●○ 한국어 텍스트 과제

민원(콜센터) 질의-응답 데이터

●○ 개요: 민원(콜센터) 질의-응답 데이터셋이란?

코로나19 확산 이후 콜센터 집단감염으로 인한 상담사 재택근무 등 시장 변화에 따른 민원콜의 증가로 콜센터 업무에 많은 어려움이 발생하였다. 이에 콜센터 업무 효율화를 위해 AI기술을 활용한 ICC(Intelligent Contact Center) 지능형 컨택센터의 필요성이 대두되었고 관련 기술 개발에 활용할 수 있는 학습 데이터셋 구축을 위해 민간기업_ 공공기관의 콜센터(민원) 상담 내역을 활용한 질의응답 학습 데이터셋을 구축하였다. 이 데이터셋은 상담사와 고객간의 질의-응답 110만쌍, 이를 녹취한 음성데이터 440시간으로 구성되어 있다.

●○ 데이터셋의 구성

본 데이터셋은 실제 상담데이터를 기반으로 음성데이터를 재가공한 텍스트데이터 110만 대화쌍, 재가공한 텍스트 데이터를 녹음한 음성데이터로 크게 2가지 형태로 나뉘어져 있다.

상담 건마다 상담시간의 차이가 있기 때문에 대화쌍을 기준으로 데이터를 구성했으며, 데이터 편향 방지 및 모델링 구축을 위해 카테고리별 데이터셋을 설계하였다.

데이터셋은 정형 데이터와 비정형데이터로 구분하였으며, 자연스러운 상담 대화 셋을 구성하기 위해 비정형데이터의 비중을 늘려 데이터를 구축하였다.

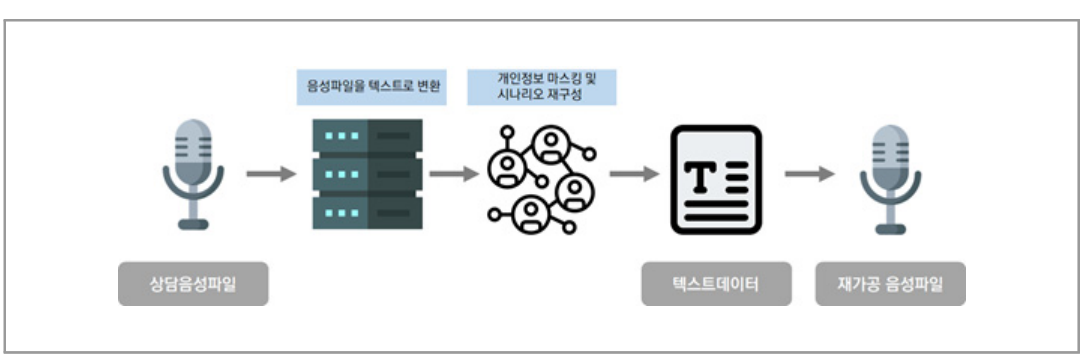


그림1 | 질의-응답 데이터 자료 형태

음성데이터를 재가공한 텍스트 데이터

- 원천 음성데이터를 STT엔진을 통해 텍스트로 변환
- 텍스트로 변환 후 개인정보 제거 및 비문 수정
- 해당 질의-응답 데이터의 개체명, 용어사전, 지식베이스 추출 후 자체 저작도구에 업로드

재가공한 텍스트 데이터를 녹음한 음성데이터

- 가공된 질의-응답 데이터를 기반으로 한 음성 녹음 데이터
- 실제 상담을 하는 듯이 녹취 진행

●○ 데이터셋의 설계 기준과 분포

kth가 자체적으로 보유하고 있는 Daisy엔진의 데이터 처리 기술(STT엔진, 상담내용 요약, 상담카테고리 자동분류, 고객정보 자동 Blur처리)을 이용해 개인정보 이슈가 없는 원천데이터를 확보하였다. 이러한 원천데이터를 다음과 같은 과정을 통해 구축하였으며 해당 데이터는 콜센터 업계의 AI 상담 품질 개선을 목적으로 데이터셋을 설계하였다.

구축단계	세부 절차	설명
1. 수집	1.1 원천 데이터 선정	원천데이터를 수집하는 대상 선정(저작권 확인, 문장 형식 검토)
	1.2 원천 데이터 수집	기존 음성데이터를 텍스트(문장) 형태로 변환
2. 정제	2.1 개인정보 삭제	데이터 내 고객의 개인정보 (성명,주소,전화번호,결제정보) 삭제
	2.2 비문 번역	STT가 제대로 변환하지 못한 음성자료 Text로 변환
3. 가공	3.1 시나리오 화	Dialog Kit을 사용해 정제데이터 기반 시나리오 제작
	3.2 음성녹음	시나리오 기반 음성데이터 재 녹음
4. 검수	4.1 완성된 시나리오 검수	개인정보 삭제 여부 확인
	4.2 일치 여부 검수	시나리오와 녹음파일 일치여부 검토
5. 활용	5.1 챗봇,콜봇	상담 관련 챗봇,콜봇 학습데이터로 활용

그림2 | 데이터 구축 프로세스

- 상담데이터를 정형, 비정형데이터로 나누어 분류 하였으며 상담데이터 중 개인을 식별할 수 있는 정보인 성명, 주민등록번호, 이메일주소, 카드번호 등의 개인정보는 삭제하여 개인정보 이슈가 없도록 데이터를 구축하였다.

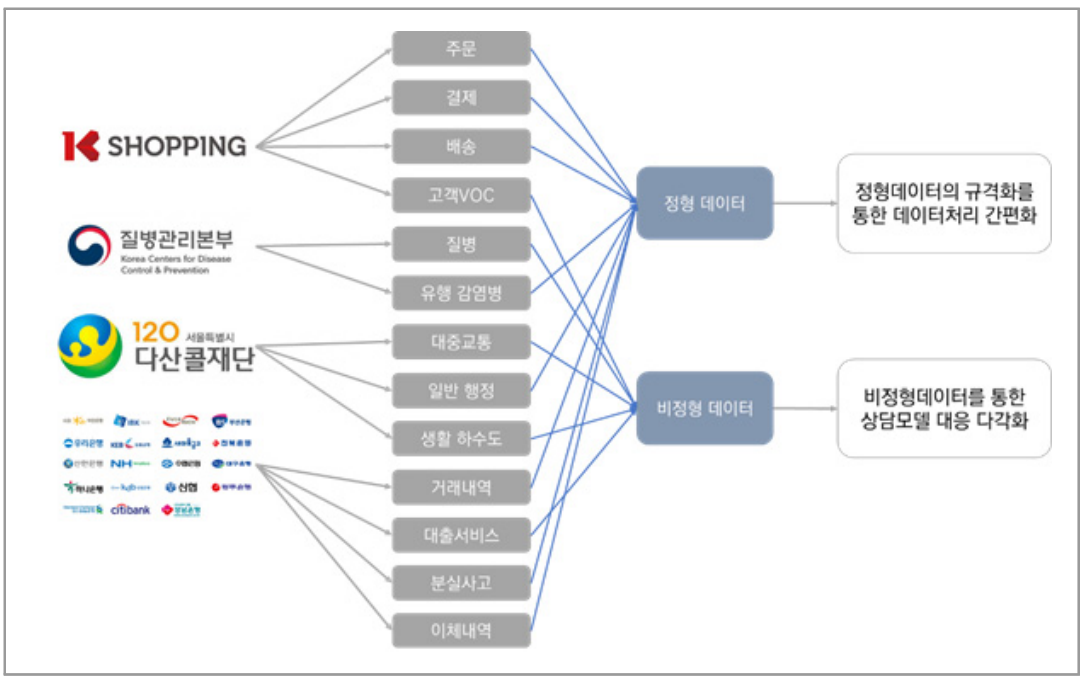


그림3 | 정형데이터, 비정형 데이터 분류

- 원천데이터 선정은 민간기업/공공기관의 저작권 문제가 해결된 질의-응답 데이터로 구성하였다.

데이터명	데이터출처	대화쌍	음성 데이터 분량	형태
민간기업	K쇼핑	55만쌍	220시간 이상	텍스트
	금융/보험	20만쌍	80시간 이상	텍스트
공공기관	질병관리본부	25만쌍	100시간 이상	텍스트
	다산콜센터	10만쌍	40시간 이상	텍스트

그림4 | 데이터셋 분포

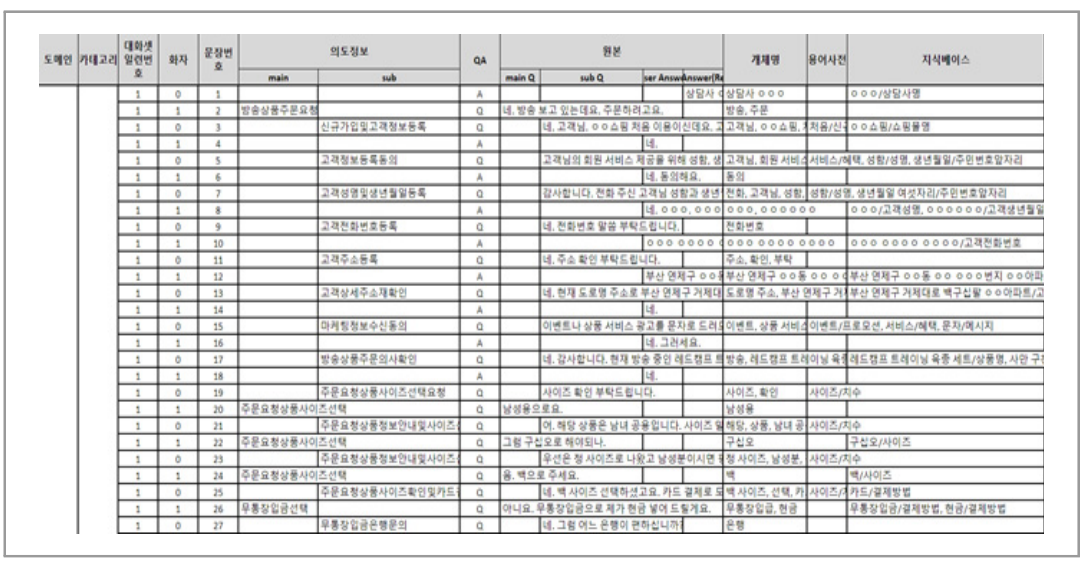
●○ 데이터 구조

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

항목		설명	예시
도메인		대분류	주문, 결제, 배송 등
카테고리		중분류	주문확인, 주문처리, 단순배송문의 등
대화셋 일련번호		데이터건 별 식별번호	데이터건 별 1,2,3..
화자		상담사와 고객을 분리하는 번호	0 : 상담사, 1 : 고객
문장번호		질의, 응답 한 문장 당 부여된 번호	문장별 1,2,3…
의도 정보	main	고객의도	방송상품 주문요청
	sub	상담사의도	고객주소등록
QA		질의-응답 구분	Q:질의, A:응답
원본	main Q	고객 질문	네 방송 보고 있는데 주문하려구요
	sub Q	상담사 질문	고객님 상세 주소 알려주시겠어요?
	User Answer	고객 응답	네, 전화번호는 ooo oooo oooo
	System Answer	상담사 응답	네 등록해드리겠습니다.
개체명		문장 중 명사 단어들을 ‘,’로 분리	주문, 삼성카드, 현금 영수증
용어사전		개체명의 유사 및 동의어	성함/성명/이름, 신청/주문
지식베이스		개체명이 해당 문장에서 쓰이는 의미를 기술	엘에이 갈비/상품명, 삼성카드/카드명

●○ 데이터 예시

이 데이터셋 예시는 모든 데이터 카테고리에 해당한다.

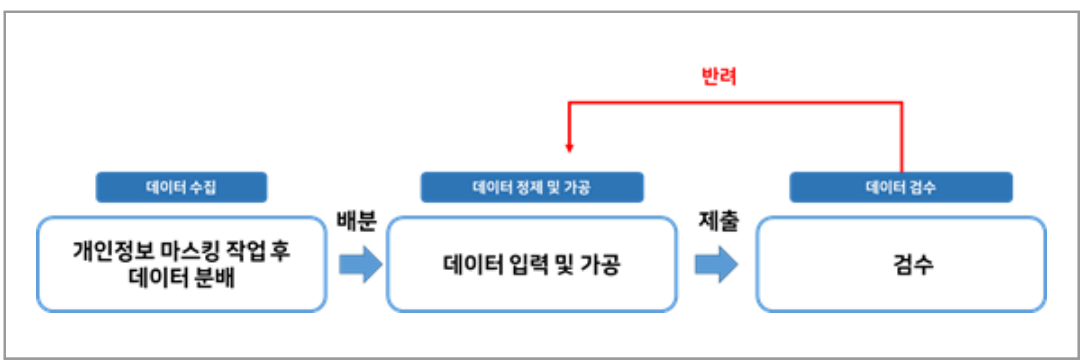


도메인	카테고리	대화셋 일련번호	화자	문장번호	의도정보		QA	원본			개체명	용어사전	지식베이스
					main	sub		main Q	sub Q	ser Answ Answer(Re			
		1	0	1			A			상담사 (	상담사 ○○○		○○○/상담사명
		1	1	2	방송상품주문요청		Q	네, 방송 보고 있는데요, 주문하려고요.			방송, 주문		
		1	0	3		신규가입및고객정보등록	Q	네, 고객님, ○○쇼핑 처음 이용이신데요, 고			고객님, ○○쇼핑, 처음용/신규		○○쇼핑/쇼핑몰명
		1	1	4			A			네,			
		1	0	5		고객정보등록동의	Q	고객님의 회원 서비스 제공을 위해 성함, 생			고객님, 회원 서비스	서비스/혜택, 성함/성명, 생년월일/주민번호앞자리	
		1	1	6			A			네, 동의해요.	동의		
		1	0	7		고객성명및생년월일등록	Q	감사합니다. 전화 주신 고객님 성함과 생년			전화, 고객님, 성함,	성함/성명, 생년월일 여섯자리/주민번호앞자리	
		1	1	8			A			네, ○○○, ○○○	○○○, ○○○○○○		○○○/고객성명, ○○○○○○/고객생년월일
		1	0	9		고객전화번호등록	Q	네, 전화번호 말씀 부탁드립니다.			전화번호		
		1	1	10			A			○○○ ○○○○ (	○○○ ○○○○ ○○○○		○○○ ○○○○ ○○○○/고객전화번호
		1	0	11		고객주소등록	Q	네, 주소 확인 부탁드립니다.			주소, 확인, 부탁		
		1	1	12			A			부산 연제구 ○○동	부산 연제구 ○○동 ○○ ○○		부산 연제구 ○○동 ○○ ○○○번지 ○○아파
		1	0	13		고객상세주소재확인	Q	네, 현재 도로명 주소로 부산 연제구 거제대			도로명 주소, 부산 연제구 거		부산 연제구 거제대로 백구십일 ○○아파트/고
		1	1	14			A			네,			
		1	0	15		마케팅정보수신동의	Q	이벤트나 상품 서비스 광고를 문자로 드려			이벤트, 상품 서비스	이벤트/프로모션, 서비스/혜택, 문자/메시지	
		1	1	16			A			네, 그러세요.			
		1	0	17		방송상품주문의사확인	Q	네, 감사합니다. 현재 방송 중인 레드캠프 트			방송, 레드캠프 트레이닝 육중	레드캠프 트레이닝 육중 세트/상품명, 사은 구	
		1	1	18			A			네,			
		1	0	19		주문요청상품사이즈선택요청	Q			사이즈 확인 부탁드립니다.	사이즈, 확인	사이즈/치수	
		1	1	20	주문요청상품사이즈선택		Q	남성용으로요.			남성용		
		1	0	21		주문요청상품정보안내및사이즈	Q		어, 해당 상품은 남녀 공용입니다. 사이즈 일		해당, 상품, 남녀 공	사이즈/치수	
		1	1	22	주문요청상품사이즈선택		Q	그럼 구십오로 해야되나.			구십오	구십오/사이즈	
		1	0	23		주문요청상품정보안내및사이즈	Q		우선은 정 사이즈로 나왔고 남성분이시면		정 사이즈, 남성분,	사이즈/치수	
		1	1	24	주문요청상품사이즈선택		Q	음, 백으로 주세요.			백	백/사이즈	
		1	0	25		주문요청상품사이즈확인및카드	Q		네, 백 사이즈 선택하셨고요. 카드 결제로 도		백 사이즈, 선택, 카	사이즈/카	카드/결제방법
		1	1	26	무통장입금선택		Q	아니요. 무통장입금으로 제가 현금 넣어 드릴게요.			무통장입금, 현금		무통장입금/결제방법, 현금/결제방법
		1	0	27		무통장입금은행문의	Q	네, 그럼 어느 은행이 편하십니까?			은행		

●○ 데이터 구축 과정

데이터 분류	데이터 출처
K쇼핑	상담데이터 녹취 및 분석을 진행하는 Daisy플랫폼 서버에서 다운로드
금융/보험	개인정보 제공 이슈로 원천데이터 획득 제한, 직접 전화문의
질병관리본부	개인정보 제공 이슈로 원천데이터 획득 제한, 직접 전화문의
다산콜센터	개인정보 제공 이슈로 원천데이터 획득 제한, 직접 전화문의

상단의 표와 같은 출처에서 데이터를 공급받아 다음과 같이 정제 및 가공 작업을 진행함.



- 정제작업
  - 쇼핑데이터 구축은 kth의 자체 상담 데이터 분석 플랫폼인 Daisy를 이용해 개인정보 자동 마스킹 및 카테고리 분류를 진행하였고 보다 정확한 데이터 구축을 위해 용역을 고용하여 2차 마스킹 작업을 통해 개인정보 이슈 해결
  - 텍스트로 변환된 상담 데이터 중 상담사들 간의 음성이 혼재되어 식별할 수 없는 데이터에 대해서는 작업을 제외함
  - 금융, 질병관리본부, 다산콜센터 녹취 데이터의 경우, 유료 STT엔진을 이용하여 텍스트로 변환한 뒤, 수동 카테고리 분류 후 작업과정에서 학습을 통하여 자동 분류 실시 이후 문의한 상담 내용 중 개인정보를 선별 및 비식별화(마스킹) 처리 하여 개인정보 이슈 해결

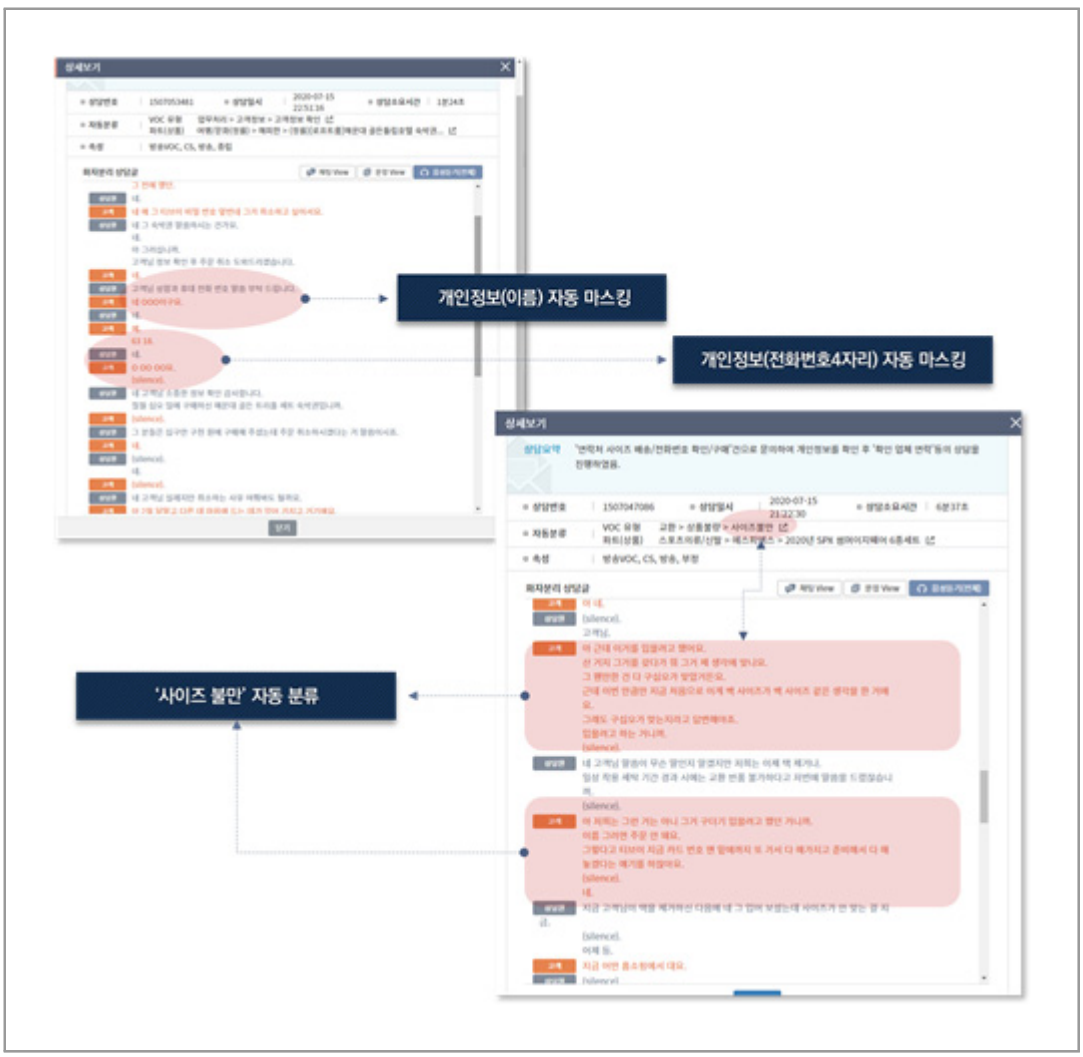


그림5 | Daisy 엔진 사용 예시

- 가공 작업
  - 상담 데이터의 문장별 분석을 통한 개체명 추출, 용어사전, 지식베이스 관리 진행
  - 작업자별로 입력 및 검수 요청한 데이터를 한꺼번에 볼 수 있는 플랫폼 사용
  - 작업이 어려운 데이터에 한해 작업 제외 함.
  - 목표량 :질문-답변으로 이루어진 데이터셋 110만쌍

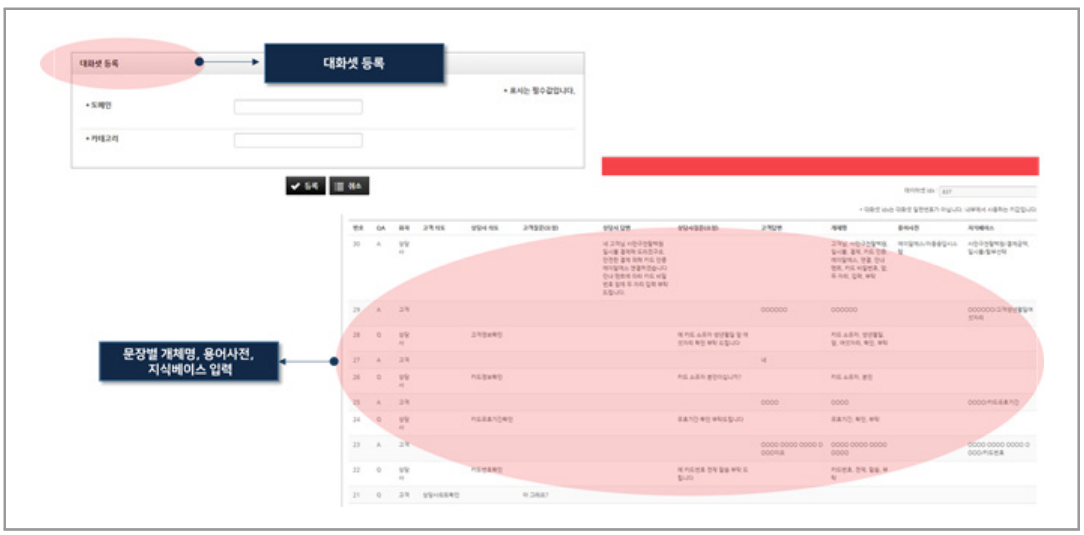


그림6 | 저작도구 사용툴

●○ 검수와 품질 확보

- 품질을 위한 교육
  - 데이터 구축 초기에 인공지능 데이터 구축 가이드라인을 만들어 크라우드소싱 인력을 교육함
  - 이 외에도 동영상 교육, 화상 교육을 통해 입력자, 전수검수자들에게 가이드라인을 제시
  - 검수과정을 통해서도 반려사유를 알려줌으로써 교육이 이루어짐

그림7 | 동영상 설명서

- 품질을 위한 교육
  - 전수 검수
    - 입력된 대화쌍이 가이드라인에 맞게 구축되었는지, 띄어쓰기 및 맞춤법이 틀린 부분이 있는지, 체크 후 반려
  - 선별 2차 검수
    - 검수된 질의 응답쌍이 작성 기준에 맞는지 체크 후 반려 및 수정 제출

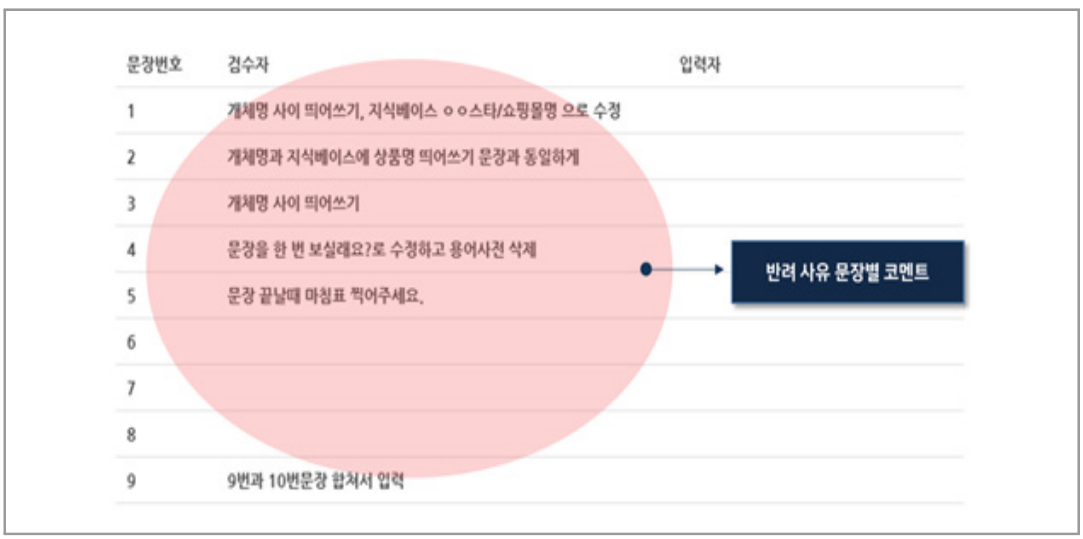


그림8 | 반려사유 코멘트

  - 샘플링된 데이터 셋 검토
    - 구축된 데이터를 초기부터 일정 기간마다 샘플링하여 주관기관으로 전송하여 구축 가이드와 학습에 유효한 데이터인지 검토 받음
    - 검토사항을 데이터 구축에 적용함

●○ 데이터 구축 담당자

수행기관(주관) : 케이티하이텔주식회사
(전화 : 02-2602-3289)