서 1종에 대해 세 명의 크라우드워커가 각각 정제한 결과를 전문관리자가 최종 비교 검토하여 최종 XML 데이터를 만듦. 이러한 프로세스에 대해서는 크라우드워커에게 사전 공지하여 보다 꼼꼼하게 작업하도록 유도함.

ㅇ 크라우드워커(해당업무 전담인력)이 크라우드워커 플랫폼을 활용하여 데이터 정제 가이드라인에 따라 텍스트 추출된 데이터의 정제 작업 진행함.

4.5 정제 도구

ㅇ 자체 개발한 크라우드워커 플랫폼(corpus.co.kr)을 통해 품질기준의 구문적 정확성과 의미적 정확성을 높이기 위한 작업 환경 구축/운영



[그림] 크라우드워커 플랫폼 로그인 화면(좌) 및 플랫폼 정제 작업 화면(우)

5. 가공(라벨링)

5.1 가공(라벨링) 절차

ㅇ 크라우드워커 플랫폼(corpus.co.kr)에서 데이터 정제 단계로 "작업 완료" 될 일감에 대해 의료 및 법률 전문분야 크라우드워커가 라벨링 가이드라인에 따라 태깅 수행

ㅇ 라벨링 작업 시 질의가 있을 경우 플랫폼에서 제공하는 메모 기능으로 크라우드 매니저에게 문의하도록 하여 라벨링의 품질과 업무 속도를 높임

5.2 가공(라벨링) 기준

5.2.1 분류(classification)

① CATEGORY
: 의료 및 법률의 세부 분야에 대해 라벨링함. 의료분야는 대분류로 기초의학, 임상의학의 2개 분류로 나누며, 상세 분류로 기초의학 분야 하위 7개 분야, 임상의학 분야 하위 21개 분야로 나눔. 법률분야는 대분류로 법률일반, 공법, 사법, 분야별법의 4개 분류로 나누며, 상세 분류로 법률일반 1개 분야, 공법 하위 5개 분야, 사법 하위 5개 분야, 분야별법 분야 하위 13개 분야로 나눔('[표] 의료 및 법률분야 분류안' 참조)

② POPULARITY : 텍스트에 대한 독해 난이도로서 3단계로 나눔.
1: 일반인이 쉽게 이해
2: 일반인이 어느 정도 이해
3: 일반인이 이해하기 어려움

③ KEYWORD : 해당 분야의 전문라벨러가 라벨링을 하면서 해당 텍스트의 키워드를 선정하여 태깅

5.2.2 정밀 라벨링

ㅇ 정밀 라벨링 프로세스는 **국립국어원 2020 개체명 분석 말뭉치 구축 지침 (ver.1.6)을 준용**

[표] 국립국어원 2020 개체명 분석 말뭉치 구축 지침의 의료 및 법률 분야와 관련된 분류 명칭

의료 분야	STUDY_FIELD(FD) 학문분야	FD_MEDICINE (의학 관련 학문 분야 및 분과) 예) 의학, 병리학, 내과
	THEORY(TR) 특정 이론, 법칙, 원리	TR_MED_PROCEDURE (의학 요법/처방, 수술 및 처치) 예) 인공호흡, 요혈자극법, 바이오피드백 TR_MED_MEASURE (신체검진/검사) 예) 맥진, 내시경검사, 혈액검사, 검안 TR_MED_OBSERVATION (문진 내용) 예) 사회적 사실, 라이프스타일, 기왕력(과거력), 가족력

	ARTIFACTS(AF) 사람에 의해 창조된 인공물	AFW_MED_DEVICE 예) 의료기기
	ORGANIZATION(OG) 기관 및 단체	OGG_MEDICINE (의학/의료 기관/단체) 예) 세계보건기구, 고려대학교 안암병원
	ANIMAL(AM) 사람을 제외한 짐승	AM_PART (동물이나 사람 몸의 한 부분(신체 부위) 명칭) 예) 목, 살, 등, 배, 몸통, 손, 입, 갈기
	TERM(TMM)	TMM_DISEASE (질병 명칭) 예) 폐결핵, 기관지확장증, 레트증후군 TMM_SYMPTOM (증상) 예) 기침, 피고름가래 TMM_DRUG (약/약품명) 예) 헤파린, 근이완제, 모기향, 구충제 TM_CELL_TISSUE_ORGAN (세포/조직/기관 명칭 (외부에서 보이지 않는 조직이나 기관)) 예) 간상세포, 각막세포, 고막, 뉴런, 늑연골, 식도, 콩팥, 비장
법률 분야	ORGANIZATION(OG) 기관 및 단체	OGG_LAW (법률 기관/단체) 예) 광주지방법원, 저작권심의조정위원회
	CIVILIZATION(CV) 문명/문화	CV_LAW (법률 명칭) 예) 민법, 형법 CV_TAX (조세 명칭) 예) 소득공제, 면허세, 부가가치세
	EVENT(EV) 특정 사건/사고/행사 명칭	EV_OTHERS (사건/사고/사태, 회의/학술대회, 전시회, 경진대회 명칭) 예) 세월호 참사, 다보스포럼
	TERM(TML)	TML_PARTY (당사자) TML_PROCEDURAL_ACTS (소송행위) TML_LITIGATION (절차) TML_PROVISION (조문) TML_CASE (판례) TML_JUDGMENT (소송결과)

5.3 가공(라벨링) 조직

o 훈련된 크라우드워커(해당업무 전담인력)가 크라우드워커 플랫폼을 활용하여 정제된 데이터의 가공 작업 진행

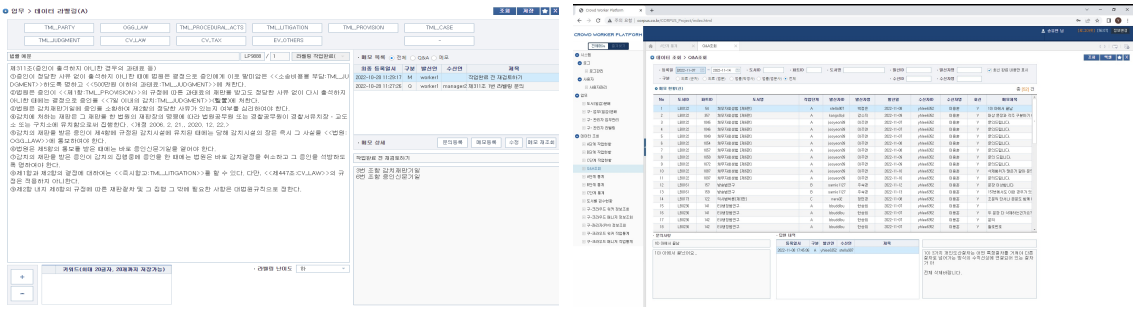
5.4 가공(라벨링) 도구

o 크라우드워커 플랫폼 활용

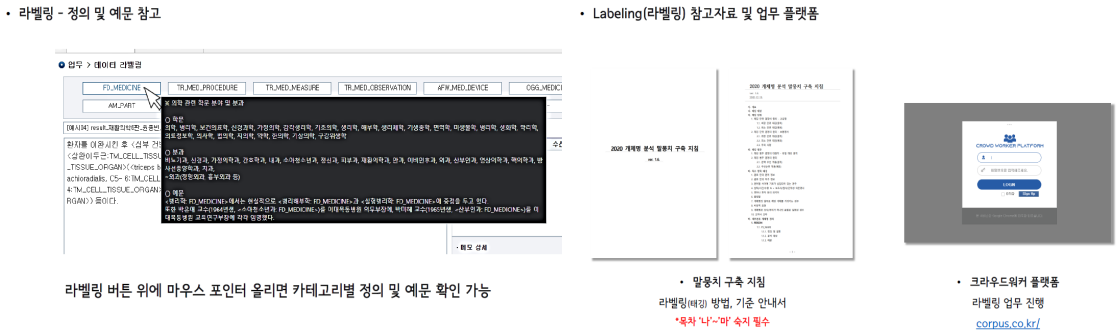
- 자체 개발한 크라우드워커 플랫폼(corpus.co.kr)에서 실제 태깅 작업과 더불어 난이도, 키워드 등의 분류 작업도 동시 수행하여 다양성, 정확성 등의 품질 지표 목표에 도달할 수 있는 작업 환경을 구축 및 운영함.

o 라벨링 가이드라인 제공

- 플랫폼 화면에서 라벨링 가이드라인 기준으로 제시된 의료/법률 분야별 카테고리를 해당하는 용어에 라벨링 작업 수행



[그림] 라벨링 작업 화면(좌) 및 메모 및 Q&A 기능(우)

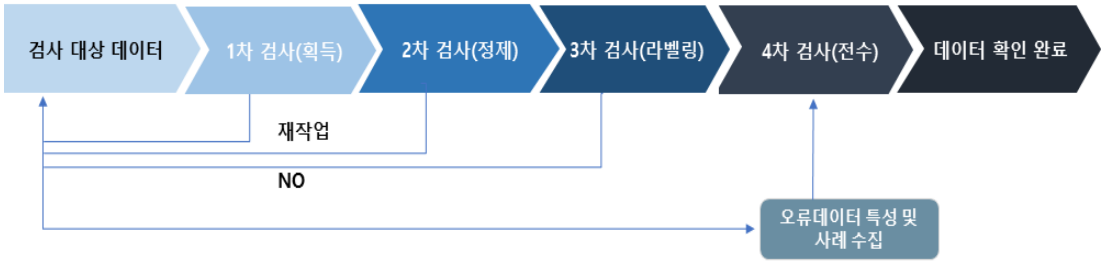


[그림] 가공 매뉴얼

6. 검사

6.1 검사 절차

○ 데이터 검사 프로세스



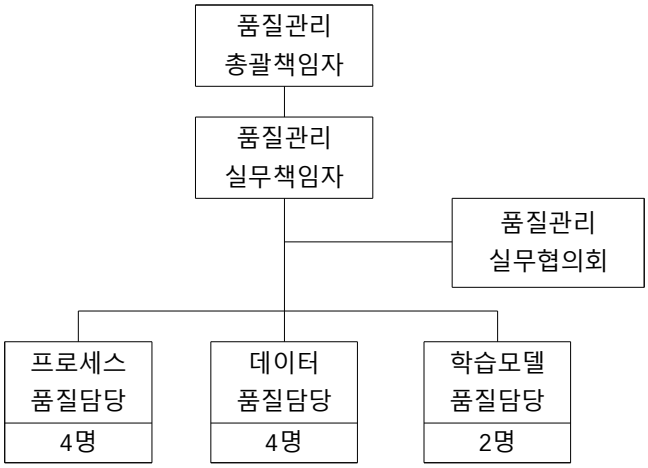
○ 검사 및 피드백 절차

- 데이터의 정확성 확보 및 빠른 검수결과 피드백을 통한 효율적인 데이터셋 구축을 위해 단계별 품질관리체계를 확립하고 리스크에 대한 사전 대응 방안 및 중점 관리 방안을 수립
- 효율적인 품질관리를 위한 다단계 품질검수를 실시
  . 1단계(획득) : 법적 토대를 중시하여 수집, 서적 중요도 평가에 의해 선정된 우선 순위 목록에 의한 수집, 원문 활용 가능한 최종본 수집
  . 2단계(정제) : 활용 가능한 문단 확보(불필요한 정보 삭제)
  . 3단계(가공) : 전문 용어 분류 기준에 적합한 적용
  . 4단계(검수) : 전문 검수자에 의한 검증, 프로토타입을 통한 기계적 자체검증 실시

6.2 검사 기준

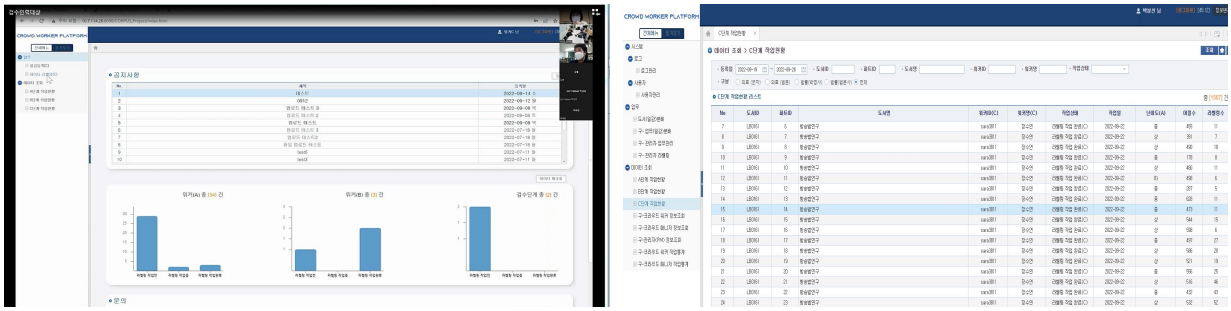
구분	지표	품질목표	품질목표 달성기준
구축 공정	적합성	95%이상	원천데이터 포맷 XML 형식 오류율 5% 이하
	유용성	95%이상	라이선스를 확보한 의료 및 법률 분야 핵심 레퍼런스 도서 선정률 95% 이상
데이터 적합성	분야별 다양성	필수요건	의료, 법률 분야별 하위 분류를 각각 10개 이상 포함 개별 하위 분류는 750만 어절(15%) 이하
	통계적 다양성	95%이상	다양성 지표와 체크리스트를 작성하여 95% 이상 만족
데이터 정확성	데이터 라벨링 정확성	0.90이상	최종관리자(매니저)가 검토완료한 일감 중 **2%**를 선별하여 의학분야 총괄책임자(의대교수) 및 법률분야총괄책임자(법학전문대학원교수)가 직접 작업한 분야별 태그 결과(정답)와 신뢰도(Kohen&s kappa)를 분석하여 0.90 이상
	데이터 검수	95%이상	샘플링 데이터의 적합성과 정확성이 별도 검수 절차 요건에 95% 이상 부합
	전문가 정성 평가	4.0이상	2인 이상의 의료 및 법률 전문가를 고용하여 샘플링 데이터의 평균 의견 점수(Mean Opinion Score)를 산출하여 4.0 이상

6.3 검사 조직

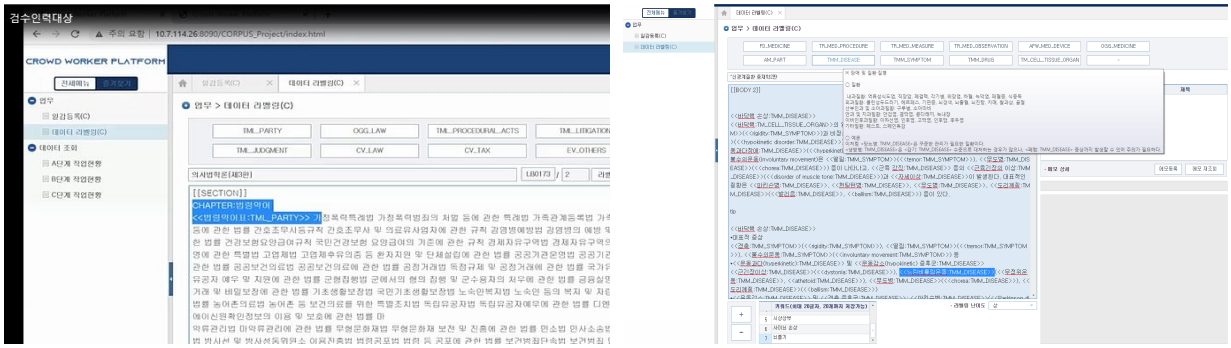


조직 구분	역할과 책임
품질관리 총괄책임자	• 인공지능 학습용 데이터의 품질관리 총괄 • 실무책임자 지정
품질관리 실무책임자	• 인공지능 학습용 데이터의 품질관리 실무 총괄 • 품질관리 계획 수립 • 품질관리 실무 수행조직 운영/품질관리 활동 • 품질관리협의체 구성 및 운영
품질관리 실무협의회	• 품질관리 관련 실무차원의 주요사항을 검토 • 품질관리 계획의 적정성 평가 /보완 • 구축과정에 발견된 데이터 품질 이슈 논의 및 해결방안 제시
프로세스 품질관리 담당	• 구축 과정(획득, 정제, 라벨링 등)별로 구분하여 담당자 지정 • 품질관리계획에 명시된 품질관리 활동의 준수 여부 확인 • 품질이슈 발생 시 이를 기록하고 개선하는 활동 수행 • 즉각적인 조치가 불가능하거나 협의가 필요한 사항은 실무책임자에게 보고
데이터 품질관리 담당	• 품질계획에 수립된 검사기준과 절차를 준수하여 품질검사 수행 및 개선활동 • 의미정확성, 구문정확성, 통계적 다양성 검사 계획 수립 및 품질점검 수행
학습모델 품질관리 담당	• 학습용 데이터로 훈련시키는데 적합한 인공지능 알고리즘의 유효성 검사 계획 수립 및 검사

6.4 검사 도구



[그림] 작업 현황 통계 화면: 각 단계별, 개인별 작업 현황을 한 눈에 확인할 수 있음(좌) 및 작업 현황 조회 화면: 단계별 작업자의 작업 목록을 조회하여 작업자별로 품질 관리(우)



[그림] 작업 화면: 라벨링 되어 있는 문서 단위로 전수 검수 진행(좌) 및 Q&A: 작업 중 질의 응답을 통해 현안 문제 실시간 해결(우)

7. 학습 모델

7.1 학습 모델 후보

7.1.1 학습 모델 후보군

구분	내 용
KM-FastText	• Gensim 라이브러리에 구현된 FastText 훈련 과정을 활용함 - Gensim은 Word2Vec, FastText 등의 다양한 자연어처리 알고리즘을 구현한 오픈 소스 프로젝트 - Gensim model로 해당 모델을 범용성 있게 공개함으로써 사용자가 손쉽게 모델을 활용하여 유사도 연산 등을 할 수 있음 - 다양한 하이퍼파라미터로 모델을 훈련하여 최적의 하이퍼파라미터를 찾기 용이 • 단어를 문자단위 n-gram으로 변환 - 이 때, 한국어의 특성을 고려하여 1~3 정도의 n값을 활용 - 구체적인 n값은 실험을 통해 결정 • skip-gram 방식으로 문자단위 n-gram 학습 - 단어가 주어졌을 때 맥락에 대한 log likelihood를 최대화 • SGD 방식으로 최적화 진행
KL-FastText	• 기본적인 방식은 KM-FastText와 동일 • 단, 법률용어의 경우 실질형태소의 길이가 상대적으로 길 수 있음 • 따라서, 의료데이터보다 높은 n 값을 사용하여 n-gram을 추출함
KM-BERT	• Devlin 외(2019)의 방법을 따라 pre-train 진행 - Masked Language Model -Next Sentence Prediction
KL-BERT	• 기본적인 방식은 KM-BERT와 동일

7.1.2 학습 모델 선정기준

구분	내용	
학습모델 후보1 (KM-FastTex)	품질지표	• Spearman rank correlation ρ (단어 쌍 유사도)
	선행연구	• 2021 JMIR Med Inform에서 namuwiki 말뭉치로 구축한 FastText 모델의 표준 의료 용어 단어 쌍 세트에 대한 Similarity ρ 값이 0.12, Relatedness ρ 값이 0.02를 기록
	지표기준	• 표준 의료 용어 단어 쌍 세트에 대해 namuwiki 말뭉치로 구축한 FastText 모델 대비 Similarity 혹은 Relatedness에 대한 ρ 값이 향상되었는지 여부를 평가
학습모델 후보2 (KL-FastText)	품질지표	• Spearman rank correlation ρ (단어 쌍 유사도)

	선행연구	• 법률 분야에 대해서는 선행 연구가 없음
	지표기준	• 표준 의료 용어 단어 쌍 세트에 대해 namuwiki 말뭉치로 구축한 FastText 모델 대비 Similarity 혹은 Relatedness에 대한 ρ 값이 향상되었는지 여부를 평가
학습모델 후보3 (KM-BERT)	품질지표	• MLM accuracy • NSP accuracy • F1-score
	선행연구	• 2021 대한의료정보학회 춘계학술대회에서 의료분야 표준 문장 쌍 데이터세트(MedSTS)를 한국어로 번역하여 평가하였을 때 KR-BERT의 Pearson correlation이 0.823, Spearman correlation이 0.811을 기록하였다고 발표 • 2021 한국컴퓨터종합학술지에 출판된 의료 개체명 인식과 정규화에 대한 연구에서 BERT-base multiligual 모델의 증상, 신체부위, 질병에 대한 개체명 인식 F1-score가 macro-avg 0.89, micro-avg 0.92를 기록
	지표기준	(1) 다른 코퍼스와의 성능 비교 • 검증을 위해 전문도서외 다른 코퍼스 1가지이상 구축 (위키피디아, 기사, 논문 등) • 대조군: KR-BERT에 다른 코퍼스 5000만 어절 Pre-training (KR-BERT_new) • 비교군: KR-BERT에 전문 도서 5000만 어절 Pre-training (KM-BERT) • 지표기준: ① MLM accuracy - 0.9이상 혹은 대조군 대비 0.05이상 향상 ② NSP accuracy - 0.9이상 혹은 대조군 대비 0.05이상 향상 (2) 분야별 태그 분류 성능 비교 • 분야별 태그가 된 5000만 어절 말뭉치를 training set 및 test set로 8:2로 나눔 • 대조군: KR-BERT를 training set로 fine tunning • 비교군: KM-BERT를 training set로 fine tunning • 지표기준: F1 score - 0.9이상 혹은 대조군 대비 0.05이상 향상
학습모델 후보4 (KL-BERT)	품질지표	• MLM accuracy • NSP accuracy • F1-score
	선행연구	• 2022 정보처리학회지에 출판된 법률분야 한국어 의미 유사판단에 관한 연구에서 Wikipedia 말뭉치로 pre-train한 BERT를 법률 분야 문장 쌍 유사도 평가 task를 fine-tuning 만든 모델의 유사 문장 라벨링 여부에 대한 Score를 평가 • NER관련 법률 분야에 대해서는 선행 연구가 없음
	지표기준	• KM-BERT와 동일한 방식으로 진행

7.2 학습 모델 개발

7.2.1 의료 전문 도서 말뭉치

ㅇ FastTEXT를 활용한 의료분야 한국어 단어 의미 유사판단 인공지능 모델 개발
- 의료분야 한국어 단어의 의미 유사도를 판단하는 FastTEXT 모델 개발
  : KM(Korean Medical)-FastTexT

ㅇ BERT를 활용한 의료분야 한국어 인공지능 모델 개발
- 의료분야 한국어 문장의 의미 유사도 판단 및 단어를 탐지하여 적절하게 분류하는 BERT(Bi-directional Encoder Representations from Transformers) 모델 개발 : KM-BERT

구분	내 용
KM-FastText	• Gensim 라이브러리에 구현된 FastText 훈련 과정을 활용함 - Gensim은 Word2Vec, FastText 등의 다양한 자연어처리 알고리즘을 구현한 오픈소스 프로젝트 - Gensim model로 해당 모델을 범용성 있게 공개함으로써 사용자가 손쉽게 모델을 활용하여 유사도 연산 등을 할 수 있음 - 다양한 하이퍼파라미터로 모델을 훈련하여 최적의 하이퍼파라미터를 찾기 용이함 • 단어를 문자단위 n-gram으로 변환 - 이 때, 한국어의 특성을 고려하여 1~3 정도의 n값을 활용 - 구체적인 n값은 실험을 통해 결정 • skip-gram 방식으로 문자단위 n-gram 학습 - 단어가 주어졌을 때 맥락에 대한 log likelihood를 최대화 • SGD 방식으로 최적화 진행
KM-BERT	• Devlin 외(2019)의 방법을 따라 pre-train 진행 - Masked Language Model - Next Sentence Prediction

7.2.2 법률 전문 도서 말뭉치

ㅇFastTEXT를 활용한 의료분야 한국어 단어 의미 유사판단 인공지능 모델 개발
- 법률 분야 두 한국어 단어의 의미 유사도를 판단하는 FastTEXT 모델 개발
  : KL(Korea Legal)-FastTexT

ㅇ BERT를 활용한 법률분야 한국어 인공지능 모델 개발
- 법률 인공지능의 주요 임무는 법률 문서 업무를 보조하는 것임. 법률 문서는 판결, 주문, 청구취지 및 기초 사실, 피고와 원고의 주장 요지, 관계 법령으로 구성되어 있음.
- 방대한 양의 법률 문서를 읽고 기초 사실과 관계 법령을 종합하여 주문과 청구취지를 작성하는 것이 판견물 작성하는 것이 판사의 업무임.
- KL-BERT는 판결문을 작성하는 임무를 수행하기 위하여 기초 사실/주장의 요지를 학습하여 주문과 청구취지를 생성하는 훈련을 함.

- 법률 분야 두 한국어 문장의 의미 유사도 판단 및 단어를 탐지하여 적절하게 분류하는 BERT(Bi-directional Encoder Representations from Transformers) 모델 개발 : KL-BERT

구분	내 용
KL-FastText	• 기본적인 방식은 KM-FastText와 동일 • 단, 법률용어의 경우 실질형태소의 길이가 상대적으로 길 수 있음 • 따라서, 의료데이터보다 높은 n 값을 사용하여 n-gram을 추출함
KL-BERT	• 기본적인 방식은 KM-BERT와 동일

제3장 데이터 활용

1. 데이터 활용

데이터명	AI 모델	모델 성능 지표	응용서비스(예시)
의료분야 전문도서 말뭉치	KM-FastText	• Medical word-pair set (Correlation)	• 질환 검색 서비스
	KM-BERT	• Masked language model (Accuracy) • Next sentence probability (Accuracy) • Named-Entity-Recognition (F1 score) • MedSTS (Pearson correlation)	• 진료과 추천서비스
법률분야 전문도서 말뭉치	KL-FastText	• Legal word-pair set (Correlation)	• 법률 검색 서비스
	KL-BERT	• Masked language model (Accuracy) • Next sentence probability (Accuracy) • Named-Entity-Recognition (F1 score)	• 판례 요약/추천 서비스 • 법률 위반 조항 탐지기 • 소비자 불공정 계약 조항 탐지기

2. 응용 서비스

ㅇ 의료분야에서는 챗봇을 통해 환자가 병원에서 퇴원한 후에도 지속적으로 의료 서비스를 제공 받을 수 있는 비대면 서비스를 제공해 줄 수 있음. 대용량 의료 말뭉치를 기반으로 개발한 인공지능 기술은 상시적인 진찰, 환자 간호 지원, 약물관련 질문에 대한 빠른 답변, 지속적인 건강 모니터링 서비스를 제공해줌으로써 원격 의료시장의 새로운 방향을 제시해줄 수 있음.

ㅇ 의료분야 데이터 말뭉치를 이용한 활용 방안- 음성 인식기술을 통한 처방 및 차트 작성을 통한 진료 프로세스 서비스 제공- 대화형 메신저 통해 환자의 증상을 미리 문진하여 진료과 추천 서비스 제공이 가능함.

ㅇ 법률 분야는 문제에 대한 정확한 판례가 존재하는 분야로서 과거의 다양한 분쟁과 그에 대한 판례가 등록되어 있고, 그 처리 과정에서 법률적 해석이라는 규칙을 적용하여야만 해결되는 구조가 일관성을 가지고 있기 때문에 관련 시장의 규모와 신규 창업자의 진입이 증가하고 있음. 이러한 이유로 인해 대용량 말뭉치를 이용한 법률 서비스 기술 개발의 필요성이 증가하고 있음. 법률 분야에서는 대용량 말뭉치를 이용해 용어 검색 서비스와 실시간 비대면 상담 서비스 개발에 활용 가능하며, 전문성과 보안성이 요구되는 일상생활의 법률 문제점을 해결해 줄 수 있는 접근성이 용이한 법률 플랫폼 개발이 가능함.

ㅇ 법률분야 데이터 말뭉치를 이용한 활용 방안- 법령, 판례, 유사 상담 정보를 기반으로 일상생활에서 사용되는 용어를 이용한 법률 검색 서비스 제공- 법적 문제가 발생한 시점에 상관없이 사용자의 질문을 실시간으로 해석하고 적합한 법률 내용을 자동 제공하는 비대면 자동 상담 서비스 제공할 수 있음.