인공지능 데이터 구축·활용 가이드라인

- 고객 주문 질의-응답 데이터 -

인공지능 데이터 구축	사업 총괄	롯데정보통신
	데이터 설계	롯데정보통신
	원천데이터 수집 및 정제	K-WON 롯데정보통신
	데이터 가공	AIMMO
	데이터 검수	AIMMO 롯데정보통신
	크라우드 소싱	ELGEN AIMMO
	저작도구 개발	ELGEN AIMMO
	AI모델 개발	ELGEN 롯데정보통신
	응용 서비스 개발	롯데정보통신
가이드라인 작성	롯데정보통신㈜	전시형 수석, 최봉우 책임
	㈜에이모	강상현 PM
	㈜케이원정보통신	이종문 부장
	㈜엘젠아이씨티	김일환 이사
가이드라인 버전	버전 1.6, 2021년 2월 26일	

목 차

1. 데이터 명세 정보

1.1 데이터 정보 요약

데이터 이름	소상공인 고객 주문 질의-응답 데이터	
활용 분야	유통, 음식점 등 다양한 분야의 무인매장, 비대면 업무 등의 환경에서 활용	
데이터 요약	1. 음식점, 슈퍼, 농수산물 등 다양한 상점에서 발화한 질의응답데이터를 포함 2. 상품, 결제, 반품, 배송 등 소상공인이 사용할 수 있는 업무 전반의 주제를 포함 3. 질의응답셋에 감정, 대화의도, 개체명 등의 정보를 추가로 표시함 4. 각 질의응답의 데이터는 단순 질의-응답으로 끝날 수도 있으며 계속 질문이 이어지는 경우 순서를 표시하여 대화의 흐름을 알 수 있게 하였음 (고객 질의응답 데이터는 상담 대화를 근간으로 생성했으나, 상담사 인사로 시작하여 본인인증 관련 대화, 연락처 및 주소 등 개인 및 민감 정보를 포함한 대화, 마무리 인사하는 대화를 제외하면 실제 질문은 한두 개 정도로서 맥락이 이어지는 문답은 거의 없음. 이어 지는 문답 별로 인텐트 대분류가 달라지는 실태를 감안하여 NIA 및 TTA와 '자연스러운 대화 흐름' 여부는 검증 범위에서 제외하기로 합의했음.)	
데이터 출처	주제분야: 상품, 결제, 반품, 배송 등 소상공인이 사용할 수 있는 업무 전반 출처 정보: 인터뷰를 통한 직접 녹취, 데이터 직접 생성, 음성데이터 기반 데이터 추출	
데이터 이력	배포버전	1.6
	개정이력	2021.02.26 TTA 요청사항 반영
	작성자/ 배포자	전시형 수석 / 최봉우 책임

1.2 데이터 포맷

가) 기본 포맷

- 미리 정의된 데이터 항목을 모두 포함하는 CSV 파일
- 질의응답 데이터 500만건 목표

나) 데이터 예시

QA번호	화자	QA여부	문장	감성	카테고리	의도	개체명	상담번호	상담내순번
1	c	Q	자장면 얼마에요?	m	중화요리 음식점	가격문의		1	1
1	s	A	칠천원 입니다.	m	중화요리 음식점	가격문의		1	2
2	c	Q	가산에 엘디씨씨 배달 되나요?	m	중화요리 음식점	배달지역문의	가산: place 엘디씨씨: org	1	3
2	s	A	네.	m	중화요리 음식점	배달지역문의		1	4
3	s	Q	몇시까지 갖다드릴까요?	m	중화요리 음식점	배달기한문의		1	5
3	c	A	다섯시까지 갖다주세요.	m	중화요리 음식점	배달기한문의	다섯시: time	1	6
3	s	A	네 갖다드릴게요.	m	중화요리 음식점	배달기한문의		1	7

1.3 어노테이션 포맷

항목	설명	타입	필수구분
IDX	질의응답 데이터 파일 내 고유 순서 번호	Num.	Y
발화자	발화자 정보 (c: 고객 s: 점원)	string	Y
발화문	대화 텍스트 정보	string	Y
카테고리	발화가 일어나는 상점 정보	string	Y
QA번호	질의응답셋을 구분하는 정보	Num.	Y
QA여부	질의문(q)인지 응답문(a)인지 표시	string	Y
감성	텍스트별 감성 정보 (m: 중립, n:부정, p:긍정)	string	Y
인텐트	질의문 기준 발화문에 내재한 의도	string	Y
개체명	NER(개체명인식)을 위한 개체 정보	string	N
상담번호	대화 상황 구분 정보	Num.	Y
상담내순번	상담 내 발화 순서 표시	Num.	Y

1.4 데이터 구성

루트 폴더	00. 초기 데이터	라벨링데이터	(상점)카테고리: 5종	284,494건	
	01. 데이터	Training	라벨링데이터	(상점)카테고리	
		Validation	라벨링데이터	(상점)카테고리	
		Test	라벨링데이터	(상점)카테고리	
		Sample	라벨링데이터	(상점)카테고리	
		구축 가이드			
	02. 저작도구				
	03. AI모델				
	04. 활용 서비스				
	~~05. 원시 데이터~~	(해당사항 없음. 개인정보 비식별화 이전의 고객 상담/대화 데이터는 공개 불가능하며 삭제해야 함.)			

1.5 데이터 통계

1.5.1 데이터 구축 규모

데이터 출처	질의응답데이터 규모 (최종 산출물 기준)	도메인
콜센터 데이터	400 만 건 질의응답	백화점, 홈쇼핑, e-commerce 등 유통 관련
녹취 데이터	100 만 건 질의응답	도.소매업, 숙박.음식업점, 수리, 기타개인서비스업, 보건업 등에 해당하는 약 20 종 상점 (한국표준산업분류 10 차 기준)

가) 콜센터 데이터 (백화점, 홈쇼핑, 수퍼, 마트, 이커머스) : 콜센터 상담 데이터를 활용한 질의/응답 데이터 구축 (400만건)

나) 소상공인 녹취 데이터 (음식점, 의료, 카페, 가전, 농산물, 수산물, 생활잡화, 정육점, 동물병원, 수퍼) : 소상공인 녹취데이터 (100만건)

다) 학습용 데이터: CSV 파일 포맷으로 생성

라) 메타 정보 데이터: CSV 파일 내에 같이 저장

1.5.2 데이터 분포 예시

- 구축 데이터 질의응답데이터: 5,000,000 건
- 데이터 출처 별 데이터(질의응답데이터) 수 (단위: 10,000건)
- 데이터 산업 구분별 데이터 비율(추정치, 한국표준산업분류 10차 기준)

데이터 출처	
콜센터	직접 녹취
400	100

콜센터별 질의응답비율(도메인)				
e-commerce	홈쇼핑	백화점	마트	그 외
50%	10%	10%	15%	15%

직접 녹취 상점 비율			
소매업	숙박 및 음식업점	제조업	보건업
30%	30%	20%	20%

- 콜센터별 질의응답 비율 선정 기준

(1) 콜센터와 협의하여 수집 가능한 음성 시간 5000시간 선정
(2) 확보 가능한 수집 원천 음성의 샘플 분석을 통해 대상 도메인과 가장 연관도 높은 순으로 수집 비율 선정
(3) 도메인 치우침 방지를 위해 나머지 원천에서도 각각 수집
- 직접 녹취 상점 비율 선정 기준
(1) 수집 가능한 상점 대상 선정 (20여종)
(2) 한국표준산업분류에 따른 업종 분류 비율을 고려하여 수집 대상 비율 조정

1.5.3 기타 활용 통계

가) 유사통계 없음

1.6 원시데이터 특성

1.6.1 대상분류

가) 콜센터 데이터 (백화점, 홈쇼핑, 수퍼, 마트, 이커머스, 건설) '실제'

나) 소상공인 녹취 데이터 (음식점, 의료, 카페, 가전, 농산물, 수산물, 생활잡화, 정육점, 동물병원, 수퍼) '실제

데이터 출처	질의응답데이터 규모 (최종 산출물 기준)	도메인	대상분류
콜센터 데이터	400 만 건 질의응답	백화점, 홈쇼핑, e-commerce 등 유통 관련	실제
녹취 데이터	100 만 건 질의응답	도.소매업, 숙박.음식업점, 수리, 기타개인서비스업, 보건업 등에 해당하는 약 20 종 상점 (한국표준산업분류 10 차 기준)	실제

1.6.2 제약조건

가) 콜센터 데이터 (백화점, 홈쇼핑, 수퍼, 마트, 이커머스, 건설) '일부 제약있음'

나) 소상공인 녹취 데이터 (음식점, 의료, 카페, 가전, 농산물, 수산물, 생활잡화, 정육점, 동물병원, 수퍼) '일부 제약있음'

데이터 출처	질의응답데이터 규모 (최종 산출물 기준)	대상분류	제약조건
콜센터 데이터	400 만 건 질의응답	실제	일부 제약있음
녹취 데이터	100 만 건 질의응답	실제	일부 제약있음

※ 일부 제약있음(semi-constrained): 녹취 데이터의 경우 점주/고객의 질의-응답을 사전 스크립트 없이 수집하나 가급적 미리 정의된 인텐트 또는 개체명이 들어갈 수 있도록 발화. 콜센터 데이터의 경우 상담원/고객의 질의-응답을 사전 스크립트 없이 수집하나 콜센터 내부 규정에 따라 발화하는 경우 있음.

1.6.3 속성

가) 콜센터 데이터 (백화점, 홈쇼핑, 수퍼, 마트, 이커머스, 건설) '400만건 대화 추출'

나) 소상공인 녹취 데이터 (음식점, 의료, 카페, 가전, 농산물, 수산물, 생활잡화, 정육점, 동물병원, 수퍼) '100만건 대화 추출

데이터 출처	원시데이터 형태	정제 데이터 형태	개인정보 비식별화 여부	화자구분
콜센터 데이터	음성 파일	텍스트 파일	Y	상담사/고객
녹취 데이터	음성 파일	텍스트 파일	Y	점주/고객

1.7 기타정보

1.7.1 포괄성

가) 콜센터 데이터의 경우, 롯데 그룹 내 그룹사 인 백화점, 홈쇼핑, 수퍼, 마트, 이커머스, 건설 등의 상담 대화로써 유통업과 관련된 범위 나타냄

나) 소상공인 녹취 데이터는 음식점, 의료, 카페, 가전, 농산물, 수산물, 생활잡화, 정육점, 동물병원, 수퍼 등의 도메인을 포함함

1.7.2 독립성

데이터 출처	원시데이터 형태	정제 데이터 형태	민감정보	잡음
콜센터 데이터	음성 파일	텍스트 파일	원시데이터에 존재할 수 있으나 비식별화	NA
녹취 데이터	음성 파일	텍스트 파일	원시데이터에 존재할 수 있으나 비식별화	원시정보에 존재할 수 있음

1.7.3 유의사항

가) 파급 효과

파급효과	설명
비대면 매장 적용 가능	현재 코로나 등 비대면 서비스 필요성 증가에 따른 소상공인의 스마트 상점화에 대응 가능

소상공인 도메인의 한국어 연구	대규모 한국어 코퍼스 구축으로 한국어 연구, 특히 자연어 연구에 사용될 수 있음
다양한 데이터 활용 범위	데이터의 수량과 품질면에서 다양한 도메인에 적용 가능한 서비스 개발이 가능
활용 서비스 개발 등 협업 가능성	유지보수 기간 문의/피드백을 통한 데이터 사용처 지원 및 협업 가능

나) 유의 사항

유의사항	설명
도메인	데이터에 포함되어 있지 않은 도메인(상점)의 경우 AI 활용에 다소 어려움 있음
개체명 정의	데이터에 정의된 개체명 항목의 정의가 활용 목적에 따라 맞지 않을 수 있음
감정 정의	긍정/부정에 대한 감정 정의가 활용 목적에 따라 상이할 수 있음

1.7.4 관련 연구

가) 관련 연구 없음

2. 데이터 구축 가이드

2.1 데이터 구축 개요

<데이터 구축 프로세스>

가) AI 학습 데이터 구축
- 질의-응답으로 구성된 텍스트 데이터 : 500만 건 이상
- 롯데정보통신 콜센터(백화점, 홈쇼핑, 이커머스 등) 데이터 활용 (5,000시간 이상)
- 케이원에서 소상공인을 대상으로 직접 녹취하여 원천 데이터 획득

나) 데이터 획득, 가공을 위한 도구 개발
- 데이터 가공을 위한 저작 도구로 STT 엔진 구축 및 활용
- 대화 내 존재하는 이름, 주소, 전화번호 등 개인정보 비식별화

다) 구축된 데이터에 기반한 AI 응용서비스 개발
- NLU 엔진 개발, 챗봇 빌더 연동 API

2.2 문제정의

2.2.1 임무 정의

가) 고객의 매장 경험 불편 최소화 및 소상공인 운영 효율화 가능한 무인점포 운영을 위한 인공지능 데이터셋과 AI서비스 구축을 목표로 함

나) 콜센터 음성파일 상담 내역에 대해 텍스트 변환과 개인정보를 검출하고 마스킹 할 수 있는 머신러닝 기술을 필요로 함

다) 콜센터 음성파일(.wav)을 음성인식 기술을 적용하여 1차로 음성을 텍스트로 변환하고, 변환된 텍스트에 대해 전사 작업을 거쳐 원천데이터를 확보함

라) 원천데이터에 대한 개인정보 검출 및 마스킹은 개인정보에 대한 기준을 정의하고, 정의한 기준에 따라 정규식 방식을 이용하여 개인정보 검출 및 마스킹 함

2.2.2 데이터 구축 유의사항

가) 개인정보, 민감정보의 경우, 1차 Rule 기반 마스킹 하며, 2차 검수로 수작업을 통해 비식별화 처리함

※ 개인정보 범위

① 고객정보: 이름, 주민번호, 전화번호, e-mail, 주소

② 상점정보: 직원명, 직원연락처, 직원e-mail

2.3 획득·정제

2.3.1 원시데이터 선정

가) 콜센터 상담 데이터

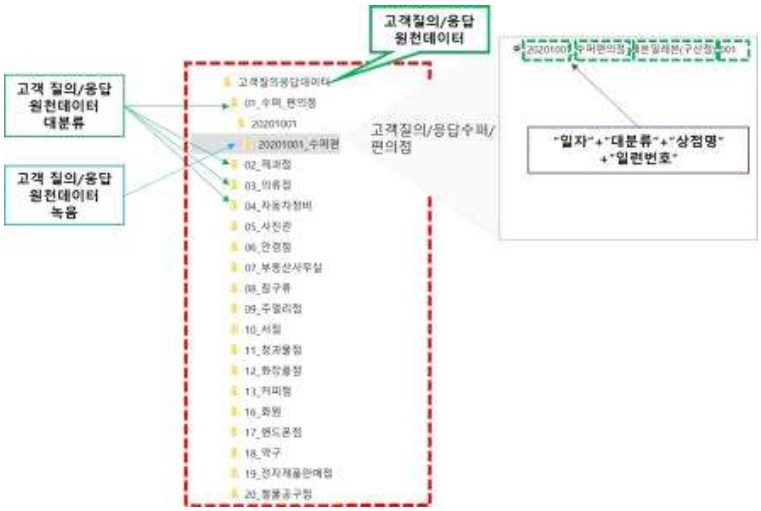
- 무인 서비스 운영 시 발생할 수 있는 다양한 유형의 질의-응답 데이터를 확보하기 위해 주관기관인 롯데정보통신㈜에서 확보하고 있는 콜센터 음성 및 텍스트 전사 데이터를 활용할 예정

나) 소상공인 녹취 데이터

- 실제 생활에 매우 밀접한 소상공인의 20개 업종에서 결제, 교환, 반품, 환불 등 약 120개 상황의 대화 직접 녹음
- 소상공인, 수집인력 직접 녹음
- 녹취 기준

파일 확장자	WAV	주파수 Frequency	16㎑
Quantization Bit	16 Bit	채널	모노
품질(Quality)	256K		

- 녹취 저장 기준

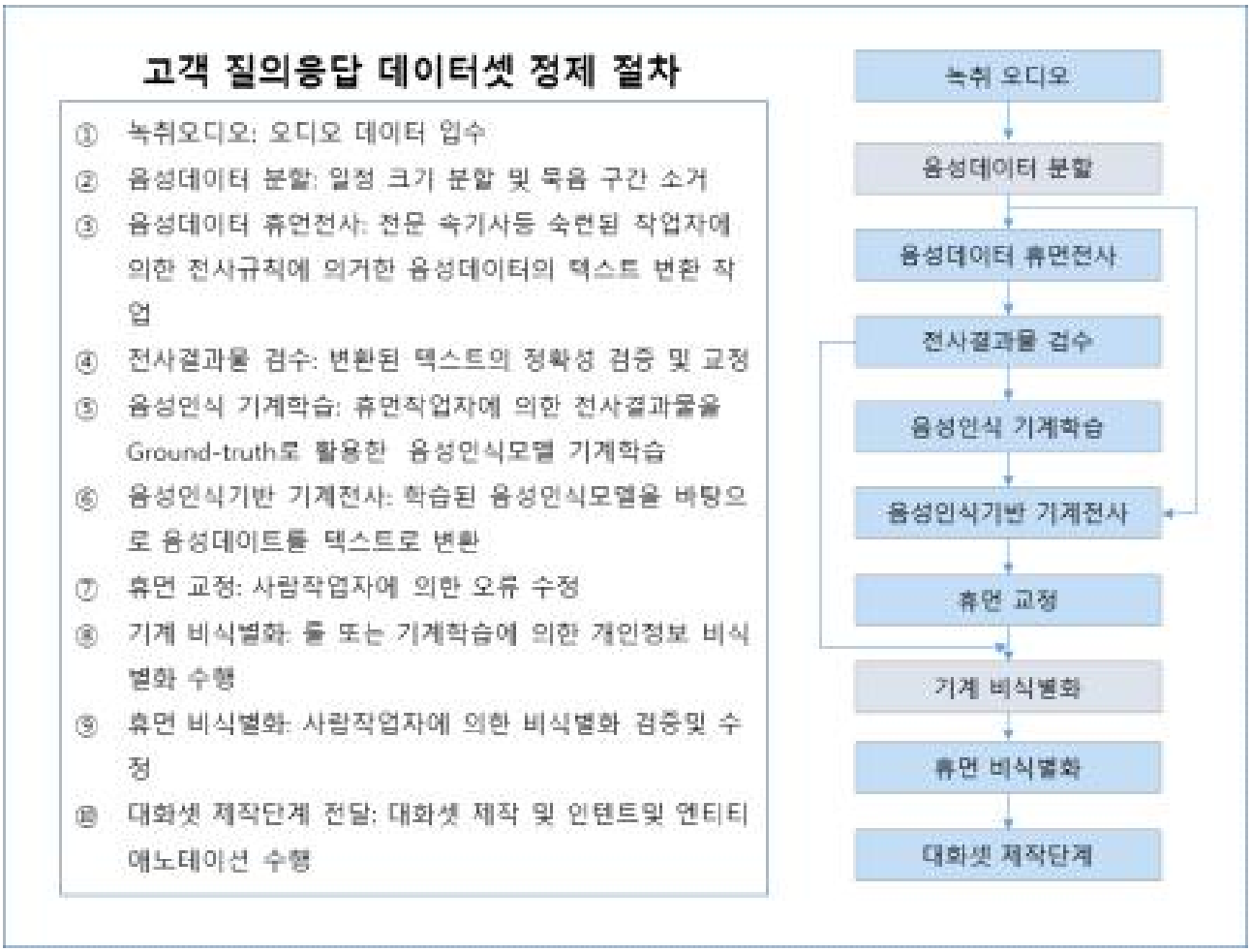


2.3.2 획득·정제 절차

가) 획득 절차

- 소상공인: 상점주로서 고객과의 결제, 교환, 반품, 활불 등 다양한 상황 대화 녹음
- 수집인력: 고객으로서 상점주와 결제, 교환, 반품, 활불 등 다양한 상황 대화 녹음
- 데이터 저장 및 검수: AI 사업 추진 센터내의 관리 인력에게 녹음 파일 전송
  관리 인력은 녹음 파일 검수 및 공유 스토리지에 원천 데이터 저장

나) 정제 절차



위의 데이터 정제과정에서 회색으로 표시된 ② 음성데이터분할과 ⑧ 기계비식별화 단계는 선택사항이다. 음성데이터 분할은 기계학습을 위한 목적인데 기계학습모델과 방법에 따라 분할이 필수적으로 요구되는 것은 아니고, 기계 비식별화는 다양하고 불규칙한 대화체 문장에서 높은 정확성을 담보하기 어렵기 때문에 비식별화 결과를 사람이 검증하고 수정하는 과정이 수반되어야 한다는 점이 선택과정에서 고려되어야 한다. 정제 과정은 크게 세가지 흐름을 내포하고 있는데 선택사항을 제외 시 각 과정의 흐름은 다음과 같다

- 음성데이터 휴먼 전사
  ① + ③ + ④ + ⑨ + ⑩
- 음성인식 기계학습
  ① + ③ + ④ + ⑤
- STT 활용 전사
  ① + ⑥ + ⑦ + ⑨ + ⑩

정제작업과정에서 시간소요가 많은 단계는 휴먼 전사 과정인데 이 과정을 수행하면서 바로 ⑨ 비식별화 마킹작업도 동시에 수행하여 단계를 축약하고 있다. 휴먼 전사 과정의 시간은 속기사와 일반 작업자의 작업속도에 있어 3~4배 정도 차이가 있어 이를 고려한 인력계획수립이 요구된다. 전사된 데이터를 바탕으로 STT 음성인식모델을 도메인 학습시켜 이를 토대로 녹취음성을 초벌 텍스트로 변환하고 이를 사람이 교정함으로써 전사에 드는 시간을 단축하는 방안을 고려할 수 있다.

2.3.3 획득·정제 기준

가) 자료 형태

- 원천 데이터 : PCM 파일 (개인정보 비식별화 이전 데이터로서 납품 대상이 아님)
- 원천 데이터인 PCM 파일을 변환한 텍스트 데이터 : TXT 파일

나) 자료의 규모

- 원천 데이터 : 소상공인 질의응답데이터 100만건, 콜센터 질의응답데이터 400만건
- 정제 데이터 : 음성이 변화된 텍스트 데이터, 질의응답데이터 500만건

다) 획득 기준(핸드폰, 핀 마이크 이용)

- 전사 가능한 범위(녹음된 내용 인식)내에서 녹음

라) 원천 데이터 품질

- 사전 지급된 핀 마이크를 핸드폰에 연결 후 대화 내용 녹음으로 데이터 품질 확보

검수 인력에 의한 녹음 파일 검수

마) 전사작업은 사전 정의된 전사규칙에 따라 수행한다. 전사규칙의 주요 내용은 다음과 같다

- 화자가 바뀔 때 행바꿈을 한다
- 화자간 말 겹침시 '/'를 사용하여 같은 행에 표기한다
- 간투어 등 대화자의 모든 말 작성
- 기타 자세한 사항은 롯데정보통신 'STT 텍스트 변환 매뉴얼' 참조 (별도 문서로 제출)

바) 기계전사용 음성인식모델의 평가는 음성인식에서 통용되는 Metric을 적용하여 성능을 수치화한다.

- CER: 문자오류율
- WER: 단어오류율

사) 비식별화는 아래의 민감정보를 대상으로 하여 수행한다.

(데이터를 제공한 롯데그룹사가 준수를 요구한 별도 문서 '개인정보 마스킹 규칙' 참조)

※ 민감정보 범위

① 고객정보: 이름, 주민번호, 전화번호, e-mail, 주소

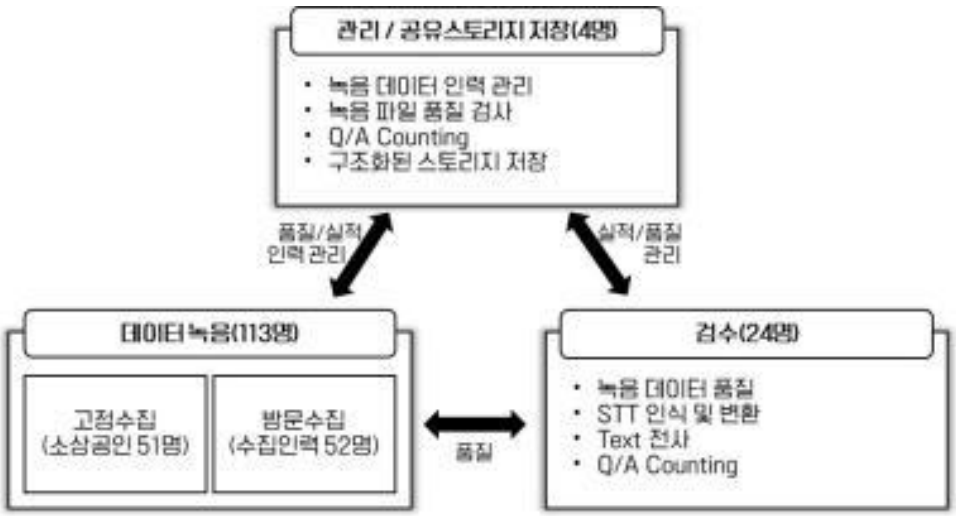
② 상점정보: 직원명, 직원연락처, 직원e-mail

아) 비속어, 사투리, 줄임말에 대한 처리

- 비속어는 정제 처리하고 그 외는 원문 그대로 처리한다.

2.3.4 획득·정제 조직

획득조직	명	설명
데이터 녹음	113	기 지급된 마이크 이용 직접 녹음
데이터 정보 입력 인력	4	공유 스토리지에 원천 데이터 저장
검수	24	저장된 녹음 데이터 검수



※ 마이크 기준

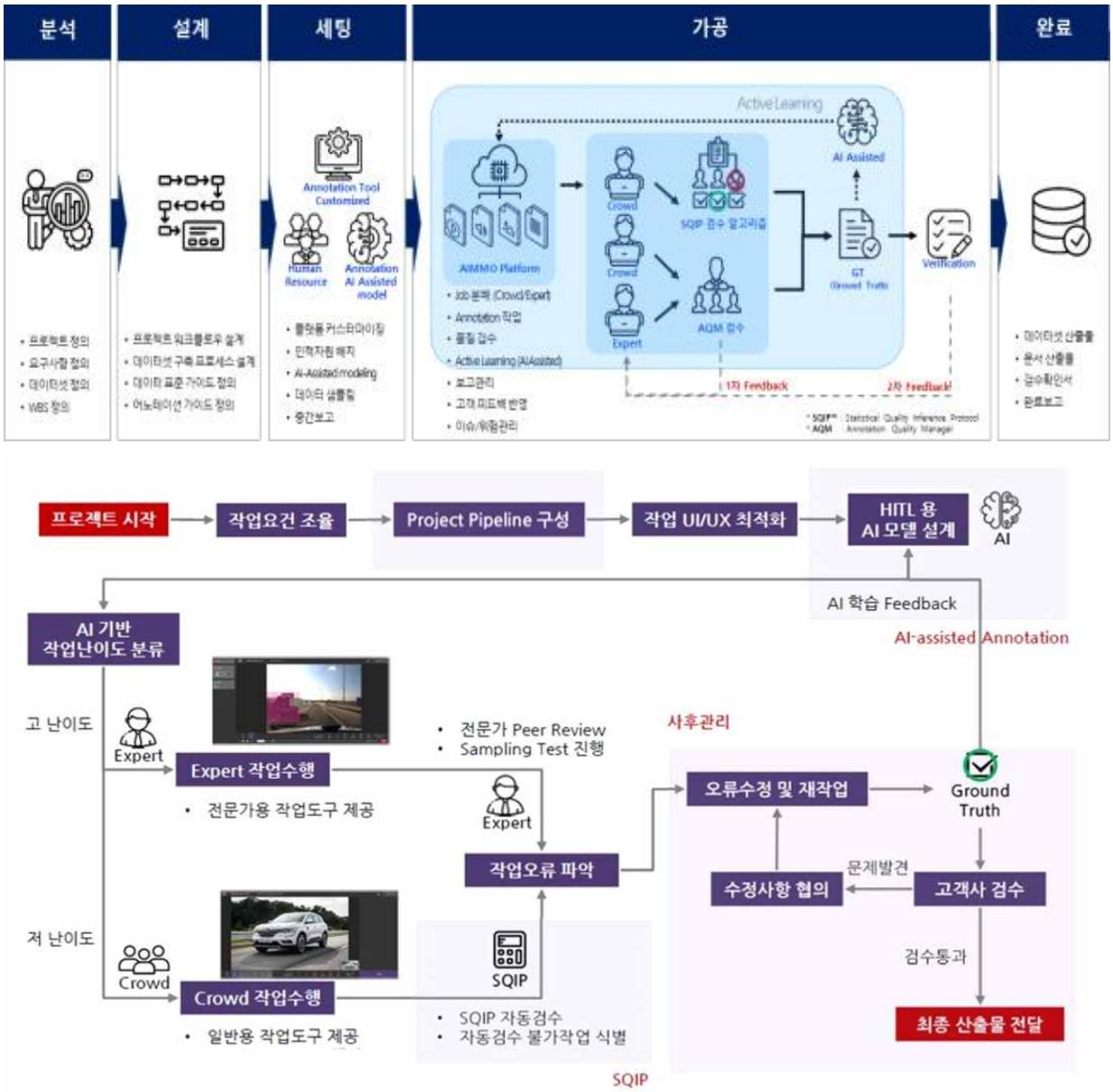
모델명	BOYA – M2	BOYA – M3
용도	아이폰용	안드로이드폰용
	핀마이크 겸용	핀마이크 겸용
지향성	전지향성	전지향성
단자	아이폰, 아이패드(라이트닝 젠더)	안드로이드(USB C 젠더)
인증	Apple MFI인증	-
Sensitive	-40+/-3dB	-40+/-3dB
Polar Pattern	76dB or more	76dB or more
Sampling Rate	24Bit/48Khz 고성능 ADC탑재	48Khz
Bit Rate	16Bit, 24Bit	16Bit
Cable Length	6m	6m
Weight	46g	44.5g

2.4 어노테이션/라벨링

2.4.1 어노테이션/라벨링 절차

가) 라벨링 절차를 '분석', '설계', '세팅', '가공', '완료' 단계로 세분화 하여 추진

<학습용 데이터 가공 프로세스>



나) 가공 플랫폼 개발 및 수정

- 해당 어노테이션 규격에 맞는 가공툴 타입으로 가공 툴 개발

다) 가공 플랫폼 세팅

- 수급 받은 데이터의 규약에 맞는 가공 데이터 설정 및 프로젝트 생성. 수집 데이터를 가공 플랫폼에 업로드.

라) 가공 샘플링

- 가공 작업 가이드라인 구축을 위한 샘플링을 내부작업자 기준으로 진행. 피드백을 취합하여 가이드라인 완성

마) 가공 진행을 위한 크라우드 소싱 인력 수급

- 플랫폼 오픈 후 크라우드 소싱 인력 투입

바) 가공 진행

사) 검수
- 일반 크라우드 작업자의 1차 검수 진행
- 내부 작업자의 2차 검수 진행
- 내부 검수 전문가의 3차 검수 진행

아) 검수 완료된 건에 대한 산출물 추출 및 포맷 변환

2.4.2 어노테이션/라벨링 기준

가) 무인매장, 비대면 업무 솔루션 등에서의 고객응대 기술 개발기관의 요구사양을 반영한 어노테이션 기준 마련

나) 관련 SW 개발업체의 요구사양에 따라 제품 및 제품의 이름과 기타정보 등에 대한 어노테이션 방법에 대한 기준 마련

다) 관련 SW 개발업체의 피드백을 통한 어노테이션 대상 기준 마련

라) 어노테이션 작업자에 의해서 판단이 어려운 객체에 대한 정보를 수집하여, 해당 내용을 가이드에서 수정 반영하는 어노테이션 기준 수

구분	설 명
음성 데이터셋	1-1) 단순 질의 응답 셋: A/B, C/D, E/F 1-2) 감성 태그: 긍정, 중립, 부정

※ 본 과제가 다양하고 세분화된 일반감정분류가 아닌 비즈니스 도메인의 상품과 서비스와 관련한 감정 분류이기에 긍정, 중립, 부정의 세 가지에 대해서만 선정 가능함. 감정 분류를 하는 목적은 어휘만으로 구분하지 못하는 상황에 대해 의도(intent)를 보다 명확하게 추출하기 위함

마) 라벨링 구성 및 기준

구분	기준
작업대상 여부	1. 문장 구성이 의미있는 QA셋에 맞는지 판단 2. Y/N 으로 QA셋 적용 여부 선택 3. N인 경우 인스턴스 추출하지 않음
Q/A 구분	1. 작업대상인 경우 Q/A 구분하여 매핑 2. 질문인 경우 Q , 답변인 경우 A 로 클래스 적용
Q/A 세트 번호	1. QA 세트에 대해 넘버링 적용을 통한 세트 구성 2. 순차적으로 넘버링 함 3. QA세트는 Q-A 구성일수도 있고, Q-Q-A, Q-Q-Q-A-A 구성 등 다양하게 구성 가능하며, 세트 번호로 QA 세트 개수를 측정함

감성	1. 긍정(P) / 부정(N) / 중립(M) 의 기준으로 적용 2. 감성 구분 - 기쁜 상태 / 즐거운 상태 / 칭찬 등이 포함된 경우 : 긍정 - 화난 상태 / 따지는 상태 / 비난 등이 포함된 경우 : 부정 - 그 외의 경우 : 중립 3. 모호한 경우 중립(M) 적용
의도(Intent)	1. 인텐트 표시는 3단계로 "대분류_소분류_발화자행동" 으로 구성 2. 인텐트 구성은 문서 말미 참고자료에 상세 기술함. 3. 각 단계별 항목 매핑
NER	1. 9가지 NER (가격(e.g. ₩, $), 수량(e.g. 개, 장), 크기(e.g. 대/중/소), 장소, 조직, 사람, 시간, 날짜, 상품명) 에 대한 클래스 적용 2. 각 속성별 원 문장의 해당하는 문구 입력 처리

바) 산출물

파일명	설 명
20200921140002_61115_REC4A.csv	원천 데이터로 수급된 전사 파일
20200921140002_61115_REC4A.csv.json	검수가 완료된 최종 가공 데이터 파일

2.4.3 어노테이션/라벨링 조직

가) 어노테이션 전담 조직을 통한 프로젝트 진척 관리 및 품질 관리 수행

획득조직	명	설명
라벨러	1000	어노테이션 저작도구인 GTaas 를 사용하여 라벨링 작업을 진행하는 인력
AQM	10	라벨러들을 관리하고 진행함에 있어서 문의사항 응대 및 1차 작업물 확인 및 관리하는 팀. 또한 최종 작업물에 대한 검수도 함께 진행
검수	1000	산출물에 대한 최종 검수 진행

나) 크라우드 작업자 난이도별(초급/중급/고급) 온/오프라인 교육 실시

- 교육 및 작업 결과에 따라 프로젝트 수행 후 검수자, 내부 관리자로 채용 기회 제공
- 라벨링 가이드 교육
  > 1차 : 동영상 교육 (https://www.youtube.com/watch?v=Pp5cENP8cpA&t=4s)
  > 2차 : 가이드 제공
  > 3차 : 화상교육 진행
  > 기타 : 내방 교육 진행 (신청자 한정)

2.4.4 어노테이션/라벨링 도구

가) 어노테이션 전문기업(에이모)이 개발한 저작도구 활용

나) (작업) 작업자 활용을 위한 편의 기능 탑재(작업이동, 화면 조절 등)

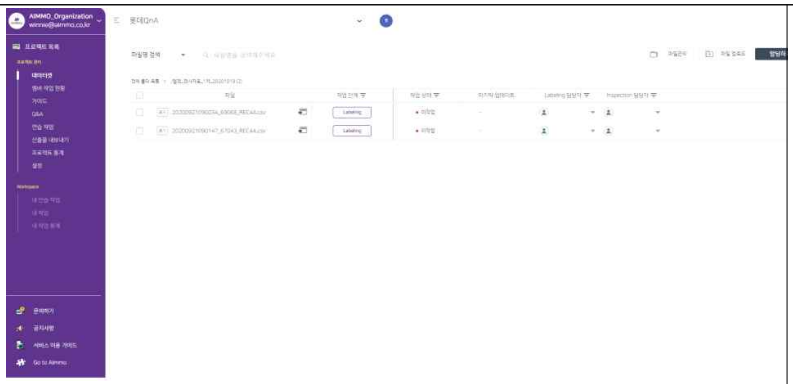
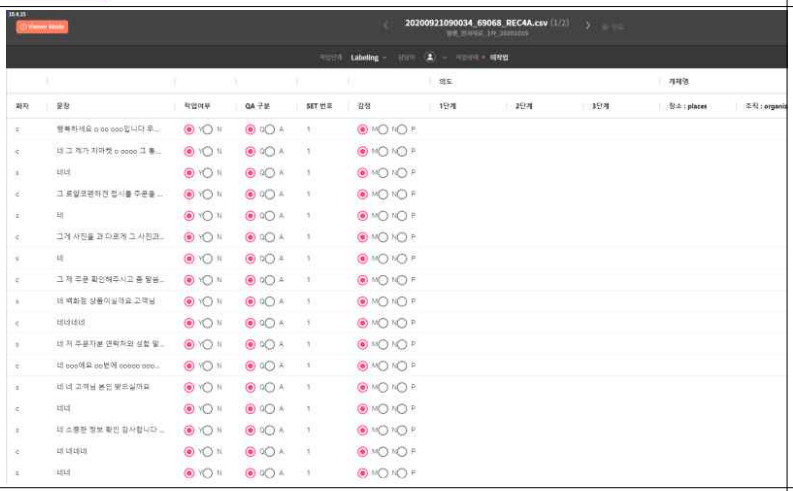
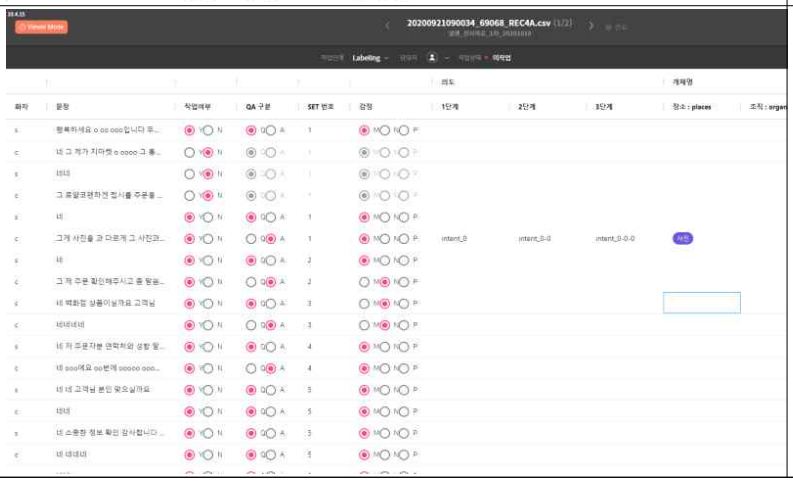
- (작업 이동) 작업後, 작업中, 작업前 파일을 간편을 간편하게 이동

- (화면 조절) 화면 상단에서 작업 화면 크기/밝기/색상, 실행 취소, 재실행 등 제공
- (가공 지원) 어노테이션 작업 편의를 위한 간편하게 클릭만으로 이미지 이동, 바운딩박스, 라벨 , 선택 삭제, 초기화 기능 등을 제공

다) (결과물) 작업 결과물을 다양한 형식 제공할 수 있도록 데이터 포맷 컨버터 기능 지원

라) (저작도구 공개) 과제를 통해 진행된 저작도구는 소스와 기술 매뉴얼(데이터셋 형태, 규모, 특성 등)을 공개하여 외부에서 활용이 가능

마) 상세 작업 방법 및 설명

	1. 작업툴에서 작업할 파일 선택
	2. 작업대에서 작업할 리스트 확인
	3. 각각의 항목에 대해 인스턴스 선택 및 입력

	4. 작업 완료 후 다음 작업으로 이동

2.5 검수

2.5.1 검수 절차

<데이터 품질관리 프로세스>

구 분	프로세스	설명
데이터 분석	대상 식별	‣ 고객사의 품질관리 요구사항을 확인 ‣ 품질관리를 수행할 대상을 구체화 및 문서화
데이터 설계	규칙 정의	‣ 품질관리 대상에 대한 프로파일링을 시행하고 ‣ 품질 측정 및 통제를 위한 지표를 설정 ‣ 설정된 품질규칙은 데이터 가공 업무규칙에 반영
데이터 가공	측정	‣ 데이터 가공 결과물 중 품질관리 대상에 대한 품질 측정
	분석	‣ 품질 측정 결과를 품질지표와 비교하여 시사점 도출 ‣ 개선이 필요한 부분에 대한 원인 및 개선방법 분석
	개선	‣ 오류의 영향도 및 시급성을 고려하여 개선 시행
	통제	‣ 품질측정-분석-개선이 선순환 구조를 이룰 수 있도록 지속적인 모니터링 수행

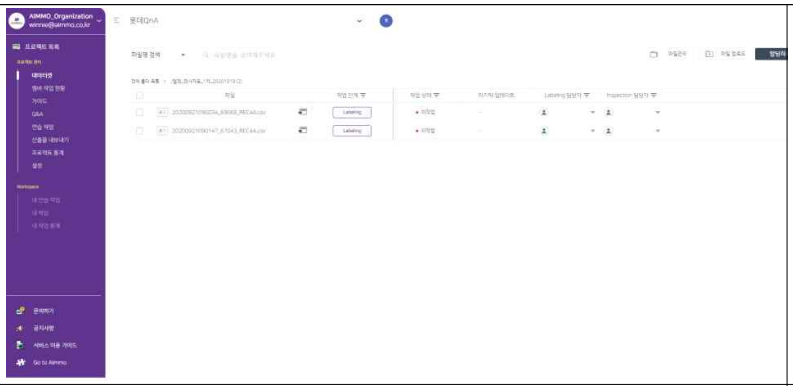
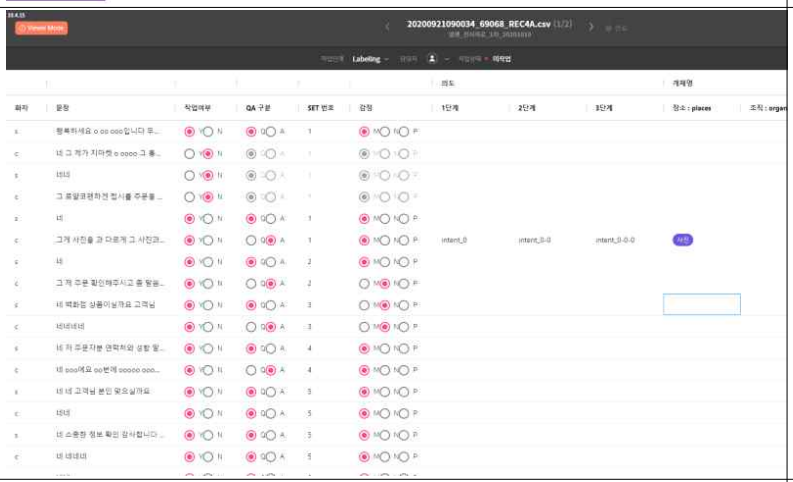
2.5.2 검수 기준

구분		측정 지표	정량 목표
정확도	구조 및 형식	어노테이션 포맷 정확도	정합률 100 %
	참값 (Ground Truth)	참값 정확도	질의응답의 정확성 100%
유효성	학습 성능	인공지능 모델 학습 검증률	Validation Accuracy 90%

2.5.3 검수 조직

획득조직	명	설명
라벨러	1000	어노테이션 저작도구인 GTaas 를 사용하여 라벨링 작업을 진행하는 인력
AQM	10	라벨러들을 관리하고 진행함에 있어서 문의사항 응대 및 1차 작업물 확인 및 관리하는 팀. 또한 최종 작업물에 대한 검수도 함께 진행

2.5.4 검수 작업 순서

	1. 검수자에게 할당되어있는, 검수할 대상 파일 선택
	2. 라벨러가 입력완료한 데이트를 보고 입 확인 후 검수 진행
	3. 검수 완료 후 완료 처리 또는 반려 처리. 반려 시 작업자에게 재할당되어, 재작업 진행 후 동일 작업 반복

2.5.5 기타 품질관리 활동

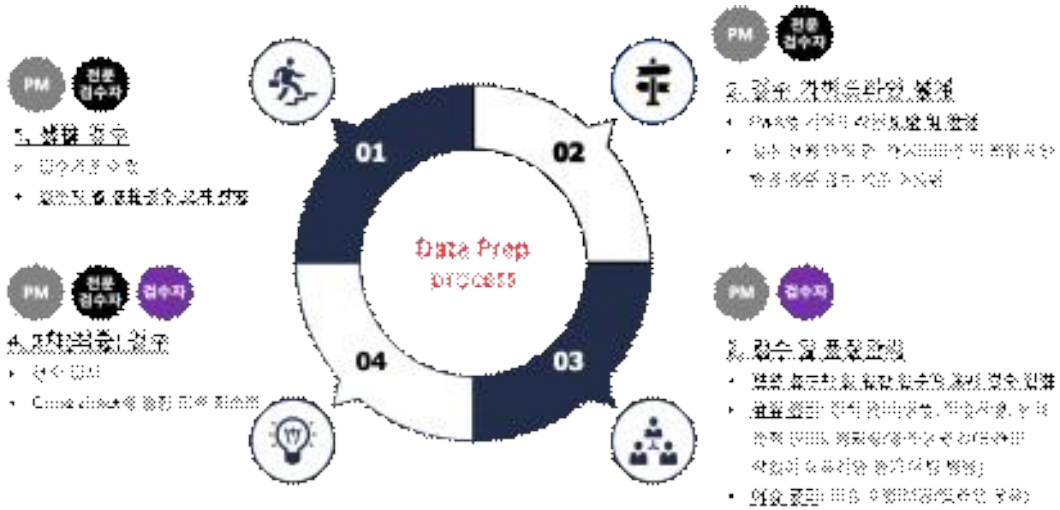
가) 전문업체를 통한 이중 검수 진행

- 질문과 답변과의 연관성과 정확성, 사용성 확보를 위해 part별 담당 인력과의 입체적인 점검을 진행하

며, 전문 인력의 지속적인 채용 및 내부 교육/리뷰 회의를 통해 검수에 투입되는 인력 역량을 향상시키도록 함.

- 검수 팀원 4인의 전수 검수 방식으로 진행하며, 데이터 구축과정에서 가이드에 미포함 되었던 특이사항들을 지속히 작업 단계에 반영할 수 있도록 진행하는 <작업자 판단 검수>를 도입하도록 함. 또한 작업 종료 후 필요시 N차까지 전문가 및 검수/작업자 인원들의 교차 검증을 통하여 데이터의 품질을 관리하도록 함.
- 검수자의 검수물 대상으로 검수 팀장의 <교차 검수> 진행 및 피드백에 따른 재검수 실시

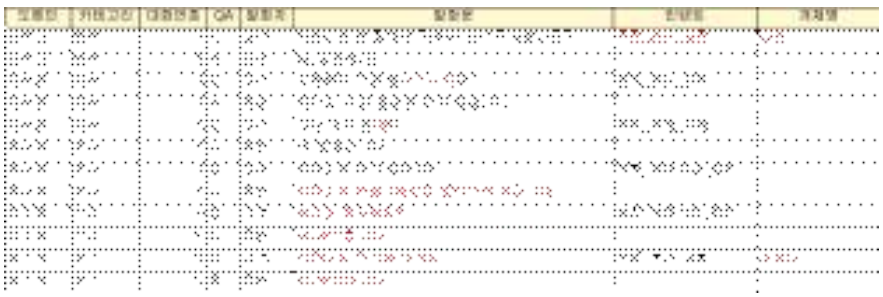
<검수 절차 개요도>



나) 이중 검수 프로세스

- <이중 검수> 프로세스는 구현 데이터의 기계학습 용이성 및 AI 서비스 구현상황을 동시에 고려하여 본 사업 구현 이후까지 고려한 검수를 목표로 함. 특히 정량적 검수의 경우, 문장경계의 구성, 맞춤법, 오탈/오기, 윤리적 적절성, 시제/높임말 호응 등 어학 전문가의 검수를 통한 검수를 진행

<검수 대상 데이터>

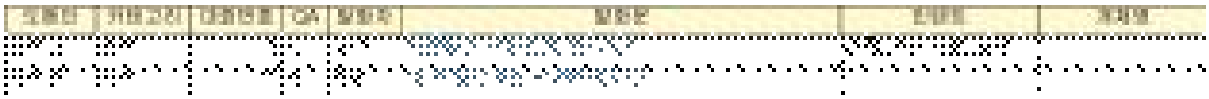


- 정성적 검수

✓ 대화 데이터의 목적 적합성 판단: 무인판매용 AI의 학습데이터로 적절한지 여부를 확인하고, 부적절한 문장은 삭제하거나 대화 흐름에 맞게 수정한다.

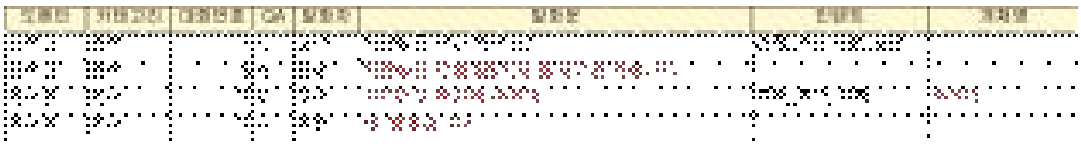
<목적 부적합 문장 예시>

<검수 및 교정 예시>

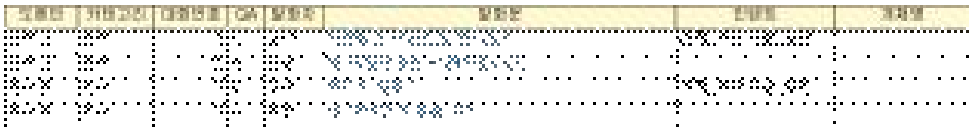


✓ 자연스러운 대화 흐름 판단: 대화의 자연스럽게 이어지는지 확인하고, 부자연스러운 대화셋은 대화 흐름에 맞게 수정한다.

<부자연스러운 대화 흐름 예시>



<검수 및 교정 예시>

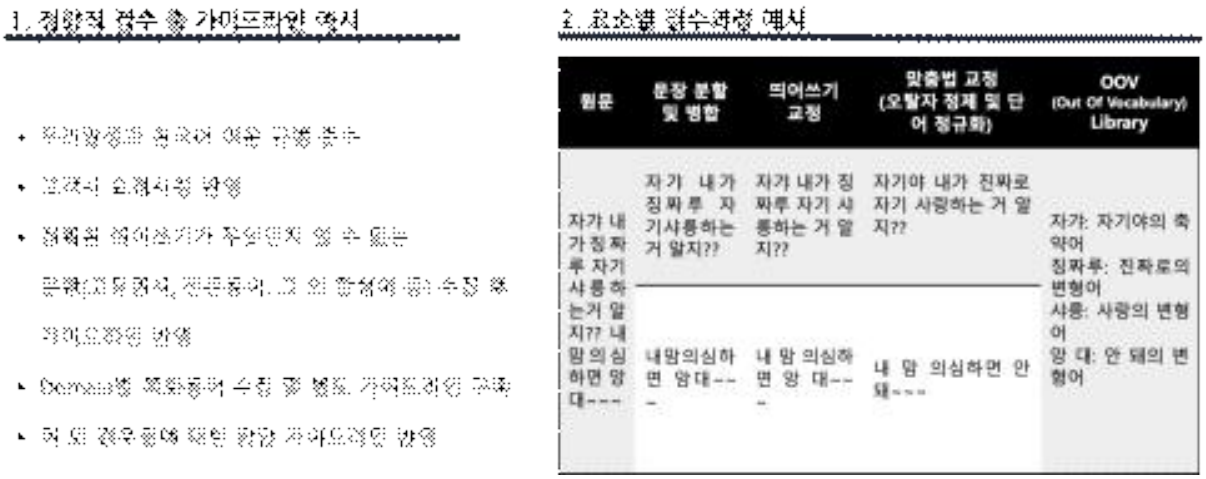


- 정량적 검수

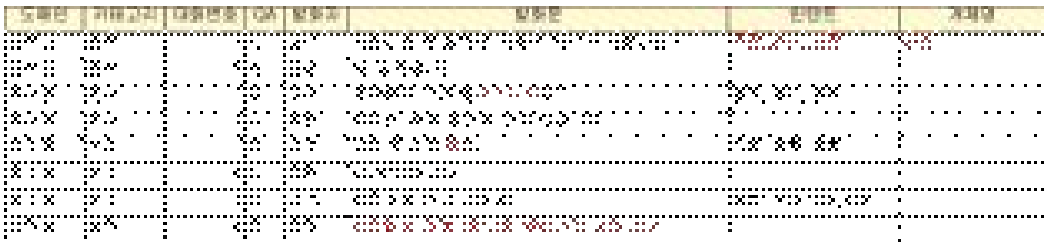
✓ 오탈자, 맞춤법, 윤리적 적절성 등 교정 전반 검수

✓ 시제/높임말 호응 등 비문 전반 검수

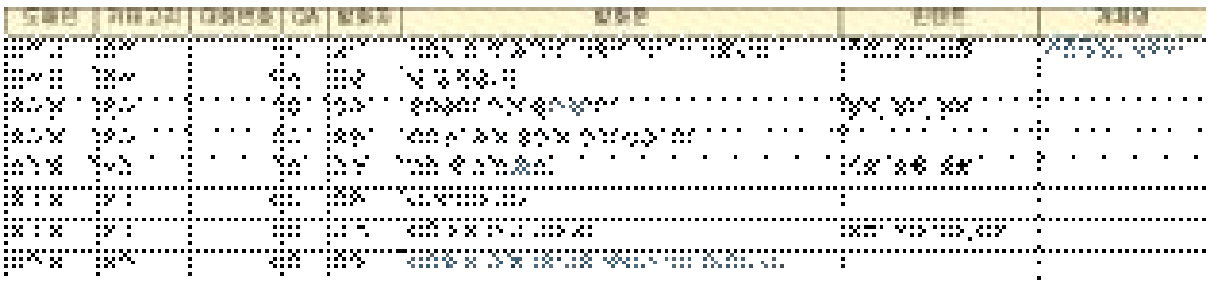
<정량적 검수의 요소별 검증 프로세스>



<정량적 평가 대상 예시>

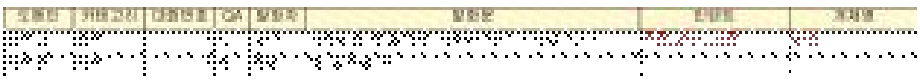


<검수 및 교정 예시>

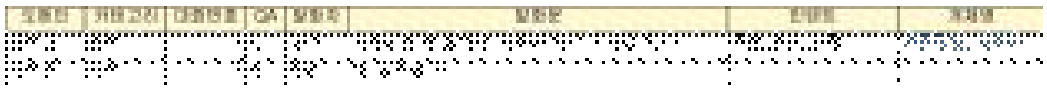


- 발화문의 인텐트 적합성 체크

<부적합한 인텐트 태깅 예시>

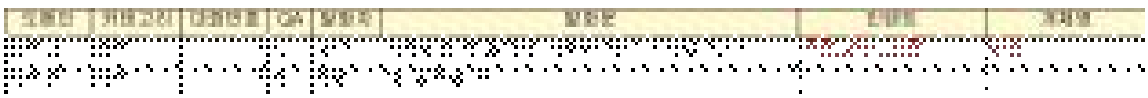


<검수 및 교정 예시>



- 개체명 누락 여부 확인 및 정확도 체크

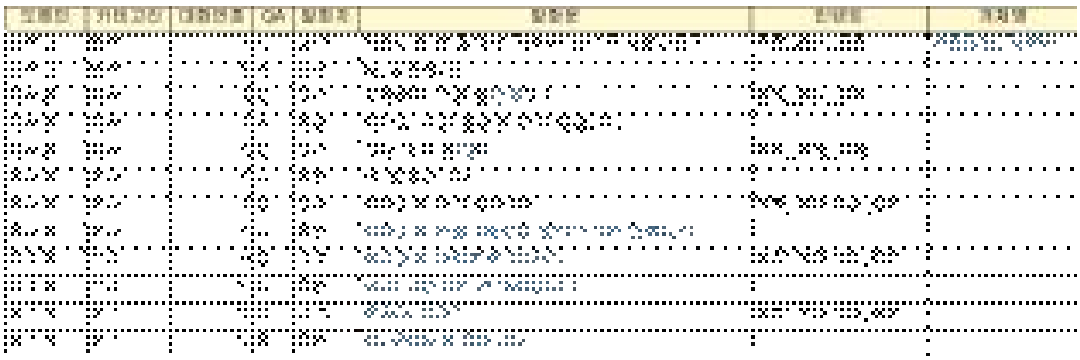
<개체명 누락 예시>



<검수 및 교정 예시>

- 검수 후 최종 산출물 예시

<최종 산출물(안)>



다) 검수단의 입체적 구성과 검수 과정의 적절성 확보

- 검수인력의 충족 기준

✓ 우수 가공자 중 일부 선별하며, 어학(국어국문 등) 전문가를 우선 투입하도록 함

✓ 학석사 4인의 팀원과 석박사 이상의 고급수준의 전문가 검수팀장으로 조직을 구성함

✓ 이를 통해 시행 착오를 최소화하는 동시에 높은 검수 품질을 제고하도록 함

✓ 검수 인력의 전문성은 기계학습 시의 효율성 및 AI 서비스 구현 시의 적절성의 두 가지 목표 달성을 위하여, 어학 전문성 뿐 아니라 데이터 가공 전문 경력자를 우선 투입하도록 함

<검수 인력의 전문성>

2.6 활용

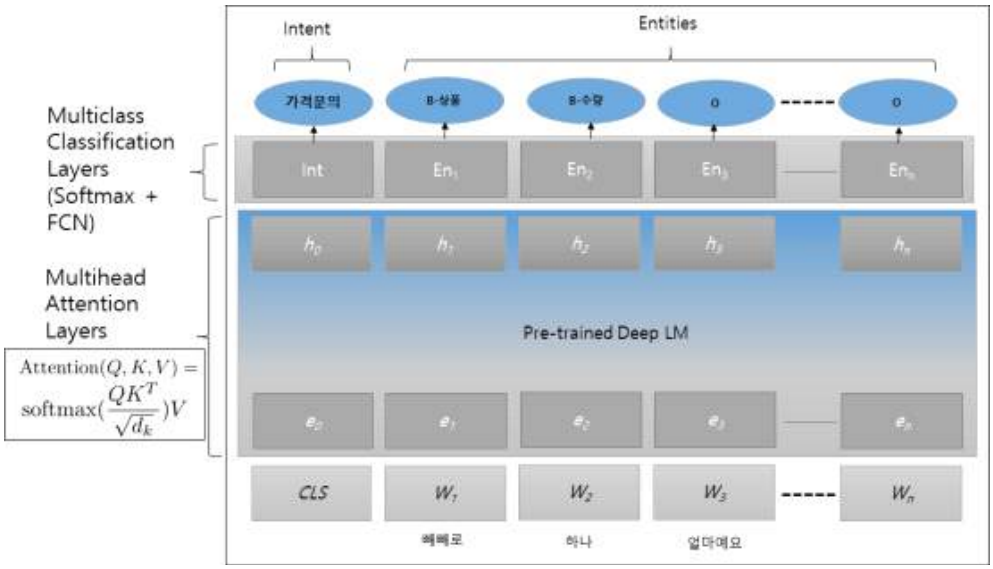
2.6.1 활용 모델

2.6.1.1 모델 학습

가) NLU 모델 구조 및 학습

- 데이터를 통해 학습시킬 NLU 모델은 아래와 같다.

< NLU 모델 >



구축된 데이터를 통해서 학습시킬 모델은 위 그림과 같이 입력된 문장에 대해서 그 문장의 의도(Intent) 와 객체들(Entities)을 분류하고 추출하는 역할을 한다. 모델의 구조는 문장내 각 토큰을 변환시킨 Q(Query), K(Key), V(Value)을 바탕으로 토큰들간 상호 관련성을 모델링하는 Multi-Head Self-Attention 메커니즘을 핵심으로 하는 Transformer encoder 구조를 활용한 BERT 나 그와 유사한 파생모델들을 주축으로 삼는다. 이 모델들을 많은 양의 한국어 데이터 상에서 사전학습 시킨 뒤 이를 한국어 언어모델(LM)로 활용하여 문장의 문맥(Context)를 고려한 표현정보가 담긴 마지막 hidden layer를 입력으로 삼는 FCN 과 Softmax 로 구성된 인텐트 및 엔티티 classification layer를 두고, 구축된 데이터에 포함된 인텐트와 엔티티와 레이블을 ground truth 로 삼아 supervised training을 통해 fine-tuning을 실시한다. 이를 위해 데이터셋을 학습(training), 검증(validation), 평가(test) set 으로 7:1.5:1.5 정도의 비율로 나눠 학습을 실시하고 accuracy 및 F1 값을 성능 metric으로 삼는다. Fine tuning 학습시 아래와 같은, 문장의 인텐트(Int)와 엔티티들($En_{1\sim n}$)의 cross entropy 목적함수를 최소화하는 파라미터값을 구한다.

$$L_\theta = -\log P(y_\theta^{Int}) \prod_{k\in\{1\cdots n\}} P(y_\theta^{En_k})$$

데이터셋을 학습, 검증, 평가 용도로 재구성할 때에 데이터셋 전체나 상점 카테고리 전체를 임의로 분리해서는 모델 성능을 높이지 못한다. 아래와 같은 기준을 참고하여 데이터를 분리해야 한다. 임의로 생성한 문장이 아니라 실제 상담대화를 가공한 데이터이기 때문에 인텐트 비율에 편차가 크기 때문이다.

1. 필요한 인텐트 대화만 추출하여 학습/검증/평가 데이터 구성
2. 필요한 상점 카테고리 및 인텐트 대화만 추출하여 학습/검증/평가 데이터 구성
3. 필요 시, QA여부 컬럼 중 'Q'(질문) 항목만 추출하여 학습/검증/평가 데이터 구성
4. 선정한 인텐트 간 건수 차이가 크면 많은 쪽의 인텐트 대화 건수를 줄여서 데이터 구성

모델 구현은 tensorflow 나 pytorch 같은 deep learning 프레임워크를 사용하여 구현하고 이 모델을 빠른 반복학습을 위해 GPU 상에서 학습시킨다. 배치사이즈는 GPU 카드의 메모리 크기를 고려하여 적절한 크기로 설정한다.

2.6.1.2 서비스 활용 시나리오

가) 챗봇 서비스 모델

- 구축된 데이터로 학습된 NLU 모델을 활용하는 챗봇 서비스 구성도는 아래와 같다.

< 챗봇 서비스 모델 >

- NLU모델과 Chatbot 서비스는 Rest API를 통해 서로 통신한다. 사용자가 문장을 입력하면 챗봇 서비스에 의해 Rest API를 통해 NLU 모델에 그 문장이 전달된다. NLU 모델은 이 문장에 대해 인텐트와 엔티티들을 추출하여 Rest API를 통해 챗봇 서비스에 결과를 회신한다.

2.6.2 데이터 제공

가) 구축 도구 가이드북 배포

- 구축 도구의 구현 방법론, 동작 순서, 지원 사항 등에 대해 상세히 기술하고, 문서 요약 작업, 검색 작업 등에 대한 상세한 설명 제공
- 구축된 데이터셋을 민간 기업 등 외부에서 인공지능 학습, 지능정보 서비스 개발 등 실제 민간 산업에 활용할 수 있도록 추상요약 학습데이터 및 학습 방법에 대한 기술 매뉴얼을 작성하여 제공할 예정

나) 데이터 셋 구축을 위한 플랫폼 소스 공개

- 오픈 소스는 최대 개발자 커뮤니티인 깃헙(GitHub)에 등록하여 유지 보수 및 이슈를 관리하고,

오픈 소스에 대해 문의 및 피드백에 대해 기술 지원이 이루어 질 수 있도록 효과적 시스템 제공

다) 학습 데이터 개방 및 확산

- 구축된 학습 데이터는 AI HUB에 등록/배포하고 챗봇 API를 함께 제공하여 AI HUB를 중심으로 학습데이터 개방 및 확산이 이루어질 수 있는 생태계를 구현

라) 국내 인공지능 기술혁신 지원 방안 (소스, 활용방법 및 매뉴얼 제공 등)

- 비전문가도 쉽게 활용할 수 있도록 지원하기 위해 구축 시나리오 및 상세 가이드를 제공하며, 데이터의 추출, 정제, 학습데이터 구축 과정, 프로토타입 개발 방법 등을 정리하여 공개하고, 구축된 데이터셋을 민간 기업 등 외부에서 AI학습, 지능정보 서비스 개발 등 실제 민간 산업에 활용할 수 있도록 추상요약 학습데이터 및 학습 방법에 대한 기술 매뉴얼을 작성하여 제공함
- 세부적으로 구축 가이드에 포함되는 사항은 아래와 같음
  소스를 포함한 기본 플랫폼
  질의응답 데이터 셋 설명 가이드
  데이터와 챗봇빌더 연계 프로토타입 개발 방법
- 과제를 통해 개발된 인공지능 서비스와 관련한 정보처리 시스템의 매뉴얼에 대해서는 국내 인공지능 기술혁신을 위한 확정된 공개장소를 통하여 공개하며, 소상공인의 상황에 맞게 상황별, 기능별 활용이 가능하도록 API 매뉴얼 제공
- 스타트업, 소규모의 중소기업은 인공지능 엔진을 도입에 필요한 비용 확보가 어렵기 때문에 본 과제를 통해 개발한 엔진을 기반으로 추가 개발을 진행하고 서비스 오픈 후 도입한 기업의 수익을 공유하는 방식으로 초기 기술 개발 비용을 절감하고 기술 이전 업체는 지속적으로 수익을 확보할 수 있는 구조가 확립되어 안정적인 수익 구조 확보 가능함

참고. 의도(Intent) 목록 v2.3

분류 1 (17)	분류 2 (96)	분류 3
AS	날짜	1. 질문 (방법, 정보획득) 2. 요청(의지표출, 가능여부질의, 행위의지표출) - 가능여부질의: ex) ~가능한가요, ~되나요 ※ 단, 구매자의 요청에 판매자가 어떠한 액션을 취할 수 있는 경우에 한함 ex) 17일까지배송가능한가요? - 요청 ex) 드라이클리닝가능한가요? - 요청X → 단순질문 - 가능여부(구체적일자 vs. 단순질문) ex) 17일까지 배송가능한가요? - 요청 ex) 언제까지 배송가능한가요? - 질문 3. 비교 4. 확인(재차 질문 어조) - ex) ~되는 거죠, ~되는 거 맞죠?, ~된 거죠
	방법	
	비용	
	시간	
	일반	
결제	방식	
	수단	
	시기	
	영수증오류	
	일반	
	재결제	
	추가	
	취소	
	할인	
교환/반품/환불	날짜	
	방법	
	비용	
	시간	
	일반	
구매	변경	
	예약	
	오류	
	제품	
	추가	
	취소	
매장	구조	
	부대시설	
	이용	
	정보	
멤버십	가입	
	사용	
	오류	
	일반	
	적립	

배송	날짜	
	방법	
	비용	
	오류	
	일반	
	지역	
	택배사	
부가서비스	날짜	
	방법	
	비용	
수납	방법	
	원무	
수술/입퇴원	방법	
	비용	
	생활	
	예약	
	정보	
예약	방법	
	변경	
	비용	
	취소	
외래	방법	
	비용	
	시간	
	예약	
	정보	
웹사이트	가입	
	사용	
	오류	
제품	가격	
	구매	
	구성	
	날짜	
	방법	
	불량	
	소재	
	시용	
	용도	
	원산지	
	일반	
	입고	
	재고	
	정보	
	추천	
	커스텀	
	품질	
	호환	
주문	변경	
	비품	
	오류	

	제품	
	추가	
	취소	
포장	방식	
	비용	
	일반	
행사	기간	
	날짜	
	유형	
	일반	
	정보	

참고. 상점 카테고리 목록

카테고리명 (14)	**비고**
가구인테리어	-
건강	-
디지털가전	-
병원	-
뷰티	-
생활잡화	-
슈퍼	-
식품	-
음식점	-
의류	-
출산육아	-
카페	-
패션	-
기타	-

붙임1 인공지능 데이터 명세서 양식

데이터 이름	소상공인 고객 주문 질의-응답 데이터	
데이터 포맷	미리 정의된 데이터 항목을 모두 포함하는 CSV 파일 질의응답 데이터 500만건 목표	
활용 분야	유통, 음식점 등 다양한 분야의 무인매장, 비대면 업무 등의 환경에서 활용	
데이터 요약	1. 음식점, 슈퍼, 농수산물 등 다양한 상점에서 발화한 질의응답데이터를 포함 2. 상품, 결제, 반품, 배송 등 소상공인이 사용할 수 있는 업무 전반의 주제를 포함 3. 질의응답셋에 감정, 대화의도, 개체명 등의 정보를 추가로 표시함 4. 각 질의응답의 데이터는 단순 질의-응답으로 끝날 수도 있으며 계속 질문이 이어지는 경우 순서를 표시하여 대화의 흐름을 알 수 있게 하였음 (고객 질의응답 데이터는 상담 대화를 근간으로 생성했으나, 상담사 인사로 시작하여 본인 인증 관련 대화, 연락처 및 주소 등 개인 및 민감 정보를 포함한 대화, 마무리 인사하는 대화를 제외하면 실제 질문은 한두 개 정도로서 맥락이 이어지는 문답은 거의 없음. 이어지는 문답 별로 인텐트 대분류가 달라지는 실태를 감안하여 NIA 및 TTA와 '자연스러운 대화 흐름' 여부는 검증 범위에서 제외하기로 합의했음.)	
데이터 출처	주제 분야: 상품, 결제, 반품, 배송 등 소상공인이 사용할 수 있는 업무 전반 출처 정보: 인터뷰를 통한 직접 녹취, 데이터 직접 생성, 음성데이터 기반 데이터 추출	
데이터 이력	배포버전	1.4
	개정이력	2021.02.26. TTA 요청사항 반영
	작성자/ 배포자	전시형 수석 / 최봉우 책임

데이터 구성

루트 폴더	00. 초기 데이터	라벨링데이터	(상점)카테고리: 5종	284,494건	
	01. 데이터	Training	라벨링데이터	(상점)카테고리	
		Validation	라벨링데이터	(상점)카테고리	
		Test	라벨링데이터	(상점)카테고리	
		Sample	라벨링데이터	(상점)카테고리	
		구축 가이드			
	02. 저작도구				
	03. AI모델				
	04. 활용 서비스				
	~~05. 원시 데이터~~	(해당사항 없음. 개인정보 비식별화 이전의 고객 상담/대화 데이터는 공개 불가능하며 삭제해야 함.)			

어노테이션 포맷

항목	설명	타입	필수구분
IDX	질의응답 데이터 파일 내 고유 순서 번호	Num.	Y
발화자	발화자 정보 (c: 고객 s: 점원)	string	Y
발화문	대화 텍스트 정보	string	Y
카테고리	발화가 일어나는 상점 정보	string	Y
QA번호	질의응답셋을 구분하는 정보	Num.	Y
QA여부	질의문(q)인지 응답문(a)인지 표시	string	Y
감성	텍스트별 감성 정보 (m: 중립, n:부정, p:긍정)	string	Y
인텐트	질의문 기준 발화문에 내재한 의도	string	Y
개체명	NER(개체명인식)을 위한 개체 정보	string	N
상담번호	대화 상황 구분 정보	Num.	Y
상담내순번	상담 내 발화 순서 표시	Num.	Y

<table>
<tr><td rowspan="3">데이터 통계</td><td>데이터 구축 규모</td><td>
<table>
<tr><th>데이터 출처</th><th>질의응답데이터 규모 (최종 산출물 기준)</th><th>도메인</th></tr>
<tr><td>콜센터 데이터</td><td>400 만 건 질의응답</td><td>백화점, 홈쇼핑, e-commerce 등 유통 관련</td></tr>
<tr><td>녹취 데이터</td><td>100 만 건 질의응답</td><td>도.소매업, 숙박.음식업점, 수리, 기타개인서비스업, 보건업 등에 해당하는 약 20 종 상점 (한국표준산업분류 10 차 기준)</td></tr>
</table>
</td></tr>
<tr><td>데이터 분포 예시</td><td>
- 구축 데이터 질의응답데이터: 5,000,000 건
- 데이터 출처 별 데이터(질의응답데이터) 수 (단위: 10,000건)
- 데이터 산업 구분별 데이터 비율(추정치, 한국표준산업분류 10차 기준)
<table>
<tr><th colspan="2">데이터 출처</th></tr>
<tr><td>콜센터</td><td>직접 녹취</td></tr>
<tr><td>400</td><td>100</td></tr>
</table>
<table>
<tr><th colspan="5">콜센터별 질의응답비율(도메인)</th></tr>
<tr><td>e-commerce</td><td>홈쇼핑</td><td>백화점</td><td>마트</td><td>그 외</td></tr>
<tr><td>50%</td><td>10%</td><td>10%</td><td>15%</td><td>15%</td></tr>
</table>
<table>
<tr><th colspan="4">직접 녹취 상점 비율</th></tr>
<tr><td>소매업</td><td>숙박 및 음식업점</td><td>제조업</td><td>보건업</td></tr>
<tr><td>30%</td><td>30%</td><td>20%</td><td>20%</td></tr>
</table>
</td></tr>
<tr><td>기타 활용 통계</td><td>유사통계 없음</td></tr>
</table>

<table>
<tr><td rowspan="4">기타 정보</td><td>대표성
(Coverage)</td><td>콜센터 데이터의 경우, 롯데 그룹 내 그룹사 인 백화점, 홈쇼핑, 수퍼, 마트, 이커머스, 건설 등의 상담 대화로써 유통업과 관련된 범위 나타냄
소상공인 녹취 데이터는 음식점, 의료, 카페, 가전, 농산물, 수산물, 생활잡화, 정육점, 동물병원, 수퍼 등의 도메인을 포함함</td></tr>
<tr><td>독립성</td><td>
<table>
<tr><th>데이터 출처</th><th>원시데이터 형태</th><th>정제 데이터 형태</th><th>민감정보</th><th>잡음</th></tr>
<tr><td>콜센터 데이터</td><td>음성 파일</td><td>텍스트 파일</td><td>원시데이터에 존재할 수 있으나 비식별화</td><td>NA</td></tr>
<tr><td>녹취 데이터</td><td>음성 파일</td><td>텍스트 파일</td><td>원시데이터에 존재할 수 있으나 비식별화</td><td>원시정보에 존재할 수 있음</td></tr>
</table>
</td></tr>
<tr><td>유의사항</td><td>
<table>
<tr><th>유의사항</th><th>설명</th></tr>
<tr><td>도메인</td><td>데이터에 포함되어 있지 않은 도메인(상점)의 경우 AI 활용에 다소 어려움 있음</td></tr>
<tr><td>개체명 정의</td><td>데이터에 정의된 개체명 항목의 정의가 활용 목적에 따라 맞지 않을 수 있음</td></tr>
<tr><td>감정 정의</td><td>긍정/부정에 대한 감정 정의가 활용 목적에 따라 상이할 수 있음</td></tr>
</table>
</td></tr>
<tr><td>관련 연구</td><td>관련 연구 없음</td></tr>
</table>