개요: 한국어대화 데이터셋이란?

한국어대화 데이터셋은 국내 소상공인 9종과 공공 민원 분야를 선정하여 챗봇에서 활용할 수 있는 한국어 대화를 약 1만건 이상을 구축하였다. 한국어 대화(Dialog) 10,000건을 수집하여 100,000문장과 도메인 및 카테고리 정보, 의도, 어휘 등을 포함하여 구축하였다.

챗봇 서비스는 공공분야와 민간 대기업을 중심으로 활성화되고 있으나 챗봇 서비스의 핵심인 대화 데이터가 공개되지 않아, 일반인 및 소상공인 등 일반 수요자를 위한 양질의 대화 데이터 제공이 요구된다. 한국어대화 데이터셋은 소상공인이나 공공 민원 분야에서 챗봇을 개발하기 위해 소요되는 데이터 구축 비용과 노력을 절감하여 챗봇 서비스 확보를 위한 진입장벽을 낮추고, 다양한 인공지능 서비스와 연계하여 활용될 수 있다.

데이터셋의 구성

데이터셋은 대화 데이터, 의도목록, 용어사전, 그리고 지식베이스 총 4가지로 구성되어있다. 대화 데이터는 Q&A형태의 손님과 점원의 대화로 문장 수를 기준으로 공공 민원 분야 11,082건과 소상공인 분야 90,413건 총 101,495건 데이터를 구축하였다. 의도목록은 시나리오를 기반으로 업종별 세부상황을 도출하여 대화 데이터의 의도 정보 태깅을 위해 구축하였다. 향후 인공지능 서비스에서 다양하고 구체적인 상황의 대화 주제를 판별하는데 활용·확장이 가능하다. 또한, 각 도메인 및 카테고리별로 동의어에 대한 용어사전(thesaurus)을 구축하여 대표어(headword)와 하위어(word)의 관계를 정의하여 활용할 수 있도록 하였고 업종 분류체계, 의도정보, 개체명(Entity) 정보, 동의어 정보 항목이 포함된 지식베이스를 구축하여 챗봇의 응답에 활용할 수 있도록 구축하였다.

데이터 종류	포함 내용	제공 방식
대화 데이터	질문과 답(10만 건)	EXCEL, JSON 포맷 파일
의도 목록	본문에서 답을 찾을 수 없는 질문(3천 건)	JSON 포맷 파일
용어사전	동의어 사전(3천 건)	JSON 포맷 파일
지식베이스	업종 및 의도 분류체계, 개체명, 동의어 등 (1만4천건)	JSON 포맷 파일

데이터셋의 설계 기준과 분포

대화데이터를 구축하기 위해 업종별 데이터 밸런스와 실제 대화데이터 수집을 고려하였다.
소상공인시장진흥재단의 소상공인 업종 분류체계를 기반으로 대화데이터 구축을 위한 소상공인 9개 분야를 선정하였고 업종 종사자 면담 및 관찰, 영업장 관찰 등 실제 고객 응대 상황 조사와 관련 업종 경험자 채용, 소상공인 감수위원을 활용한 의견 수렴을 통해 주요 대화 주제를 선정하였다. 공공 분야는 시정 민원 상담 중 차량등록, 여권, 대중교통 및 교통, 상수도 관련 4개의 세부 분야에 대하여 대화데이터를 수집하고

구축하였다. 데이터 수집은 전문 시나리오 작가와 업종 전문가로 구성된 팀을 통해 일관된 기준 하에 작성된 시나리오를 기반으로 현장 음성 녹취, 온/오프라인 FAQ 및 CS 교육 매뉴얼 수집, 크라우드 소싱 방법을 활용하였다.

그림 1 데이터셋 구성 개요

소상공인 분야의 대화 특성상 점원이 특정한 업무를 달성할 수 있는 대화 시스템이 요구되므로 대화 데이터는 이에 기반하여 목적 지향 대화 시스템으로 설계되었다. 목적 지향 대화 시스템은 사용자의 입력이 어떤 의도인가를 파악하고, 이해한 사용자 의도에 따라 시스템 응답 또는 행동을 하는 것을 말한다. 만약 사용자 입력이 '필수'임에도 값이 없을 경우, 이를 입력받기 위한 '되문기 질문(Slot-Filling)'을 사용하여 사용자와 대화를 시도하도록 설계하였다.

소상공인 9종과 공공 민원 1분야가 균등하도록 구축하기 위하여, 각 업종별 하나의 대화 주제를 대상으로 대화를 나눌 때 최소 10문장 이상 대화가 이루어지도록 구축하였다.

구분	소상공인								
업종/분야	일반 음식점	의복/ 의류	학원	종합 소매점	생활 서비스	카페	숙박업	여가/ 오락	부동산업
구축량	15,726	15,826	4,773	14,949	11,087	7,859	7,113	4,949	8,131

합계	90,413

구분	공공 민원			
업종/분야	차량등록	상수도	여권	교통
구축량	7,490	1,452	806	1,334
합계	11,082			

데이터 구조

대화데이터의 기본 구조는 Q&A(질의/응답)로 구성되며 각 문장은 사용자(손님) 질문(Main Question), 메인 질문에 추가적으로 필요한 시스템(점원)의 서브 질문(Sub Question), 서브 질문에 대한 사용자(손님) 응답(User Answer), 시스템(점원) 최종 응답(System Answer)로 구성된다.

유형/화자	손님	점원
질문(요청)	Main Question	Sub-question
답변(응답)	User Answer	System Answer

각각의 질문(Q)는 의도 정보를 포함하고 있으며, 화자 구분에 따라 메인 의도(사용자가 점원에게 질문, 요청하는 내용), 하위 의도(메인 의도를 처리하기 위해 점원이 사용자에게 정보를 요구하는 질문)으로 구분된다.

의도정보		문장			
메인	하위	메인 질문	서브 질문	사용자 답변	시스템 답변
주문		아메리카노 한 잔 주세요			
	온도		따뜻한 걸로 드릴까요?		
				아이스로 주세요	
	사이즈		사이즈는 어떻게 하시겠어요?		
				레귤러로 주세요	
가격		얼마예요?			
					4500원입니다.
	결제 방식		결제는 어떻게 하시겠어요?		
				현금이요	
	현금 영수증		현금영수증 해드릴까요?		
				네	

	현금 영수증		현금영수증 번호 어떻게 되나요?		
				010-1234-1234입니다	

데이터 예시

한국어대화 데이터셋은 아래 예시와 같이 대화데이터, 의도목록, 용어사전, 지식베이스 등으로 구축되었다.

1) 화자 분리: 1-손님, 0-점원

2) 의도정보: main intent와 sub intent1, 2로 구분되며, 하나의 문장은 하나의 의도로 태깅(두 개 이상의 의도인 경우 대표 의도로 태깅).

3) 원본데이터: Q는 mainQ와 mainQ를 실행하기 위한 subQ로 구분, A는 user answer와 system answer (챗봇의 response)로 구분

4) 개체명 추출: 사용자의 의도가 반영된 키워드로 고객과 점원의 발화 문장에서 추출, 고유명사와 복합명사(인명, 지명, 기관명, 업소명, 메뉴명 등), 수식표현 등으로 구성됨

5) 용어사전: 개체명 추출을 통해 가져온 용어/대표순으로 등록

6)지식베이스: 개체명 대표어와 그 개체명이 소속된 카테고리로 구분

대화데이터 SAMPLE

대화내용(수정한 부분) (띄어쓰기교정)	Data - ID	도메인 ID	도메인	카테고리	화자	문장번호	의도정보		QA	질답연계	원본				개체명	용어사전	지식베이스
							main	sub			mainQ	subQ	User Answer	System Answer (Response)			
혹시 예약되나요?	16333324	A	음식점	홀서빙 음식점	1	1	예약 문의		Q		혹시 예약되나요?				예약		
네 가능합니다	16333324	A	음식점	홀서빙 음식점	0	2	예약 문의		A					네 가능합니다			
이번 주 토요일에 예약하려고 하는데요	16333324	A	음식점	홀서빙 음식점	1	3	예약 문의		Q		이번 주 토요일에 예약하려고 하는데요				이번 주, 토요일, 예약		이번 주 /예약일 토요일/예약일
네 몇 시로 예약해 드릴까요?	16333324	A	음식점	홀서빙 음식점	0	4	예약 문의	예약시간	Q			네 몇 시로 예약해 드릴까요?			시, 예약		
오후 6시로 해주세요	16333324	A	음식점	홀서빙 음식점	1	5	예약 문의	예약시간	A				오후 6시로 해주세요		오후 6시		오후 6시/시간
네 몇 분 오시나요?	16333324	A	음식점	홀서빙 음식점	0	6	예약 문의	인원	Q			네 몇 분 오시나요?			분		명/인원
4명 예약해주세요	16333324	A	음식점	홀서빙 음식점	1	7	예약 문의	인원	A				4명 예약해주세요		4명, 예약		
네 이번 주 토요일 오후 6시 4분 맞으시죠?	16333324	A	음식점	홀서빙 음식점	0	8	예약 문의	예약확인	Q			네 이번 주 토요일 오후 6시 4분 맞으시죠?			이번 주, 토요일, 오후 6시 4분		이번 주/예약일 토요일/예약일 오후 6시/시간
네 맞아요	16333324	A	음식점	홀서빙 음식점	1	9	예약 문의	예약확인	A				네 맞아요				
네 예약되셨습니다	16333324	A	음식점	홀서빙 음식점	0	10	예약 문의		A	3				네 예약되셨습니다	예약		

그림 2 대화데이터 예시

의도목록 SAMPLE

도메인	카테고리	번호	Main Intent	Sub Intent	example
카페	카페	1	일반 주문	사이즈	카페라떼 컵 사이즈는 뭘로 드릴까요?
			일반 주문	컵 종류	그럼 머그잔으로 제공해드려도 괜찮을까요?
			일반 주문	수량	네 몇 잔 드릴까요?
			일반 주문	제조 시간	지금 주문이 밀려있어서 10분 정도 기다리셔야 하는데 괜찮으세요?
			일반 주문	주문 확인	사이즈 업 아이스 아메리카노에 샷 추가 맞으시죠?
			일반 주문	음료 온도	핫으로 드릴까요 아이스로 드릴까요?
			일반 주문	원두 선택	신맛과 고소한 맛 원두 중에 어떤 걸로 드릴까요?
			일반 주문	토핑	휘핑크림 올려드릴까요?
			일반 주문	픽업 안내	진동벨 울리면 받으러 오세요
			일반 주문	커팅	와플은 잘라드릴까요?
			일반 주문	포장 유무	가지고 가시나요?
		2	주문 내용 변경		아메리카노를 카페모카로 변경 가능한가요?
			주문 내용 변경	추가 금액 안내	가능합니다만 가격이 천원 더 올라가게 됩니다
		3	메뉴 문의		차 종류는 어떤 게 있나요?
		4	메뉴 추천 요구		생크림 케이크는 뭐가 제일 맛있어요?
			메뉴 추천 요구	음료 당도	달달한 음료 찾으세요?
		5	메뉴 추가 문의		혹시 한잔 더 추가되나요?
		6	옵션 추가 요구		
			옵션 추가 요구	추가 금액 안내	네 500원 추가됩니다
		7	가격 문의		
		8	제조 시간 문의		네 몇 분 정도 걸릴까요?
		9	단체 주문 문의	예약 시간	단체 주문 가능한가요?
		10	메뉴판 요구		혹시 메뉴판 주실 수 있나요?
		11	리필 문의		아메리카노 리필되나요?
		12	주문 취소		그렇게 기다릴 수가 없어서 주문취소 할게요 주문취소 가능한가요?
...	...	...	...	...	...

그림 1 의도목록 데이터 예시

용어사전 SAMPLE

도메인	카테고리	단어	Headword
A 음식점	공통	금주	이번주
		이번 주	이번주
		분	명
		성함	이름
		일행분	일행
		말씀	말
		메뉴판	메뉴
		음료 할인권	음료할인권
		인	명
		창가	창가자리
		베이비체어	유아용의자
		주차비용	주차비
		공짜	무료
		이번주	이번주
		창가석	창가자리
		국물 음식	국물음식
		저녁	오후
		킬로그램	kg
		킬로그람	kg
		영업 중	영업중
		추천 메뉴	추천메뉴
		접시	그릇
		카드 결제	카드결제
		일	일행

지식베이스 SAMPLE

도메인	카테고리	Main Intent	Sub Intent	Entity						
A 음식점	홀서빙음식점	예약문의	시간	오후 6시	3시	20분	오후 10시	5분	30분	오전 8시
	홀서빙음식점	예약문의	인원	7명	5명	4명	10인	6명	16명	3명
	홀서빙음식점	빈 자리 문의	인원	7명	5명	4명	10인	6명	16명	3명
	홀서빙음식점	예약 문의	이름	김철수	홍길동					
	홀서빙음식점	예약 자리 문의	이름	김철수	홍길동					
	홀서빙음식점	식사 주문	음료	콜라	커피	스프라이트	사이다	탄산음료	마운틴듀	와인
	홀서빙음식점	주문 변경	음료	콜라	커피	스프라이트	사이다	탄산음료	마운틴듀	와인
	셀프서비스음식점	메뉴 주문	음료	콜라	커피	스프라이트	사이다	탄산음료	마운틴듀	와인
	배달음식점	식사 배달 요청	음료	콜라	커피	스프라이트	사이다	탄산음료	마운틴듀	와인
	홀서빙음식점	식사 주문	메뉴	참치 스페셜 세트	a세트	a세트	음료	세트메뉴 1번	이벤트 런치 세트 4인	런치세트
	셀프서비스음식점	메뉴 주문 문의	메뉴	참치 스페셜 세트	a세트	a세트	음료	세트메뉴 1번	이벤트 런치 세트 4인	런치세트
	배달음식점	배달시간 문의	메뉴	참치 스페셜 세트	a세트	a세트	음료	세트메뉴 1번	이벤트 런치 세트 4인	런치세트
	배달음식점	식사 배달 요청	메뉴	참치 스페셜 세트	a세트	a세트	음료	세트메뉴 1번	이벤트 런치 세트 4인	런치세트
	배달음식점	배달 가능 문의	메뉴	참치 스페셜 세트	a세트	a세트	음료	세트메뉴 1번	이벤트 런치 세트 4인	런치세트
	홀서빙음식점	예약 문의	연락처	02-1234-5678	010-1234-5678					
	홀서빙음식점	예약 문의	날짜	지난주	오늘	내일	당일	이번주	이번주	2월 14일
	셀프서비스음식점	케첩/소스 요구	수량	1	2잔	7개	4개	2개	1개	3개
	셀프서비스음식점	음료 주문	결제	카드결제	분할결제	신용카드	삼성페이	쿠폰	쿠폰결제	계좌이체
	배달음식점	식사 배달 요청	결제	카드결제	분할결제	신용카드	삼성페이	쿠폰	쿠폰결제	계좌이체

그림 4 용어사전 및 지식베이스 예시

데이터 구축 과정

데이터 구축은 2018년 10월부터 2019년 3월까지 1만건의 데이터를 수집하여 10만건의 대화데이터를 구축하였다. 실제 녹취된 음성데이터의 경우 음성인식 STT엔진을 통해 전사 후 화자/문장 분리 및 형태소분석, 메타데이터 정의 등의 데이터 정제를 거쳐 원천데이터 데이터베이스에 저장하였다.

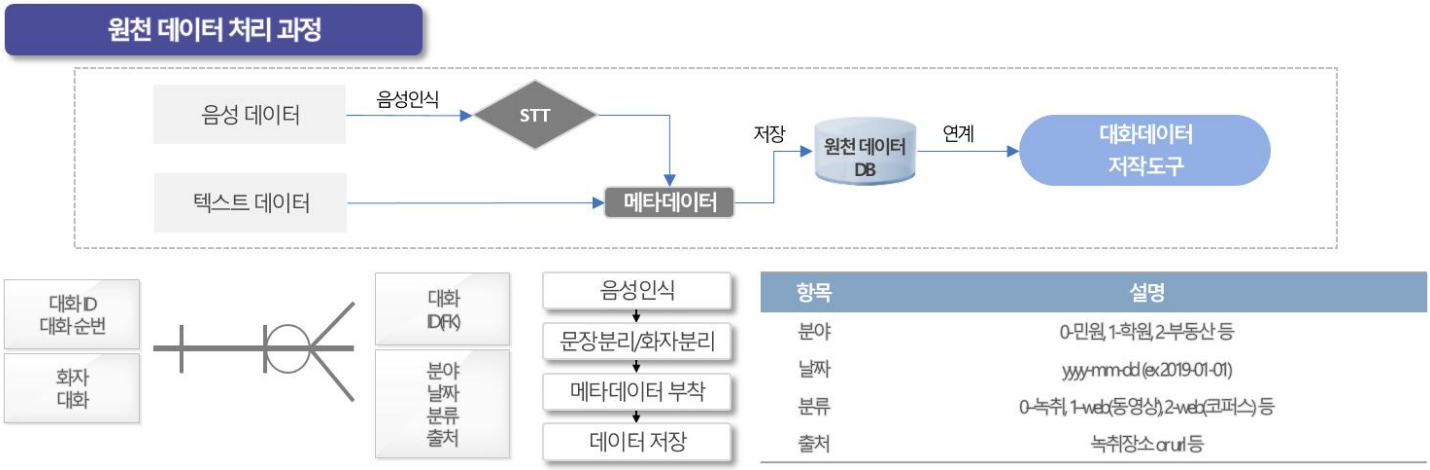
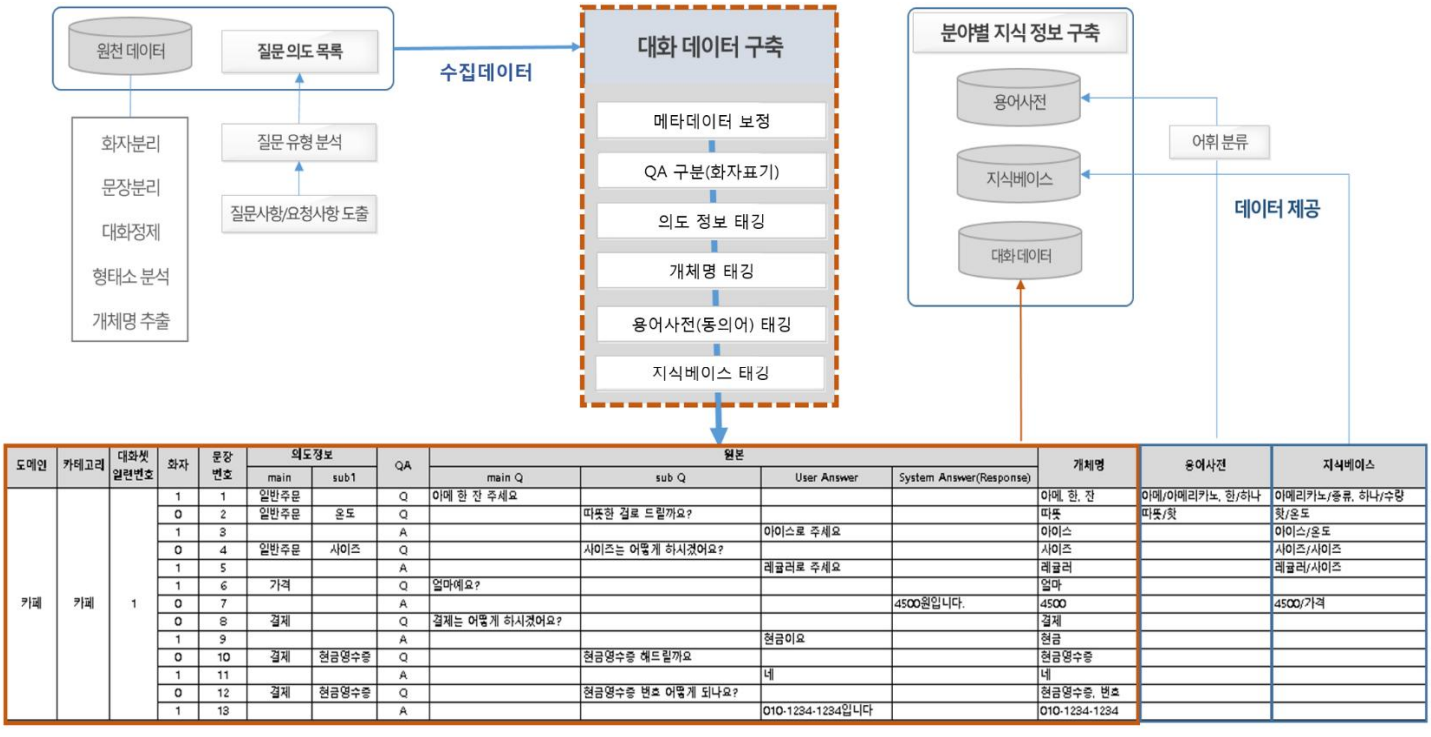


그림 5 원천데이터 처리과정

저장된 원천데이터들은 질문의도목록을 참조하여 의도정보 및 개체명, 동의어 등을 태깅하여 대화 데이터셋으로 구축되었고 개체명과 도메인 및 카테고리 분류체계는 용어사전과 지식베이스에 저장하였다. 대화데이터의 문장을 구축할 때 화자의 발화문장은 완성형 문장으로 구축하였으며 손님과 점원의 발화를 전사한 것으로 원본(raw 대화데이터)의 대화 의도를 최대한 유지할 수 있도록 구어체 표현과 사투리 표현을 그대로 유지하였다.



도메인	카테고리	대화셋 일련번호	화자	문장 번호	의도정보		QA	원본				개체명	용어사전	지식베이스
					main	sub1		main Q	sub Q	User Answer	System Answer(Response)			
카페	카페	1	1	1	일반주문		Q	이메 한 잔 주세요				이메, 한, 잔	이메/이메리카노, 한/하나	이메리카노/종류, 하나/수량
			0	2	일반주문	온도	Q		따뜻한 걸로 드릴까요?			따뜻	따뜻/핫	핫/온도
			1	3			A			아이스로 주세요		아이스		아이스/온도
			0	4	일반주문	사이즈	Q		사이즈는 어떻게 하시겠어요?			사이즈		사이즈/사이즈
			1	5			A			레귤러로 주세요		레귤러		레귤러/사이즈
			1	6	가격		Q	얼마예요?				얼마		
			0	7			A				4500원입니다.	4500		4500/가격
			0	8	결제		Q	결제는 어떻게 하시겠어요?				결제		
			1	9			A			현금이요		현금		
			0	10	결제	현금영수증	Q		현금영수증 해드릴까요			현금영수증		
			1	11			A			네		네		
			0	12	결제	현금영수증	Q		현금영수증 번호 어떻게 되나요?			현금영수증, 번호		
			1	13			A			010-1234-1234입니다		010-1234-1234		

그림 6 대화데이터 수집 및 데이터 정제

그림 7 데이터 구축을 위한 저작도구

일관된 데이터 구축과 구축한 데이터에 대한 효율적인 관리를 위하여 대화데이터 저작도구를 개발하여 데이터구축에 활용하였다. 저작도구는 대화 데이터 목록 검색 및 조회, 데이터 필터링이 가능하고 텍스트 데이터에 대한 화자/문장 분리 결과 검수/보정 및 형태소 분석결과를 확인할 수 있다. 또한, 문장별로 의도목록을 참조하여 의도정보를 입력할 수 있고 용어사전 및 지식베이스도 활용할 수 있다. 구축된 대화 데이터셋은 XML/JSON 형태로 업로드 및 다운로드 할 수 있으며, 데이터 시각화를 통해 도메인 별 통계 정보 및 데이터 트리뷰 등을 제공받을 수 있다.

검수와 품질 확보

구축한 데이터셋에 대한 검수와 품질 확보를 위하여 교차 품질 검수 체계 및 가이드라인을 수립하여 아래와 같이 정합성 검수, 내용적 검수, 정량적 검수를 수행하였다.

정합성 검수	- 문장 교차 검수 - 구조적 오류 검수 - 누락값 보정 - 맞춤법 검사
내용적 검수	- 다중 작업자를 통한 의도 정보 태깅 결과 확인 - 구분 및 통합이 필요한 의도 목록 도출 후 의도 목록 수정 및 보완 - 용어 사전 및 지식베이스 구축 정보 확인
정량적 검수	- 기구축 문장 수에 대한 정량적 목표 달성 여부 확인 - 의도별 문장수 검토 후 10문장 이하의 의도에 대한 추가 문장 구성

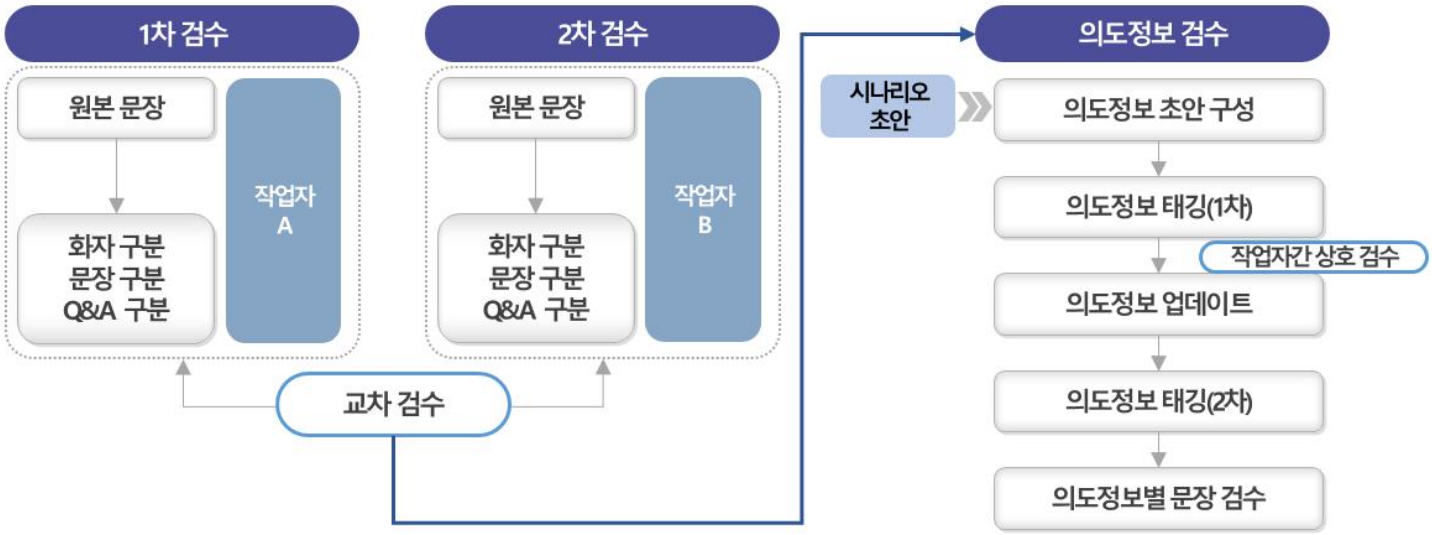


그림 8 품질 확보를 위한 교차 품질 검수 체계

1,2차 문장 교차 검수를 통해 원본 문장과 화자 및 문장, Q/A 구분 내용이 맞지 않는 경우 1차 완료된 작업에 대해 다른 작업자가 교차 재검토를 수행하였다. 의도정보 검수를 위하여 시나리오 초안에서 도출된 업종별 세부상황을 활용하여 의도정보 목록 초안 작성 후 의도정보 태깅(1차) 및 검수하였고 의도별로 대화 데이터 구성하여 다중 작업자들을 통해 의도에 매치가 안 되는 문장, 구분 및 통합이 필요한 의도 목록을 도출하였다. 이를 기반으로, 의도정보 목록 수정 및 보완, 업데이트하여 문장별 의도정보 태깅(2차) 및 검수하였고10개 문장 이하로 구성된 의도는 추가 문장 입력을 통해 의도별로 최소 10문장 이상의 문장 수를 확보하였다.

그 외에 대화데이터의 구조적 오류 검수 및 데이터 구축 과정에서 발생한 누락 값(null) 보정을 진행하였고 지식베이스 및 용어사전 태깅 정보 정리 및 중복 데이터를 제거하여 데이터셋의 품질을 확보하였다.

데이터 구축 담당자

수행기관(주관) : 한국과학기술정보연구원