

21년도 인공지능 학습용 데이터 구축 가이드라인

< 웹데이터 기반 데이터셋 >

※ 지정공모 과제의 경우 소분류 단계인 “세부 데이터명” 건별로 작성하여 주시기 바랍니다. 단 세부 데이터셋이 공통적이고 복수인 경우 부제에 제목 명기 가능.

인공지능 데이터 구축	사업 총괄	솔트룩스
	데이터 설계	솔트룩스
	데이터 수집 및 정제	알토비전
	데이터 가공	비플라이소프트, 이노그루, 오픈큐비트
	데이터 검수	솔트룩스, 소리자바, 알체라
	클라우드 소싱	비플라이소프트, 이노그루, 오픈큐비트 솔트룩스, 소리자바, 알체라
	저작도구 개발	솔트룩스
	AI모델 개발	솔트룩스
가이드라인 작성	솔트룩스	방재준
	솔트룩스	오상중
가이드라인 버전	ver 1.0 ('22. 1. 15)	

목 차

1. 데이터 명세 정보 1

1.1 데이터 정보 요약 1

1.2 데이터 포맷 2

1.3 어노테이션 포맷 3

1.4 데이터 구성 3

1.5 데이터 통계 4

1.6 원시데이터 특성 6

1.7 기타정보 6

2. 데이터 구축 가이드 8

2.1 데이터 구축 개요 8

2.2 문제정의 8

2.3 수집·정제 9

2.4 어노테이션/라벨링 11

2.5 검수 15

2.6 활용 15

1. 데이터 명세 정보

1.1 데이터 정보 요약

데이터 이름	웹데이터 기반 대규모 한국어 말뭉치 데이터	
활용 분야	BERT와 GPT같은 대규모 한국어 서비스, 한국어 문법 교정기 고도화, 한국어 문서 핵심 요약 서비스, 신조어 이해 사전 서비스 고도화 등에 활용	
데이터 요약	웹사이트 기반 대용량의 텍스트 데이터를 수집 후, 전사도구를 활용하여 구축된 타이틀, 단락 제목, 본문 텍스트가 구조화된 10억 어절의 AI 학습 데이터 셋	
데이터 출처	머니투데이, 뉴스1, 뉴시스, 스타뉴스, 메가뉴스 기사 등	
데이터 이력	배포버전	ver 1.0 ('22. 1. 15)
	개정이력	신규
	작성자/ 배포자	솔트룩스 방재준 / 솔트룩스 오상중

1.2 데이터 포맷

항목		설명
Header		
	identifier	식별자
	name	파일명
	category	카테고리 종류
	type	타입 구분
	source_file	소스 파일명
	source	데이터 출처 구분
	subject	데이터 카테고리명
라벨링 결과		
named_entity		개체명 라벨링 결과 배열
	title	데이터셋 제목
	subtitle	데이터셋 소제목
	content	데이터셋 본문
	id	개체명 일련번호
	text	개체명 텍스트
	tag	개체명 구분

1.3 어노테이션 포맷

단계	수준1	수준2	수준3	수준4	데이터 타입	필수값 여부	설명	유효값/허용범위/예시
라벨링	header				Object	Y	데이터셋	
		identifier			string	Y	데이터셋 식별자	ex) 데이터유형, 목적, 순번
		name			string	Y	데이터셋 이름	
		category			number	Y	데이터셋 카테고리	0: 텍스트
		type			number	Y	데이터셋 타입	0: 텍스트
		source_file			string	Y	소스 파일명	1.1. 파일명 부여 방식 참조
	named_entity	source			string	Y	데이터 출처	0: 뉴스 1: 게시판 2: 위키
		subject			string	Y	데이터 주제	source_file 항목의 3~4번째 문자
					List	Y	개체명/신조어 목록	
		title			List	Y	제목	
			sentence		string	Y	제목 내용	
			labels		List	Y	제목 분석 목록	
				id	number	Y	제목 분석 일련번호	
				text	string	Y	제목 분석 항목	
			tag	string	Y	Y	제목 분석 태그	개체명/신조어 태그
		subtitle			List		소제목	
			sentence		string		소제목 내용	
			labels		List		소제목 분석 목록	
				id	number		소제목 분석 일련번호	
				text	string		소제목 분석 항목	
			tag	string			소제목 분석 태그	개체명/신조어 태그
		content			List		본문	
			sentence		string	Y	본문 내용	
			labels		List	Y	본문 분석 목록	
				id	number	Y	본문 분석 일련번호	
				text	string	Y	본문 분석 항목	
			tag	string	Y	Y	본문 분석 태그	개체명/신조어 태그
		board			string		게시판	태그 작업 대상
		writer			string		작성자	태그 작업 대상
		write_date			string	Y	작성일시	태그 작업 대상, yyyy-MM-dd
		url			string		출처 URL	태그 작업 대상
		source_site			string		출처 사이트	태그 작업 대상

1.4 데이터 구성



<데이터 저장소 실제 예시>

- 파일명 부여 규칙

말뭉치 유형 구분	데이터 유형 분류
B: 대규모 말뭉치	W: 웹데이터 기반

1.5 데이터 통계

1.5.1 데이터 구축 규모

- 원천데이터 : 원시데이터(웹데이터)로부터 비문, 비속어, 편향성, 비식별화 등 정제된 JSON 구조의 텍스트 데이터 10억 어절 이상
- 라벨링 데이터 : 원천데이터로부터 JSON구조의 구문에 맞게 개체명과 신조어에 대하여 태깅된 텍스트 데이터 10억 어절 이상

1.5.2 데이터 분포

1. 구축 데이터 : 1,175,464,020 어절

2. 카테고리 및 출처별 분포

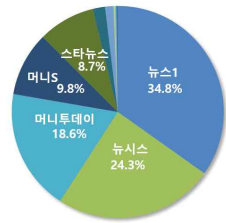
분류 코드	분류명	파일 수
CU	문화/패션/뷰티	1,535
EC	경제	2,280
EN	연예	2,752
ID	산업	3,420
IN	국제	3,371
LI	라이프스타일	3,962
TL	여행레저	3,732
HB	취미	3,178

HE	건강	3,562	68,648,804	5.8%
PO	정치	4,403	93,955,104	8.0%
SC	IT/과학	3,672	41,970,255	3.6%
SG	사회일반	4,501	74,946,093	6.4%
AC	사건사고	4,063	65,350,073	5.6%
WO	여성복지	4,153	49,970,850	4.3%
ED	교육	6,140	57,060,422	4.9%
LC	지역	5,996	57,954,012	4.9%
SP	스포츠	5,031	62,942,643	5.4%
합 계		65,751	1,175,464,020	100.0%

지디넷	1.9%
더리더	1.3%
코스모폴리탄	0.2%
엘르	0.2%
디스이즈게임	0.2%
하퍼스 바자	0.1%
합 계	100.0%



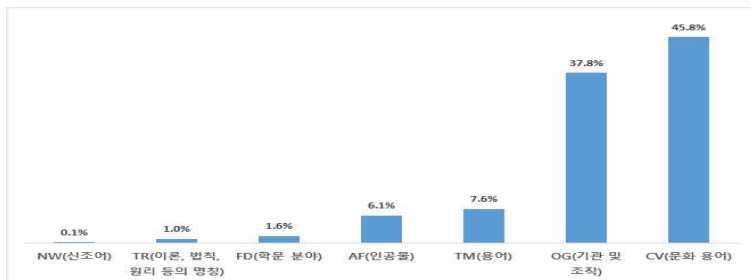
[카테고리별 어절 분포 그래프]



[출처별 분포 그래프]

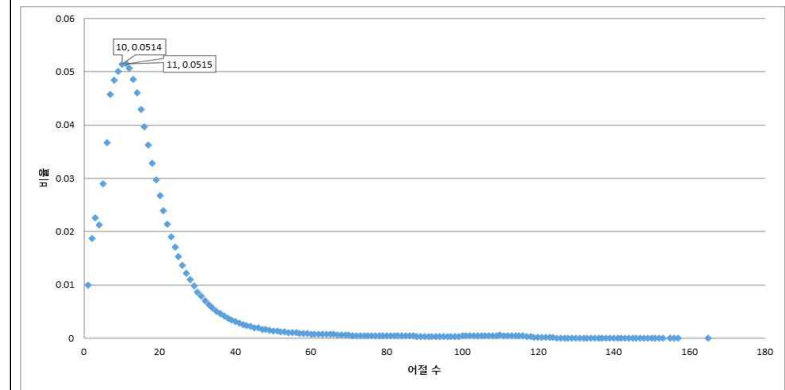
3. 개체명 및 신조어 태그별 분포

- CV(문화 용어)의 태그 수가 가장 높음



4. 문장 길이 분포(어절 기준)

- 가장 일반적인 문장의 어절 수는 10~11 어절



1.6 원시데이터 특성

1.6.1 대상분류

- 작성 시기, 출처, 주제(카테고리)

1.6.2 제약조건

- 제약 없음 : 저작권 활용 계약 체결 후 데이터 수집

1.6.3 속성

- 텍스트

1.7 기타정보

1.7.1 포괄성

- 다양한 온라인 뉴스 채널 크롤링

1.7.2 독립성

- 국립국어원, NIA, 모두의 말뭉치, AI Hub 등과 같이 기 구축/공개된 말뭉치와의 중복확인 결과 중복 없음

1.7.3 유의사항

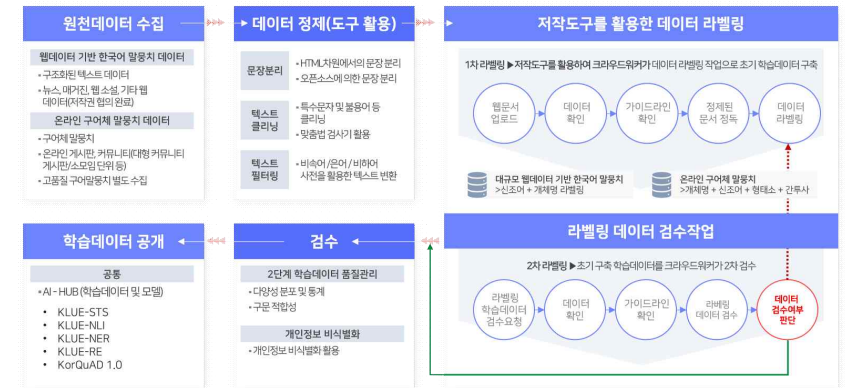
- 2010년 이전 데이터를 포함하고 있음

1.7.4 관련 연구

- 데이터 활용한 논문 없음

2. 데이터 구축 가이드

2.1 데이터 구축 개요



[그림1] AI 학습데이터 구축 프로세스

2.2 문제정의

2.2.1 임무 정의

- 데이터 경제로의 패러다임 변화
 - 4차 산업혁명 시대로 급속 진입하면서 제조업, 서비스업 중심의 한국경제는 도태의 위기에 직면하게 됨. 특히 코로나19로 인한 극심한 경기 침체와 함께 데이터 경제로의 패러다임 전환이라는 이중 사업을 해결해야 하는 시점
 - 데이터를 기반으로 한 인공지능의 시대가 도래함에 따라 인공지능 시대의 석유라고 일컫는 기초 데이터의 국적 차원의 확보 및 제공이 글로벌적인 경쟁력 확보의 필수 요소이며, 데이터 확보가 이루어져야 비로소 디지털 시대로의 전환기를 맞을 수많은 기업과 스타트업 그리고 국가 공공 행정 서비스의 미래 선도형 경제 실현이 가능한 시점에 도달함
 - 구글, 아마존 등 글로벌 IT 대기업은 빅데이터의 축적과 함께 다양한 AI 혁신기술을 공개하며 수많은 형태의 새로운 산업과 서비스 영역을 개척하며 선보이고 있어 벌써부터 "디지털 독과점"이란 비판을 받고 있는 수준으로 앞서 나가고 있음
- 대규모 한국어 말뭉치 데이터가 필요한 이유
 - 모든 디지털 산업의 기초가 될 데이터는 80% 이상이 텍스트, 음성, 영상 등으로 되어 있음. 특히 딥러닝 언어모델 개발을 위하여는 대용량 텍스트 데이터가 필요하며, 현재 웹 기반의 다양한 한국어 데이터가 부족하여 대규모 웹데이터 기반 한국어 데이터 구축이 필요함
 - 따라서 대규모 웹데이터 기반 한국어 데이터 구축 타이틀, 단락 제목, 본문 텍스트가 구조화된 텍스트 데이터를 수집하여 범용 용어 및 고유명사, 우리말샘 용어 데이터 확장에 활용하는 대규모 한국어 말뭉치 데이터 구축 사업은 인공지능 학습용 데이터 구축 사업의 근간이 되는 중요한 부분이라 할 수 있음

2.2.2 데이터 구축 유의사항

- 저작권 이용 허락 계약
 - 메가 뉴스의 경우 보유 미디어(ZDnet, CNET, 뉴스엔게임)의 국내 작성 기사 저작권은 메가 뉴스에 있으므로 회사 측과의 저작권 사용 계약을 체결 하였음으로 저작권 문제 해결함
- 개인정보 비식별화 및 불용어 클리닝
 - 원천데이터의 개인정보 비식별화는 개인정보보호법에 따라 1차, 2차로 나누어 정제하여 마스킹 처리하고, 비속어 및 특정 인물 비하 발언 등의 불용어는 별도 사전을 구축하여 정제한다.
- 데이터 수집 시 민감정보 삭제
 - 원시데이터의 수집 시 비방글, 성차별, 정치 편향성, 종교 등 민감정보에 대하여 데이터 수집 후 삭제하고 원천데이터로 가공함

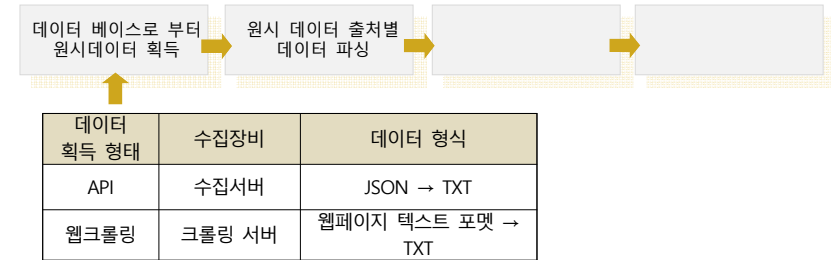
2.3 수집·정제

2.3.1 원시데이터 선정



- 국립국어원, NIA, 모두의 말뭉치, AI Hub 등과 같이 기 구축/공개된 말뭉치와의 중복성을 확인하고, 구축되지 않은 원시 데이터를 수집
- 저작권 확보를 통해 콘텐츠를 제공 받는 경우 계약서상에 중복 문서 검토를 수행한 결과물을 제공 받을 수 있도록 계약사항에 명시
- 데이터를 공급 받는 기관을 통해 수집 계약 시 데이터 수집 분야, 수집기간 등을 명시한 자료를 별첨 문서로 확보

데이터 포맷	json, txt, csv			
데이터 획득 규모	타이틀, 단락 제목, 본문 텍스트가 구조화된 텍스트의 인공지능 학습 데이터 12억 어절			
데이터 출처	메가뉴스, 클라우드 소스를 통한 웹 기반 원시 데이터			
수집 방법	저작권 협약을 통해 다양한 뉴스 웹사이트에서 크롤링을 통해 프레임의 일부를 추출하여 텍스트 수집			
수집 카테고리	분류코드	분류명	분류코드	분류명
	SC	IT/과학	AC	사건사고
	CU	문화/패션/뷰티	HE	건강
	IN	국제	HB	취미
	WO	여성복지	SG	사회일반
	ED	교육	TL	여행레저
	LC	지역	PO	정치
	LI	라이프스타일	EC	경제
	SP	스포츠	EN	연예
			ID	산업

2.3.2 수집·정제 절차



2.3.3 수집·정제 기준

기준		내 용																																				
수집 기준		<ul style="list-style-type: none">- 17개 카테고리에서 균등한 분포로 수집하여 다양성 확보- 최근 3개년 이내 작성된 데이터의 비중을 80% 이상으로 수집- 국립국어원, NIA, 모두의 말뭉치, AI Hub 등에서의 구축 데이터와 중복되지 않은 데이터 수집																																				
정제 기준	주요 데이터 분리	<ul style="list-style-type: none">- 딥러닝 언어 모델을 학습시키기 위한 대용량의 데이터를 구조화된 형식으로 만들기 위해 데이터를 분리. 웹 데이터를 제목, 소제목, 본문 형식으로 분리.																																				
	텍스트 클리닝	<ul style="list-style-type: none">- 웹으로부터 수집된 비정형 데이터는 수많은 특수문자 및 불용어를 포함. 이를 미리 사전 구축 및 규칙을 정의하여 대체 텍스트로 변환. <div><div><div>비속어 사전을 통한 1차 자동 정제</div><div>클라우드 워커를 통한 2차 수동 정제</div></div><div><ul style="list-style-type: none">▪ 현재까지 2,715개의 비속어(욕설, 혐오표현) 리스트 사전 구축 → 지속적으로 업데이트 예정▪ 두 글자 이상의 비속어 필터링 → 초기 1어절 비속어도 포함되어 있었으나 오류가 많아 두 글자 이상 비속어만 필터링▪ 비속어는 '비속어' 로 대체 → 예) 솔직히 그 상황이면 저라도 (비속어) 소리가 절로 날듯요.</div></div>																																				
			<div><div>1차 자동 정제된 데이터를 클라우드 워커가 수동으로 추가 확인</div><div>오직용 된 경우는 기존 데이터 단어로 수정, 자동으로 필터링 되지 않은 데이터는 직접 삭제 작업 진행</div></div>																																			
		<div>[텍스트 클리닝 정제 프로세스]</div>																																				
	개인정보 비식별화	<ul style="list-style-type: none">- 자유롭게 작성이 가능한 웹 데이터의 특성 상 개인정보가 포함될 수 있음. 이를 미리 성명 사전 구축 및 규칙을 정의하여 대체 텍스트로 변환. <table><tr><th>구분</th><th>표시</th><th>마크업</th><th>구분</th><th>표시</th><th>마크업</th></tr><tr><td>이름</td><td>@이름</td><td>&name&</td><td>건강보험번호</td><td>@건강보험번호</td><td>&health-insurance-num&</td></tr><tr><td>상호명</td><td>@상호명</td><td>&company-name&</td><td>계좌번호</td><td>@계좌번호</td><td>&bank-account-num&</td></tr><tr><td>주민등록번호</td><td>@주민번호</td><td>&social-security-num&</td><td>여권번호</td><td>@여권번호</td><td>&passport-num&</td></tr><tr><td>카드번호</td><td>@카드번호</td><td>&card-num&</td><td>주소</td><td>@주소</td><td>&address&</td></tr><tr><td>자동차번호</td><td>@자동차번호</td><td>&car-num&</td><td>전화번호</td><td>@전화번호</td><td>&tel-num&</td></tr></table> <div>[비식별화 대상 대체 텍스트]</div>	구분	표시	마크업	구분	표시	마크업	이름	@이름	&name&	건강보험번호	@건강보험번호	&health-insurance-num&	상호명	@상호명	&company-name&	계좌번호	@계좌번호	&bank-account-num&	주민등록번호	@주민번호	&social-security-num&	여권번호	@여권번호	&passport-num&	카드번호	@카드번호	&card-num&	주소	@주소	&address&	자동차번호	@자동차번호	&car-num&	전화번호	@전화번호	&tel-num&
구분	표시	마크업	구분	표시	마크업																																	
이름	@이름	&name&	건강보험번호	@건강보험번호	&health-insurance-num&																																	
상호명	@상호명	&company-name&	계좌번호	@계좌번호	&bank-account-num&																																	
주민등록번호	@주민번호	&social-security-num&	여권번호	@여권번호	&passport-num&																																	
카드번호	@카드번호	&card-num&	주소	@주소	&address&																																	
자동차번호	@자동차번호	&car-num&	전화번호	@전화번호	&tel-num&																																	

2.4 어노테이션/라벨링

2.4.1 어노테이션/라벨링 절차

- (1) 프로세스
- 클라우드 워커가 솔트룩스의 대규모 말뭉치 구축 저작도구를 활용하여 개체명, 신조어를 데이터 라벨링하고 구축 관리자가 검증
 - 라벨링 작업 방식은 반자동 방식
 - 범용 용어 및 고유명사는 자사의 개체명 인식기를 사용하여 작업자가 대략적인 후보를 확인할 수 있도록 이용



(2) 프로세스별 구축 상세

1. 솔트룩스의 대규모 한국어 말뭉치 구축 저작도구에서 클라우드 워커가 접속하여 프로젝트 할당

• 말뭉치 구축 탭에서 분석하고자 하는 프로젝트를 클릭합니다.

2. 프로젝트별 분석 탭 클릭

• 말뭉치 프로젝트에 들어가면 다음 그림과 같은 화면이 나타납니다.
• 다음 화면에서 분석하고자 하는 탭을 클릭하여 이동합니다.

3. 개체명, 신조어 탭에서 저작도구 엔진을 통한 태깅 확인

• 메시로 "신조어 분석"을 진행하면, 저작도구 엔진을 통하여 다음 그림과 같이 태깅이 되어 있습니다.

4. 개체명, 신조어 탭에서 미태깅된 부분은 클라우드 워커가 수작업으로 태깅 진행

• 엔진을 통하여 태깅이 되어 있지 않은 부분은 수작업으로 태깅을 진행해야 합니다.
• 태깅을 진행할 부분 '태깅' 탭이 비어있습니다. 이를 클릭하면 커서가 나타납니다.
• 커서가 나타되면 (ex. 해당명수화/WV) 를 입력하여 태깅을 합니다.

5. 클라우드 워커가 작업 완료 후 검증 관리자에게 승인 요청

6. 말뭉치 구축 검증 관리자가 검증 후 재작업 요청(반려) 또는 승인 처리

2.4.2 어노테이션/라벨링 기준

- (1) 개체명/ 신조어 인식 클래스
- 개체명 태그 6종은 TTA 표준인 개체명 태그 세트 및 태깅 말뭉치를 따른다.
 - 그 외 1종은 신조어로 표준국어대사전과 우리말샘에 없는 단어를 대상으로 한다.

분류(태그)	정의
ORGANIZATION (OG)	기관 및 조직
ARTIFACTS (AF)	인공물
CIVILIZATION (CV)	문화 용어
STUDY_FIELD (FD)	학문 분야
THEORY (TR)	이론, 법칙, 원리 등의 명칭
TERM (TM)	용어
NEW_WORD (NW)	신조어

(2) 태깅 규칙

가. 태그 선정 원칙

- ① 유일 태깅 원칙
- 개체명이 무조건 1:1 조건으로 하나의 분류에 해당하는 것은 아니므로 복수의 태깅이 가능한 개체명이라면 문맥을 파악하여 하나의 태그만을 할당해야 한다.
EX . 오늘 청와대에서 발표한... → <청와대/AF/OG> (X) / <청와대/OG> (O)
- ② 문맥 우선 적용
- 개체명 중에서는 문맥에 따라 둘 이상의 의미로 해석되거나, 동음이의어가 있을 수 있으므로 문장에서 그 의미를 고려하여 태그를 할당해야 한다.
EX . 시금치 재배 산업은... → <시금치/CV> (X) / <시금치/O>

: 시금치가 음식 재료가 아닌 PLANT로 쓰이므로 CV 태그를 할당하지 않는다.

EX. 현대 자동차에서 가장 많이 팔린 차는... → 현대 <자동차/AF> (X) / <현대 자동차/OG> (O)

: 현시대의 자동차가 아니라 현대자동차 기업을 의미하므로 현대와 자동차를 구분하지 않고 전체를 태깅한다.

나. 태깅 단위 원칙

① 최장 단위 태깅

- 복수 어절로 이루어져 있으나 그 어절 전체가 하나의 대상을 지칭할 경우 그 어절의 전체를 하나로 태깅한다.

EX. 신고전학파 성장론 → <신고전학파/FD> <성장론/TR> (X) / <신고전학파 성장론/TR> (O)

EX. 안데스 문명 → <안데스 문명/CV> (O)

- 어절 단위로 나누었을 때 특정 개체를 지칭하지 못하게 되는 경우 최장 단위 태깅 원칙을 적용한다.

EX. 스타벅스 센터필드점 → <스타벅스/OG> 센터필드점 (X) / <스타벅스 센터필드점/OG> (O)

EX. 스포츠 클라이밍 → 스포츠 <클라이밍/CV> (X) / <스포츠 클라이밍/CV> (O)

EX. 대한민국 축구대표팀 → 대한민국 <축구대표팀/OG> (X) / <대한민국 축구대표팀/OG> (O)

EX. 미얀마 대사관 → 미얀마 <대사관/OG> (X) / <미얀마 대사관/OG> (O)

② 최소 단위 태깅

- 어절 단위로 나누었을 때 전체 어절 구의 의미가 유지되는 경우에는 나누어 태깅한다.

- 지역명에 위치한 건축물, 조직은 건축물과 조직만 따로 태깅한다.

EX. 여의도 63스퀘어 → <여의도 63스퀘어/AF> (X) / 여의도 <63스퀘어/AF>

- 특정 행사나 상이 날짜와 횟수와 함께 쓰일 경우 구분하여 태깅한다.

EX. 제 34회 납세자의 날 대통령상 → 제 34회 납세자의 날 <대통령상/CV> (O)

- 각각 다른 개체명일 때 동일한 태그를 사용하더라도 각각 태깅한다.

EX. 척추뼈 디스크 → <척추뼈 디스크/TM> (X) / <척추뼈/TM> <디스크/TM> (O)

③ 그 외 복합어·합성어·파생어 등 태깅

- 접미사 '-들'은 선행 요소의 의미에 영향을 주지 않으므로 '-들'을 제외하고 태깅한다.

EX. 학생들 → <학생/CV>들

- 접미사 '-남'이 결합된 단어 전체가 [우리말샘]에 등재된 경우 전체를 태깅하나, 그렇지 않은 경우는 최장 단위 태깅 원칙 혹은 최소 단위 태깅 원칙을 따른다.

EX. 선생님 → <선생님/CV>

EX. 사장님 → <사장/CV>님

- 원문의 띄어쓰기와 관계없이 [우리말샘]에 하나의 단어로 등재되었거나 '^', '-' 기호로 연결되어 있는 경우 전체를 통합하여 태깅한다.

※ [우리말샘] 붙여쓰기 허용 기호(^)는 좌우의 단위를 서로 띄어 쓰는 것이 원칙이되 붙여 쓸 수 있는 표제어에 쓰이며, 전문어와 고유 명사의 띄어쓰기를 표기하는 데 쓰임)

※ [우리말샘] 불임표(-)는 독립된 단어 표제어의 직접 구성 성분을 분석한 결과를 표시하며 한 표제어에 한 번만 씌. 둘 이상의 구성 성분으로 이루어진 표제어의 경우에도 최종 분석 결과만 보여 줌. 불임표(-)가 쓰인 표제어는 띄어 쓰는 것이 허용되지 않으며, 가운뎃점(•) 이외의 다른 기호와는 함께 쓰이지 않음.

다. 태깅 유의사항

① 괄호 안의 원어 정보

- 괄호가 개체명 사이에 위치하여 괄호의 앞뒤를 합쳐야 완전한 개체명이 되는 경우 그 괄호까지 포함하여 태깅한다.

EX. 진(Jin)애어 → <진(Jin)애어/OG>

- 괄호가 개체명 끝에 위치하여 괄호 정보를 제외하여도 개체명 인식에 문제가 없을 경우 그괄호 정보를 제외하고 태깅한다.

EX. 청운교(橋) → <청운교/AF>

EX. 세종로(路) → <세종로/AF>

- 괄호 안의 정보가 개체명의 줄임말일 경우 각각 태깅한다.

EX. 방송통신위원회(방송위) → <방송통신위원회/OG>(<방송위/OG>)

EX. 한국방송통신대학교(이하 방송대) → <한국방송통신대학교/OG>(<이하 방송대/OG>)

- 괄호 안의 정보가 개체명의 원어 병기이거나 전체일 경우 각각 태깅한다.

EX. 대동여지도(大東輿地圖) → <대동여지도/AF>(<大東輿地圖/AF>)

EX. 엑소브레인(Exobrain) → <엑소브레인/TM>(<Exobrain/TM>)

② 괄호 안의 부가 정보

- 괄호 안의 부가 정보가 개체명일 경우 괄호 안의 정보도 각각 태깅한다.

EX. 원근법(대기 원근법, 선 원근법) → <원근법/TR>(<대기 원근법/TR>, <선 원근법/TR>)

③ 영어나 한자 등의 외국어

- 문장 안에서 원어로만 쓰인 기관명, 상품명, 작품명이 있을 경우 태깅한다.

- 단, 문장 혹은 원시 데이터 전체가 외국어로 쓰인 경우에는 태깅하지 않는다.

- 단, 외국어의 경우 고유명만 태깅하고, 보통명사는 태깅하지 않는다.

EX. Google에 입사한 이래로 → <Google/OG>에 입사한 이래로

EX. Galaxy Z 플립3 사전예약 → <Galaxy Z 플립3/AF> 사전예약

④ 줄임말

- 카카오톡, 스타벅스와 같이 알려진 줄임말의 경우 그 축약된 형태 그대로 태깅한다.

EX. 내 카톡 봤어? → 내 <카톡/TM> 봤어?

EX. 점심 먹고 스벅 가자 → 점심 먹고 <스벅/OG> 가자

⑤ 개체명 + 조사/어미 형태의 준말

- 개체명에 조사가 준말이 되어 하나의 음절로 구현된 경우 그 형태를 수정하여 형태소가 제거된 개체명만을 태깅한다.

EX. 지금 훔쳐 마시고 있어 → 지금 <훔쳐/CV>를 마시고 있어

EX. 비행기데 → <비행기/AF>인데

⑥ 오자 혹은 오타

- 원시 데이터에서 오타로 판단되는 경우, 라벨링 대상에 포함되면 오타 그대로 태깅한다.

EX. 하양책 → <하양책/TM>

EX. 학교 연양사 → 학교 <연양사/CV>

⑦ 기호

- 개체명에 가운뎃점이나 기타 기호가 포함되어 있는 경우 그 가운뎃점이나 기호를 함께 태깅

한다.

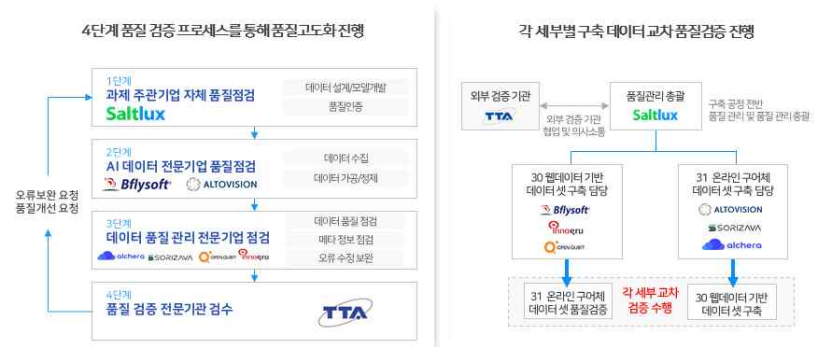
EX. 이-통장 → <이-통장/CV>

EX. 장-차관 → <장-차관/CV>

2.5 검수

2.5.1 검수 절차

- 각 세부별 구축 참여 기관이 데이터 교차 검수를 진행하여 품질 고도화
- 데이터 검수 과정에서 클라우드 워커 및 관리자가 비식별화 및 불용어 등을 재검증



2.5.2 검수 기준

- 비식별화 : 정제 및 라벨링 단계에서 비식별화가 되지 않은 개인정보를 검수자가 최종 확인
- 불용어 : 정제 및 라벨링 단계에서 정제되지 않은 비속어, 혐오표현 등의 불용어를 검수자가 최종 확인

2.6 활용

2.6.1 활용 모델

2.6.1.1 모델 학습

- 언어모델 Pretraining

언어모델 ELECTRA는 BERT에서 입력 문장의 일부분의 단어를 예측하고 맞추는 학습 목표인 MLM(Masked Language Model)을 개선하였습니다. MLM이 입력 문장의 일부만 사용하는 한계점을 해결하기 위해서, ELECTRA는 입력 문장의 모든 토큰을 활용하여 학습하는 방식을 사용합니다. 이를 위해, BERT와 똑같이 MLM을 수행한 후에, MLM 모델이 생성한 단어가 제대로 추론하였는지 여부를 판별하는 학습 목표 RTD(Replaced Token Detection)을 학습합니다. 이를 통해서 ELECTRA는 BERT보다 더 빠른 시간을 사용하여 학습할 수 있으며, 자연어 이해의 다양한 문제에 대하여 모두 BERT를 뛰어넘는 성능을 보였습니다.

ELECTRA는 BERT와 같이 Transformers 네트워크 기반으로 구성되어 있습니다. Transformers는 특정 단

어에 더 주목하게 만들도록 구성된 인공신경망 네트워크로서 attention mechanism을 통해서 구성되어 있습니다. ELECTRA는 MLM 학습 목표를 수행하는 Transformers 네트워크의 Generator를 통합합니다. 이를 통해 Generator는 양방향의 문맥을 동시에 학습합니다. 이후 통과하는 Transformers 네트워크의 Discriminator는 RTD 학습 목표를 수행합니다. Discriminator는 Generator의 결과를 활용하여 학습을 수행하며 다양한 유형의 입력 문장의 양방향 문맥을 학습하는 동시에 모든 부분에 대하여 피드백을 받아 더 효율적으로 학습을 수행합니다. 모델 구조에서 Generator는 Discriminator보다 작은 규모의 네트워크를 사용하기에 학습 속도에 큰 영향을 미치지 않습니다.

ELECTRA를 학습하기 위한 데이터셋을 구성하기 위해서, 말뭉치 구축 사업의 텍스트 데이터를 약 30% 이상을 사용하여 언어모델을 학습 하기 위한 데이터 형태로 생성하였습니다.

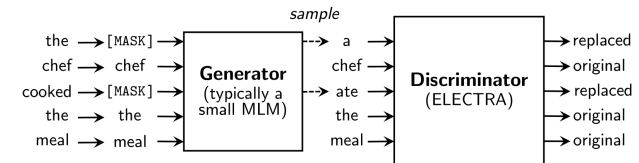


그림 20 ELECTRA 언어모델 학습 방식

- 언어모델 Finetuning

Pretrained ELECTRA은 기 학습된 언어모델로서 기계 독해, 문장 분류, 감성 분석, 개체명 인식, 자연어 추론 등과 같이 자연어 이해에 대한 다양한 문제의 학습 데이터를 이용하여 파인튜닝을 수행할 수 있습니다. 영어 자연어 이해 벤치마크 데이터셋을 이용하여 파인튜닝을 수행하는 ELECTRA의 학습 구조는 아래의 그림과 같습니다.

언어모델을 파인튜닝하기 위한 데이터셋을 구성할 때의 비율은 일반적으로 학습:검증:평가=8:1:1의 비율을 사용합니다. 언어모델을 평가하기 위한 데이터셋인 KLUE Benchmark와 Korquad 데이터를 보면 이미 학습 데이터셋이 분리되어 잘 구축되어 있기 때문에, 이를 사용하는 것으로 합니다.

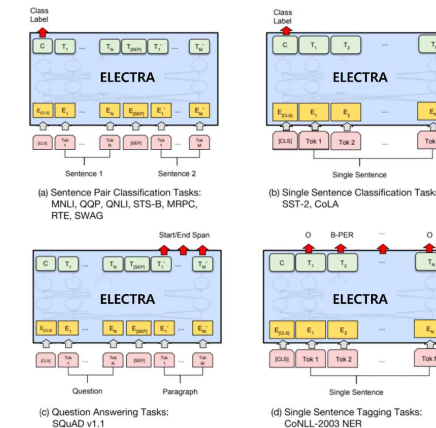


그림 21 언어모델 파인튜닝 학습 방식

2.6.1.2 서비스 활용 시나리오

학습을 완료한 모델은 기계 독해, 문장 분류, 감성 분석, 개체명 인식, 자연어 추론 등의 다양한 자연어 이해 문제를 해결할 수 있습니다. 이에 따라서 텍스트로 구성된 데이터가 주어진다면, 텍스트를 이용한 다양한 종류의 자연어 처리 서비스를 만들 수 있습니다.

- KLUE-NLI를 학습한 모델 서비스

KLUE Benchmark의 자연어 추론을 이용하여 학습시키기 위해서 공개된 데이터셋은 아래 표와 같습니다. 학습 데이터와 평가 고른 분포를 가지도록 생성되었습니다. KLUE benchmark 데이터셋이 어떻게 만들어졌는지 더 자세하게 알고 싶다면, “KLUE: Korean Language Understanding Evaluation” 논문을 참고하시면 됩니다.

Source	Train	Dev
Wikitree	3838	450
Policy	3833	450
wikinews	3824	450
Wikipedia	3780	450
Nsmc	4899	600
Airbnb	4824	600
Overall	24998	3000

KLUE-NLI 데이터를 이용하여 모델을 생성하기 위해서 특별한 전처리 과정은 진행하지 않았습니다. 모델은 “premise”와 “hypothesis”를 입력받아 “gold_label”을 예측하기 위해서 학습됩니다. KLUE-NLI 학습 모델에 들어가기 직전에 입력 데이터는 언어모델 토큰화 과정에 의해서 언어모델에 따라서 다른 입력 데이터 처리가 일어날 수 있습니다.

```
{
  "guid": "klue-nli-v1_dev_00000",
  "source": "airbnb",
  "premise": "출연자들은 발코니가 있는 방이면 발코니에서 휴식이 가능합니다.",
  "hypothesis": "어떤 방에서도 휴식은 금지됩니다.",
  "gold_label": "contradiction",
  "author": "contradiction",
  "label1": "contradiction",
  "label2": "contradiction",
  "label3": "contradiction",
  "label4": "contradiction",
  "label5": "contradiction"
},
```

KLUE-NLI 데이터를 학습하기 위한 언어모델의 아키텍처는 아래 그림과 같습니다. 이 모델을 학습하기 위해 수행되는 처리과정은 데이터 전처리, 모델 학습, 모델 평가로 3가지로 볼 수 있습니다. 데이터 전처리는 언어모델을 사용하기 위해서 필요로 하는 토큰화 과정을 제외하면 사용되지 않습니다. 토큰화 과정은 언어모델에 포함된 어휘 사전에 의해서 입력 문장을 토큰 단위의 시퀀스 데이터로 재구성하는 것을 의미합니다.

KLUE-NLI를 학습을 진행한 결과, 1세부 모델 기준 f1-score 75.98의 평가 결과가 나오는 것을 확인했습니다. 자연어 추론이 가능한 서비스를 만들 수 있는 것을 확인하였습니다.

- KLUE-STs를 학습한 모델 서비스

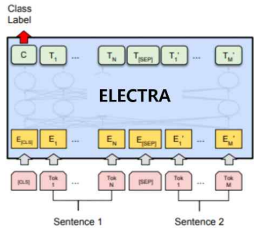
KLUE Benchmark의 텍스트 유사성을 학습시키기 위해서 공개된 데이터셋은 아래 표와 같습니다. 학습 데이터와 평가 고른 분포를 가지도록 생성되었습니다. KLUE benchmark 데이터셋이 어떻게 만들어졌는지 더 자세하게 알고 싶다면, “KLUE: Korean Language Understanding Evaluation” 논문을 참고하시면 됩니다.

Source	Train	Dev
Airbnb	5371	255
Policy	2344	132
Parakqc	3953	132
Overall	11668	519

KLUE-STs 데이터를 이용하여 모델을 생성하기 위해서 특별한 전처리 과정은 진행하지 않았습니다. 모델은 “sentence1”와 “sentence2”를 입력받아 “binary-label”을 예측하기 위해서 학습됩니다. KLUE-STs 학습 모델에 들어가기 직전에 입력 데이터는 언어모델 토큰화 과정에 의해서 언어모델에 따라서 다른 입력 데이터 처리가 일어날 수 있습니다.

```
{
  "guid": "klue-sts-v1_dev_00000",
  "source": "airbnb-rtt",
  "sentence1": "무엇보다도 호스트분들이 너무 친절하셨습니다.",
  "sentence2": "무엇보다도, 호스트들은 매우 친절했습니다.",
  "labels": {
    "label": 4.9,
    "real-label": 4.857142857142857,
    "binary-label": 1
  },
  "annotations": {
    "agreement": "0:0:0:0:1:6",
    "annotators": [
      "17",
      "07",
      "10",
      "12",
      "19",
      "14",
      "15"
    ],
    "annotations": [
      5,
      5,
      5,
      5,
      4,
      5,
      5
    ]
  }
},
```

KLUE-STs 데이터를 학습하기 위한 언어모델의 아키텍처는 아래 그림과 같습니다. 이 모델을 학습하기 위해 수행되는 처리과정은 데이터 전처리, 모델 학습, 모델 평가로 3가지로 볼 수 있습니다. 데이터 전처리는 언어모델을 사용하기 위해서 필요로 하는 토큰화 과정을 제외하면 사용되지 않습니다. 토큰화 과정은 언어모델에 포함된 어휘 사전에 의해서 입력 문장을 토큰 단위의 시퀀스 데이터로 재구성하는 것을 의미합니다.



KLUE-STs를 학습을 진행한 결과, 1세부 모델 기준 f1-score 81.17의 평가 결과가 나오는 것을 확인했습니다. 의미역 텍스트 유사성 결과를 활용 가능한 서비스를 만들 수 있는 것을 확인하였습니다.

- KLUE-RE를 학습한 모델 서비스

KLUE Benchmark의 관계 추출을 학습시키기 위해서 공개된 데이터셋은 아래 표와 같습니다. 학습 데이터와 평가 고른 분포를 가지도록 생성되었습니다. KLUE benchmark 데이터셋이 어떻게 만들어졌는지 더 자세하게 알고 싶다면, “KLUE: Korean Language Understanding Evaluation” 논문을 참고하시면 됩니다.

Source	Train	Dev
Wikipedia	21620	3621
Wikitree	10672	4088
Policy	178	56
Overall	32470	7765

KLUE-RE 데이터를 이용하여 모델을 생성하기 위해서 특별한 전처리 과정은 진행하지 않았습니다. 모델은 “sentence”를 입력받아 “label”을 예측하기 위해서 학습됩니다. KLUE-RE 학습 모델에 들어가기 직전에 입력 데이터는 언어모델 토큰화 과정에 의해서 언어모델에 따라서 다른 입력 데이터 처리가 일어날 수 있습니다.

```
{
  "guid": "klue-re-v1_dev_00000",
  "sentence": "20대 남성 A(20)씨가 아버지 치료비를 위해 B(30)씨가 모야돈 돈을 훔쳐 인터넷 방송 B에게 '별장선'으로 쓴 사실이 알려졌다.",
  "subject_entity": {
    "word": "A",
    "start_idx": 7,
    "end_idx": 9,
    "type": "PER"
  },
  "object_entity": {
    "word": "B",
    "start_idx": 29,
    "end_idx": 30,
    "type": "ORG"
  },
  "label": "no_relation",
  "source": "wikitree"
},
```

KLUE-RE 데이터를 학습하기 위한 언어모델의 아키텍처는 아래 그림과 같습니다. 이 모델을 학습하기 위해 수행되는 처리과정은 데이터 전처리, 모델 학습, 모델 평가로 3가지로 볼 수 있습니다. 데이터 전처리는 언어모델을 사용하기 위해서 필요로 하는 토큰화 과정을 제외하면 사용되지 않습니다. 토큰화 과정은 언어모델에 포함된 어휘 사전에 의해서 입력 문장을 토큰 단위의 시퀀스 데이터로 재구성하는 것을 의미합니다.

KLUE-RE를 학습을 진행한 결과, 1세부 모델 기준 f1-score 62.58의 평가 결과가 나오는 것을 확인했습니다. 관계 추출 결과를 활용 가능한 서비스를 만들 수 있는 것을 확인하였습니다.

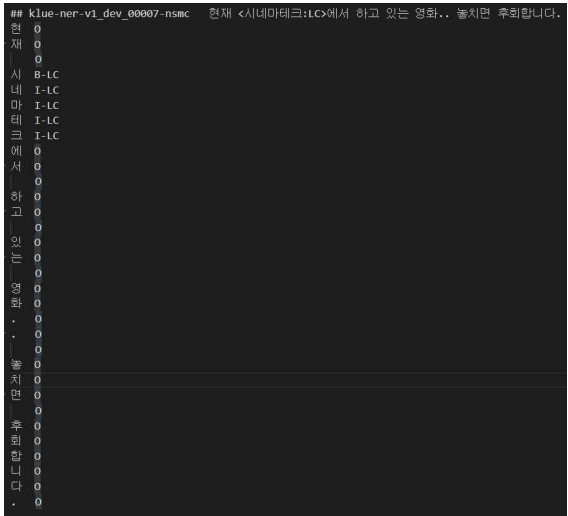
- KLUE-NER를 학습한 모델 서비스

KLUE Benchmark의 개체명 인식을 학습시키기 위해서 공개된 데이터셋은 아래 표와 같습니다. 학습 데이터와 평가 고른 분포를 가지도록 생성되었습니다. KLUE benchmark 데이터셋이 어떻게 만들

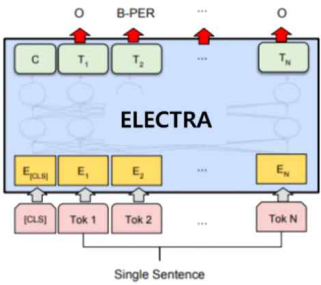
어졌는지 더 자세하게 알고 싶다면, “KLUE: Korean Language Understanding Evaluation” 논문을 참고하시면 됩니다.

Source	Train	Dev
Wikitree	11435	2534
Nsmc	9573	2466
Overall	21008	5000

KLUE-NER 데이터를 이용하여 모델을 생성하기 위해서 특별한 전처리 과정은 진행하지 않았습니다. 모델은 문장을 입력받아 토큰별 개체명 태그를 예측하기 위해서 학습됩니다. KLUE-NER 학습 모델에 들어가기 직전에 입력 데이터는 언어모델 토큰화 과정에 의해서 언어모델에 따라서 다른 입력 데이터 처리가 일어날 수 있습니다.



KLUE-NER 데이터를 학습하기 위한 언어모델의 아키텍처는 아래 그림과 같습니다. 이 모델을 학습하기 위해 수행되는 처리과정은 데이터 전처리, 모델 학습, 모델 평가로 3가지로 볼 수 있습니다. 데이터 전처리는 언어모델을 사용하기 위해서 필요로 하는 토큰화 과정을 제외하면 사용되지 않습니다. 토큰화 과정은 언어모델에 포함된 어휘 사전에 의해서 입력 문장을 토큰 단위의 시퀀스 데이터로 재구성하는 것을 의미합니다.



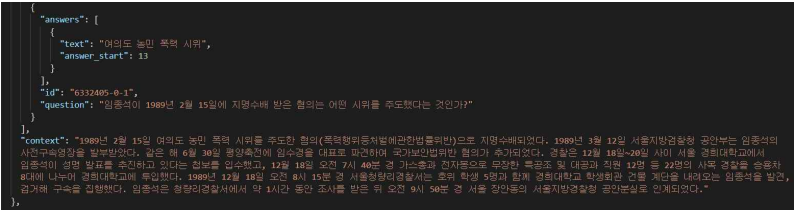
KLUE-NER를 학습을 진행한 결과, 1세부 모델 기준 f1-score 89.85의 평가 결과가 나오는 것을 확인했습니다. 개체명 인식 결과를 활용 가능한 서비스를 만들 수 있는 것을 확인하였습니다.

- KorQUAD 1.0를 학습한 모델 서비스

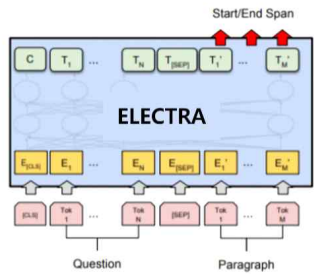
KorQUAD 1.0의 기계 독해를 학습시키기 위해서 공개된 데이터셋은 아래 표와 같습니다. 학습 데이터와 평가 데이터가 문서마다 다수의 질문을 포함하도록 생성되었습니다. KorQUAD 1.0 데이터셋이 어떻게 만들어졌는지 더 자세하게 알고 싶다면, “<https://korquad.github.io/KorQuad%201.0/>” 사이트를 참고하시면 됩니다.

	TRAIN	DEV
문서	1,420	140
질문	60,407	5,774

KorQUAD 1.0 데이터를 이용하여 모델을 생성하기 위해서 특별한 전처리 과정은 진행하지 않습니다. 모델은 “context”와 “question”을 입력받아 “answers”를 포함한 텍스트 범위를 예측하기 위해서 학습됩니다. KorQUAD 1.0 학습 모델에 들어가기 직전에 입력 데이터는 언어모델 토큰화 과정에 의해서 언어모델에 따라서 다른 입력 데이터 처리가 일어날 수 있습니다.



KorQUAD 1.0 데이터를 학습하기 위한 언어모델의 아키텍처는 아래 그림과 같습니다. 이 모델을 학습하기 위해 수행되는 처리과정은 데이터 전처리, 모델 학습, 모델 평가로 3가지로 볼 수 있습니다. 데이터 전처리는 언어모델을 사용하기 위해서 필요로 하는 토큰화 과정을 제외하면 사용되지 않았습니다. 토큰화 과정은 언어모델에 포함된 어휘 사전에 의해서 입력 문장을 토큰 단위의 시퀀스 데이터로 재구성하는 것을 의미합니다.



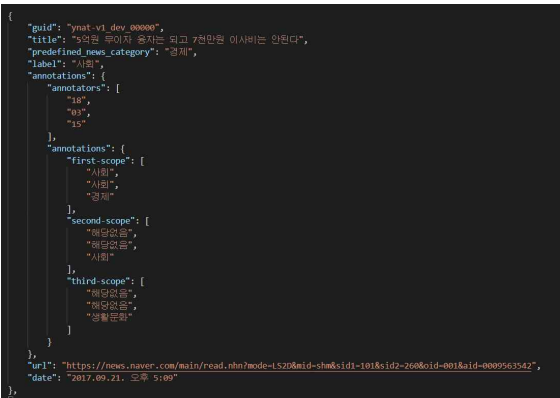
KorQUAD 1.0를 학습을 진행한 결과, 1세부 모델 기준 f1-score 89.72의 평가 결과가 나오는 것을 확인했습니다. 기계 독해 결과를 활용 가능한 서비스를 만들 수 있는 것을 확인하였습니다.

- KLUE-TC를 학습한 모델 서비스

KLUE Benchmark의 주제 분류를 학습시키기 위해서 공개된 데이터셋은 아래 표와 같습니다. 학습 데이터와 평가 고른 분포를 가지도록 생성되었습니다. KLUE benchmark 데이터셋이 어떻게 만들어졌는지 더 자세하게 알고 싶다면, “KLUE: Korean Language Understanding Evaluation” 논문을 참고하시면 됩니다.

Topic	Train	Dev
Politics	7379	750
Economy	6118	1268
Society	5133	3740
Culture	5731	1387
World	8320	776
IT/Science	5235	587
Sport	7742	599
Total	45678	9107

KLUE-TC 데이터를 이용하여 모델을 생성하기 위해서 특별한 전처리 과정은 진행하지 않았습니다. 모델은 “title” 를 입력받아 “label” 을 예측하기 위해서 학습됩니다. KLUE-TC 학습 모델에 들어가기 직전에 입력 데이터는 언어모델 토큰화 과정에 의해서 언어모델에 따라서 다른 입력 데이터 처리가 일어날 수 있습니다.



KLUE-TC 데이터를 학습하기 위한 언어모델의 아키텍처는 아래 그림과 같습니다. 이 모델을 학습하기 위해 수행되는 처리과정은 데이터 전처리, 모델 학습, 모델 평가로 3가지로 볼 수 있습니다. 데이터 전처리는 언어모델을 사용하기 위해서 필요로 하는 토큰화 과정을 제외하면 사용되지 않았습니다. 토큰화 과정은 언어모델에 포함된 어휘 사전에 의해서 입력 문장을 토큰 단위의 시퀀스 데이터로 재구성하는 것을 의미합니다.

KLUE-TC를 학습을 진행한 결과, 1세부 모델 기준 f1-score 82.45의 평가 결과가 나오는 것을 확인했습니다. 주제 분류 결과를 활용 가능한 서비스를 만들 수 있는 것을 확인하였습니다.

2.6.2 데이터 제공

- 1) AI Hub(www.aihub.or.kr) 활용
 - 구축된 데이터를 제공하는 AI Hub 페이지에서 대규모 한국어 말뭉치 AI 학습용 데이터의 필요성과 구축 내용, 데이터셋 구조, 예시 등에 대해 자세한 내용을 제공하여 누구나 쉽게 데이터를 활용할 수 있는 환경 마련
 - 해당 데이터를 이용한 국민, 기관, 기업 등의 피드백을 수렴하고 결과를 분석하여 활용 방향 제고
- 2) 솔트룩스 AI Cloud(saltlux.ai) 활용
 - 솔트룩스의 3세대 AI 클라우드 서비스(saltlux.ai)는 현재 43개의 AI 모델을 사용할 수 있는 서비스를 제공하고 있음
 - 대규모 한국어 말뭉치 AI 학습데이터의 활용한 AI 모델은 솔트룩스의 AI 클라우드 서비스 자원으로 무상으로 3년간 공개를 지원