●○ 상품 이미지 및 고객 주문질의 응답 데이터 과제

소상공인 고객 주문 질의-응답 데이터

●○ 개요: 고객 주문 질의응답 데이터셋이란?

자연어처리(NLP, Natural Language Processing) 기술이 발전하면서 비용효율적으로 24시간 365일 고객을 응대하는 서비스에 대한 수요가 늘고 있다. 고객을 응대하는 기술은 챗봇 기반 질의응답이나 자동응답 서비스로 구현하는 것이 보편적인 추세이다. 이러한 서비스의 기반기술인 인공지능 모델을 학습하는 용도인 고객 주문 질의응답 데이터셋을 롯데정보통신 컨소시엄이 구축했으며, 500만 건의 한국어 질문과 대답으로 구성되어 있다.

고객 주문 질의응답 데이터셋은 직접 인공지능 모델을 학습하는 데에는 물론 카카오 채널과 같이 상점/기업을 대상으로 한 기존 챗봇 서비스에도 활용 가능하다. 한국에서 인기가 높은 카카오톡 기반 챗봇 서비스(카카오 I 오픈빌더)와 더불어 네이버 CLOVA 챗봇, IBM의 Watson Assitant와 같은 서비스에서 고객 의도를 잘 파악하도록 인공지능을 훈련하는 데에 쓸 수 있다.

고객 응대 챗봇 기획자는 고객 주문 질의응답 데이터셋에서 상점 카테고리와 대화의도 등을 기준으로 데이터를 선별하여 인공지능 모델 학습 데이터로 구성하거나 상용 챗봇 서비스에 학습 데이터를 업로드하는 방식으로 사용하면 된다. (챗봇 서비스 별로 한번에 올릴 수 있는 학습 데이터 건수는 다양하다.)

고객 질의응답 챗봇 사례는 아래를 참고할 수 있다.

그림1 | 매장 위치 질문에 대답하는 챗봇 사례

주요 챗봇 서비스의 학습 데이터 관련 도움말은 아래와 같다.

- 카카오 i 오픈빌더:
  https://i.kakao.com/docs/tutorial-chatbot-key-features#머신러닝-발화-입력
- 네이버 CLOVA Chatbot:
  https://docs.ncloud.com/ko/chatbot/chatbot-3-2.html#대화-관리
- IBM Watson Assistance:
  https://cloud.ibm.com/docs/assistant?topic=assistant-intent-recommendations

●○ 데이터셋의 구성

본 데이터셋은 소상공인이 활용할 수 있는 주제로 질의-응답을 이루는 문장 500만건으로 이루어져 있으며, 각 문장 별로 대화순번, 화자구분, QA구분, 감성구분, 상점카테고리, 의도, 개체명 태깅을 부가하여 구성했다.

총 데이터셋 500만건은 인공지능 모델 학습용도로는 한번에 활용하지 못할 만큼 거대한 양으로서, 연구자나 챗봇 기획자는 상점 카테고리, 대화 의도 등으로 데이터를 선별하여 사용해야 적절하다.

이 데이터셋을 통해 고객이 질문하는 의도를 파악할 수 있으며 개체명 인식을 통해 보다 상세한 후속 액션을 지정하는 자동화가 가능하다. 또한, 챗봇 서비스와는 별개로 문맥을 파악하거나 개체명 인식기(NER, Named Entity Recognition)와 같이 범용적인 자연어처리(NLP, Natural Language Processing) 연구에도 유용하다.

●○ 데이터셋의 설계 기준과 분포

데이터셋을 설계할 때 가장 중요하게 고려했던 점은 가급적 여러 가지 소상공인 상점에서 활용할 수 있도록 하는 다양성 확보이다. 상점 카테고리 별로 질의-응답을 구성할 때 적절한 의도 기준을 만들었고, 데이터를 수집하면서 골고루 데이터가 분포되도록 모니터링하여 학습 시에 발생하곤 하는 데이터 편향성을 최소화하도록 했다.

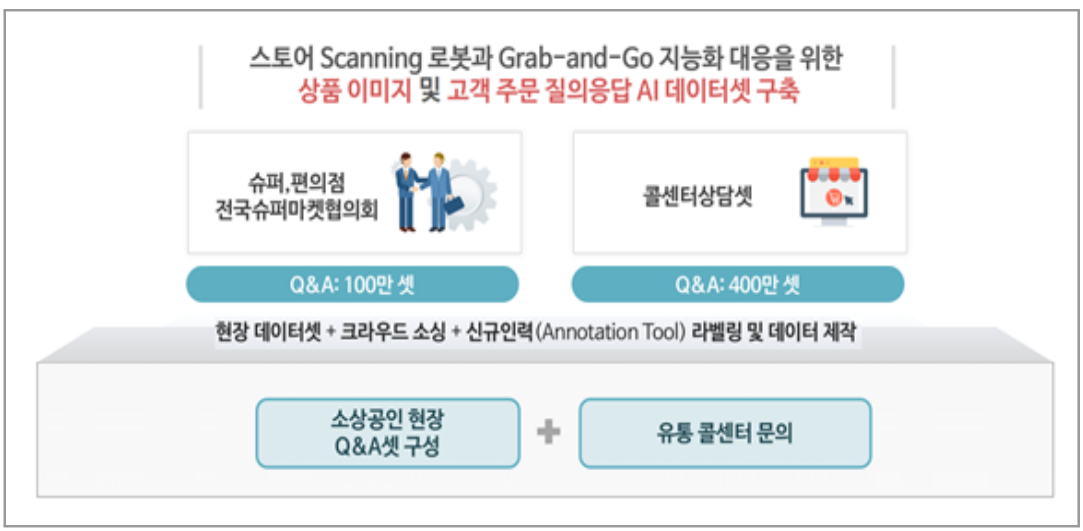


그림2 | 데이터셋 구성 개요

질의-응답의 구성 원칙과 주요 특징은 다음과 같다.

- 대화: 소상공인 상점 대화에서 100만 건, 콜센터 상담 녹취에서 400만 건의 데이터셋을 최종적으로 확보하고자 대화를 전사했다. 대화의 다양성을 확보하기 위해, 소상공인 협회를 통해 다양한 상점을 소개받았으며 콜센터 상담 데이터 역시 유통사 중에서도 소상공인 상점과 유관한 대화만 취합했다.

- 비식별화: 대화 중에 언급하는 개인정보, 민감정보는 1차적으로 보안서약서를 제출한 크라우드 소싱 인력이 비식별화 처리를 했고, 관련한 인공지능 모델 및 정규식 기반 프로그램을 통해 개인정보, 민감정보 유무를 판별했으며, 최종적으로 크라우드 소싱 인력이 비식별화 유무를 검수했다. 인공지능 학습 데이터로 사용해야 하므로 비식별화 후에는 '홍길동', '02-1234-1234', 롯데정보통신 주소와 같이 약정한 어휘로 대체했다.
- 의도 파악 및 활용: 실제 음성대화의 특성에 따라 질의와 응답이 한 문장으로 끝나지 않을 때가 상당히 많으므로, 대화 순번과 QA 구분 필드를 두어서 질의-응답 문장군을 식별하도록 했다. 또한 상점 카테고리 별로 의도 목록을 따로 두고 있으므로, 챗봇 서비스 기획자와 연구자는 목적에 부합하는 상점 카테고리와 의도 필드로 데이터를 선별하여 변용하기를 권장한다. 데이터셋을 엔터티(NER) 학습 용도로 활용할 때에도 의도 필드를 기준으로 데이터를 선별하는 게 적절하다.

●○ 데이터 구조

데이터셋은 아래와 같은 구조로 구성했다.

항목	형식	설명	비고
QA번호	수치	문장 순번	
대화순번	수치	대화 내 순번	
화자	문자(1)	상점(s) / 고객(c) 구분자	
QA여부	문자(1)	질의(q) / 응답(a) 구분자	
문장	텍스트	대화 내용	"가게 문 언제 여나요?"
감성	문자(1)	긍정(p) / 부정(n) / 중립(m) 구분자	
카테고리	텍스트	상점 구분	중화요리 음식점
의도	문자(1)	대화 전반 질의 의도	영업시간문의
개체명	텍스트	개체명 구분자 (예: 장소, 조직, 시간, 날짜)	
상담번호	수치	상담 전체내역을 포함하는 식별자	

※ 의도, 개체명 태그는 공란일 수 있다. (예: 언제 끝나요? - 개체명 없음 / 요즘 추운데요. - 유의미한 의도 없음)

●○ 데이터 예시

엑셀(CSV)을 감안한 데이터셋은 아래와 같으며, 수요에 따라 JSON 형식으로 제공하고자 한다.

QA 번호	대화 순번	화자	QA 여부	문장	감성	카테고리	의도	개체명	상담 번호
1	1	c	Q	자장면 얼마에요?	m	중화요리 음식점	가격문의		1
2	2	s	A	칠천원 입니다.	m	중화요리 음식점	가격문의		1
3	1	c	Q	가산에 엘디씨씨 배달 되나요?	m	중화요리 음식점	배달지역문의	가산: place 엘디씨씨: org	1
4	2	s	A	네.	m	중화요리 음식점	배달지역문의		1
5	1	s	Q	몇 시까지 갖다 드릴까요?	m	중화요리 음식점	배달기한문의		1
6	2	c	A	다섯시까지 갖다 주세요.	m	중화요리 음식점	배달기한문의	다섯시: time	1
7	3	s	A	네 갖다 드릴게요.	m	중화요리 음식점	배달기한문의		1

※ 고객과 상점 간 전체 상담(상담번호 기준) 내에 질의응답이 여러 차례 발생하는 현실을 반영했다.

●○ 데이터 구축 과정

데이터 구축은 2020년 10월부터 12월까지 소상공인 상점에서 100만 건, 콜센터 상담 이력에서 400만 건 확보를 목표로 신규 녹취한 후 전사하거나 기존에 전사한 텍스트 데이터를 가공하여 생성했다. 개인정보, 민감정보는 크라우드 소싱 인원이 비식별화한 내역을 개인민감정보 NER 및 정규식을 활용한 검수 프로그램과 크라우드 소싱 인원이 병행하여 검수했다.

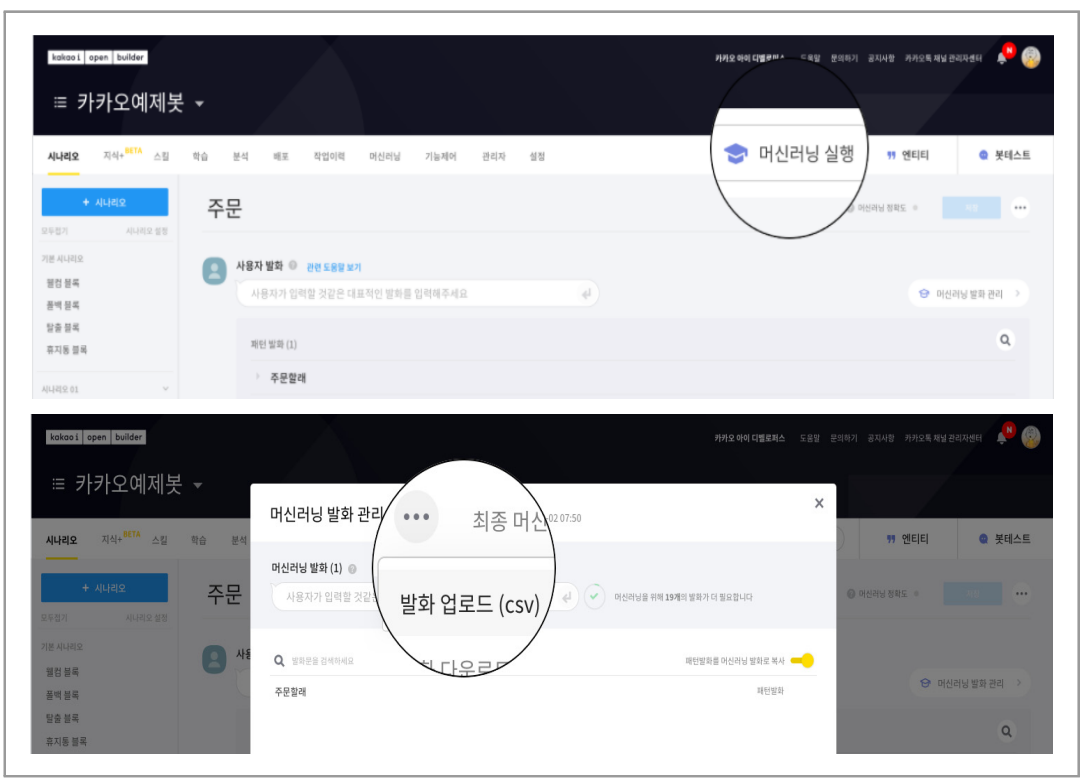
구축 절차는 아래와 같다.

그림3 | 데이터 수집, 비식별화, 가공, 검수 과정

그림4 | 데이터 비식별화 수행

위와 같이 확보한 상담 대화를 통해 500만 건의 질의-응답 데이터를 추출하여 감성, 개체명을 태깅했다. 소상공인 상담 대화의 특색에 의해 감성 태그 분포가 고르지는 못해도 의도를 대화 상황에 맞게 추출하는 데에는 효과적이다.

상용 챗봇 서비스를 이용한다면, 아래와 같이 의도 별 대화를 목적에 따라 분류하여 업로드하여 학습을 하게 할 수 있다. (예: 카카오톡 - 머신러닝 블록, 네이버 - 대화유형)



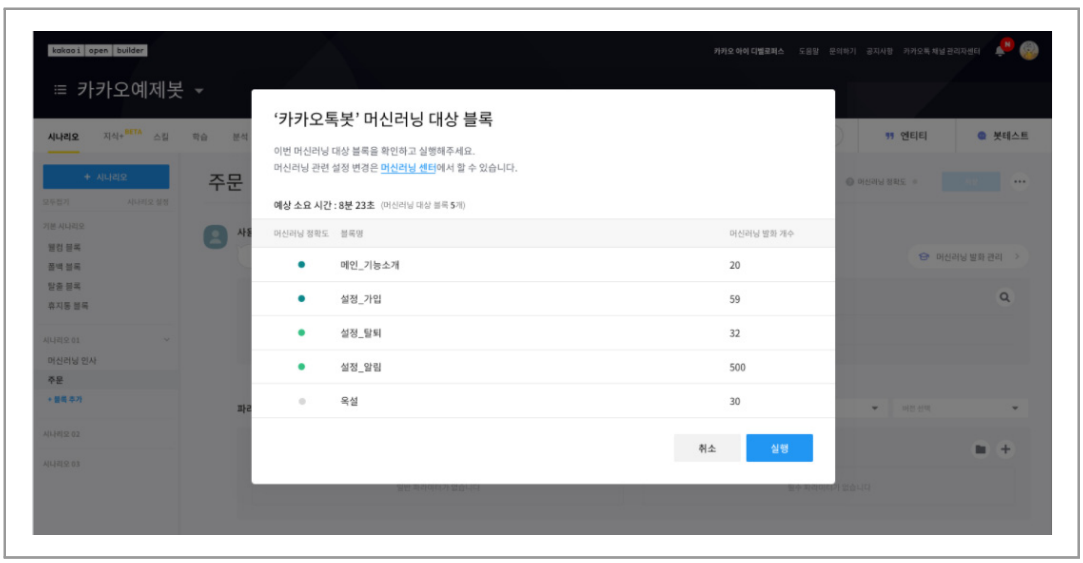


그림5 | 카카오톡 i 빌더 머신러닝 기능 (의도 별로 '머신러닝 블록'을 생성하여 CSV 파일로 학습 데이터 업로드)

실제로 고객에게 대응하는 챗봇 서비스를 만든다면, 상점 대화이므로 상품명을 인식하게 하므로, 모든 상품명을 개체명(챗봇 서비스에 따라 '엔티티'로도 표현한다.)으로 지정하는 작업이 필요하다. 해당 데이터셋에는 짜장면이나 짬뽕 같이 흔한 상품명은 개체명으로서 지정이 되어있지만, 상당수 상점은 별도로 상품명을 개체명(NER)으로 지정해야 서비스가 가능하다. 다행히 상용 챗봇 서비스는 아래와 같이 상품명을 지정할 수 있도록 기능을 제공하고 있다.

그림6 | 네이버 CLOVA Chatbot ('엔티티' 기능으로 상품명 개체를 입력 및 업로드하는 기능 제공)

검수와 품질 확보

대량의 데이터를 높은 품질로 생성하기 위하여, 단가는 낮지만 품질 관리가 어려운 크라우드소싱 방식의 데이터생성 작업을 보완하기 위한 검수 프로세스의 정립은 데이터셋 구축에 매우 중요한 의미를 갖는다. 이 데이터셋에서는 참여 업체 별로 검수를 진행하되, 외부 품질관리 검증 조직을 통해서도 검수를 진행했다.

크라우드 소싱 인력의 수작업 검수 외에는 아래와 같이 자동화 검수를 진행했다.

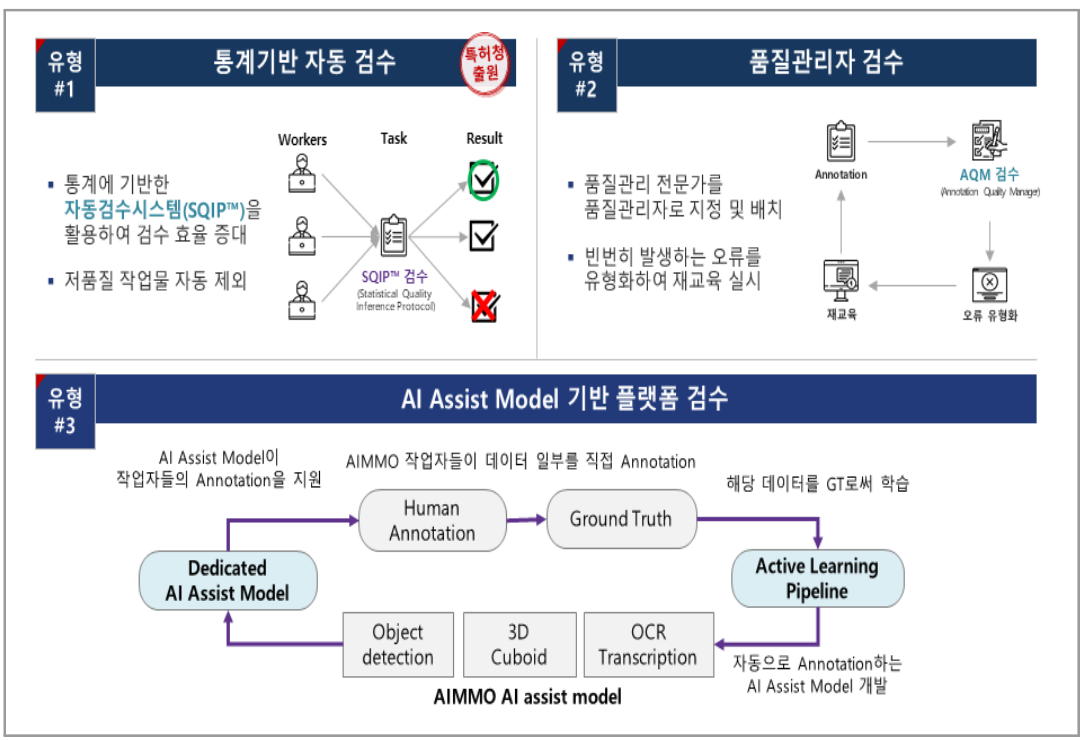


그림7 | 데이터 검수 유형 상세

전문 품질관리 조직과 검증된 품질관리 프로세스를 활용하여 단계 별로 검증을 수행했다.

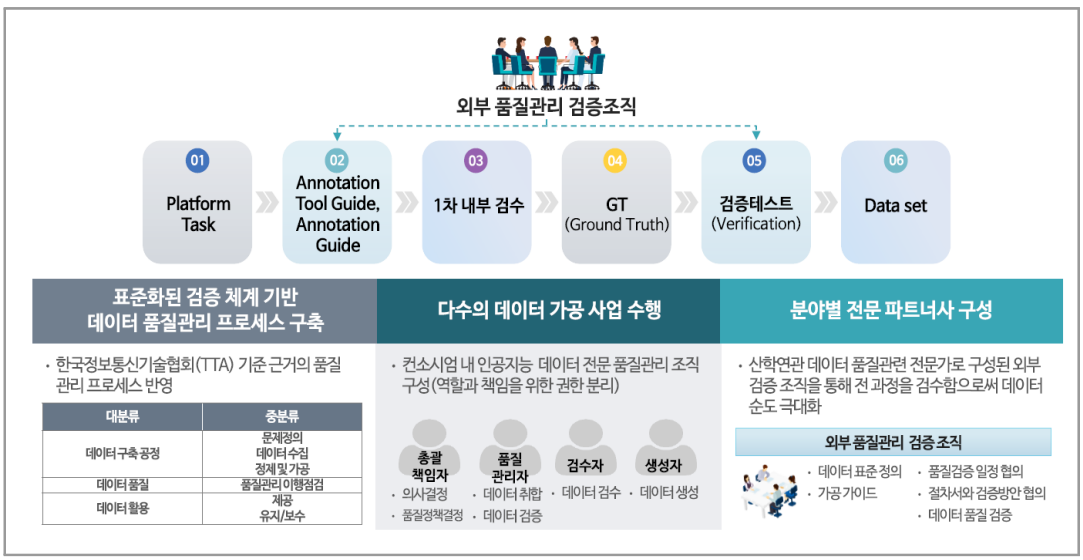


그림8 | 데이터 가공 과정 품질관리 검수 과정

●○ 데이터 구축 담당자

수행기관(컨소시움사) : ㈜롯데정보통신

(전화: 2626-4000, 이메일: aidata@lotte.net)