

한국어 학습자를 위한 상황별 존댓말 질의응답 데이터셋 구축

Construction of a Context-Specific Honorifics Question Answering Dataset for Korean Language Learners



존댓말 연구소: 허성은, 손규민, 김영호, 방민우, 박창혁, 임성원

1. 연구 배경 및 목적



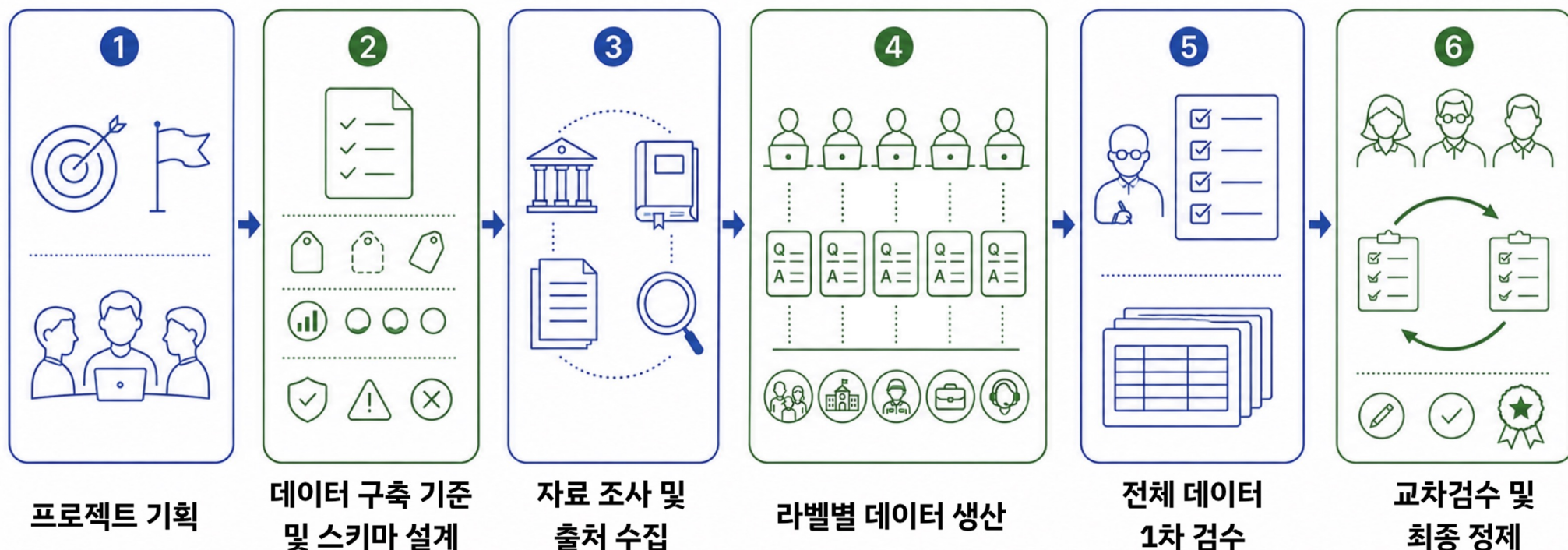
배경

- 존댓말은 상황과 인간관계에 따라 표현이 달라지는 한국어의 핵심 요소
- 그러나 기존 학습 자료는 실제 상황 중심 데이터가 부족

목적

- 한국어 중·고급 학습자를 대상으로 상황별 존댓말 질의응답(QA) 데이터셋을 구축
- 향후 한국어 교육을 위한 AI 시스템 학습에 활용

2. 데이터 구축 과정 및 설계 원칙



설계 원칙

- 화자·청자·관계·장소가 포함된 **1문항 1질문**
- 핵심 답 → 규범 → 실사용 → 예문** 순서로 답변 구조화
- 국립국어원·표준국어대사전·KCI 논문 등 공식·학술 출처 기록
- 작성자 외 검수자가 오류 태그와 사유를 기록하여 반려·수정

라벨 스키마

- 상황:** 가족·친족 · 학교·학원 · 군대 · 직장·비즈니스 · 서비스업
- 난이도:** 낮음(일상적 존대) · 보통(높임 일치·기본 압존법) · 높음(간접 높임·복합 맥락)
- 안전성:** 비속어·은어 · 개인정보 · 혐오 조장 · 편향 · Pass
- 오류 태그:** 높임법 · 상황/관계 · 내용 · 출처/근거 · 설명 품질 오류

구축 및 검수

- 5명의 작성자 × 200건 = 총 1,000건
- 1차: 전담 검수자 1인이 전체 1,000건 검수
- 2차: 작성자 5인이 다른 상황 라벨 200건씩 교차검수

[가족 · 친족] [난이도=보통] [Pass]

Q.
부모님이 편찮으실 때 “어디가 아프세요?”
대신 써야 할 높임 표현이 있나요?

일반 표현 → 권장 표현
“어디가 아프세요?” → “어디 편찮으세요?”

A.
‘편찮으시다’는 ‘아프다’의 높임말로, 웃어른에게 더 예의 바른 표현이다.
다만 ‘아프세요’도 문법적으로 틀린 표현은 아니다.
예: “어머니, 요즘 어디 편찮으신 데는 없으세요?”

출처: 표준국어대사전

3. 데이터셋 구성 및 통계

1,000건

상황별 존댓말
QA

5개

상황 라벨

200건

라벨당 데이터

100%

출처 기록률

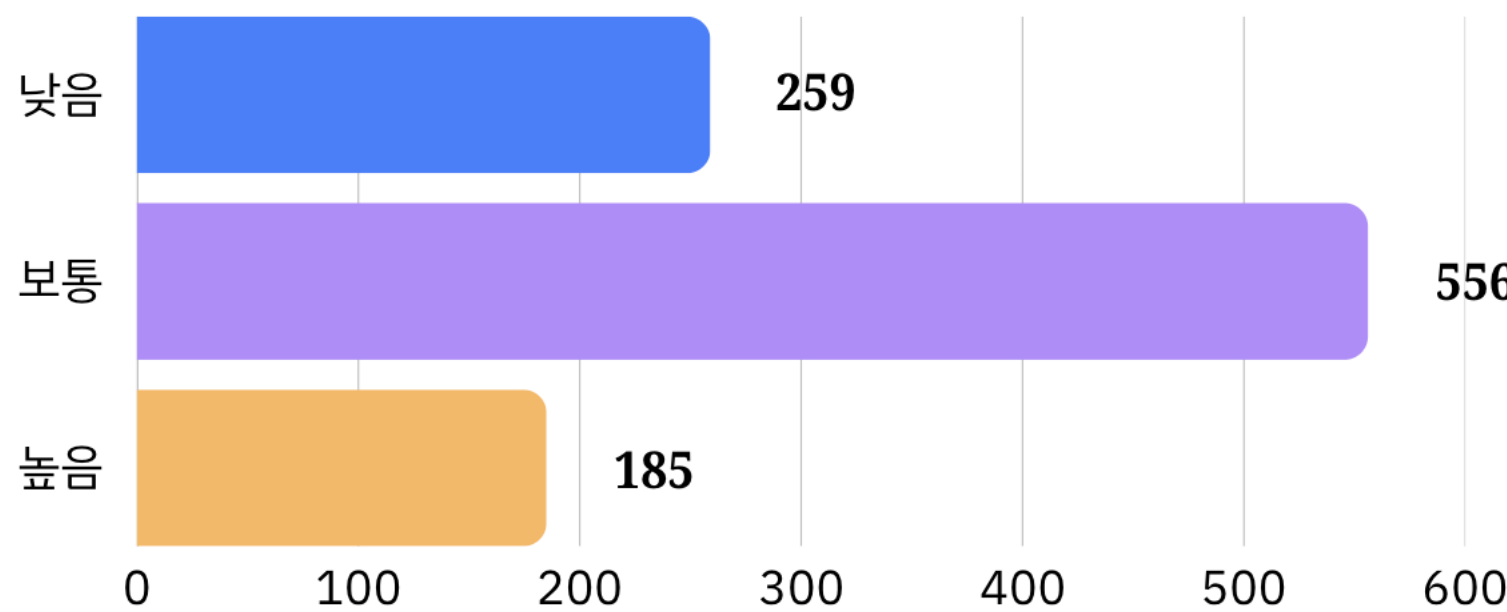
데이터 스키마

구분	주요 필드	역할
기본 정보	ID, 언어, 상황 라벨	문항 식별 및 상황 구분
학습 정보	난이도, 질문, 답변	학습에 활용되는 핵심 내용
근거 정보	출처명, URL, 확인일, 권리 상태	답변 근거와 이용 조건 기록
품질 정보	안전성, 오류 태그, 검수 상태	품질 및 위험 요소 관리
작업 정보	작성자, 검수자, 수정 이력	구축 과정 추적

상황 라벨별 구성

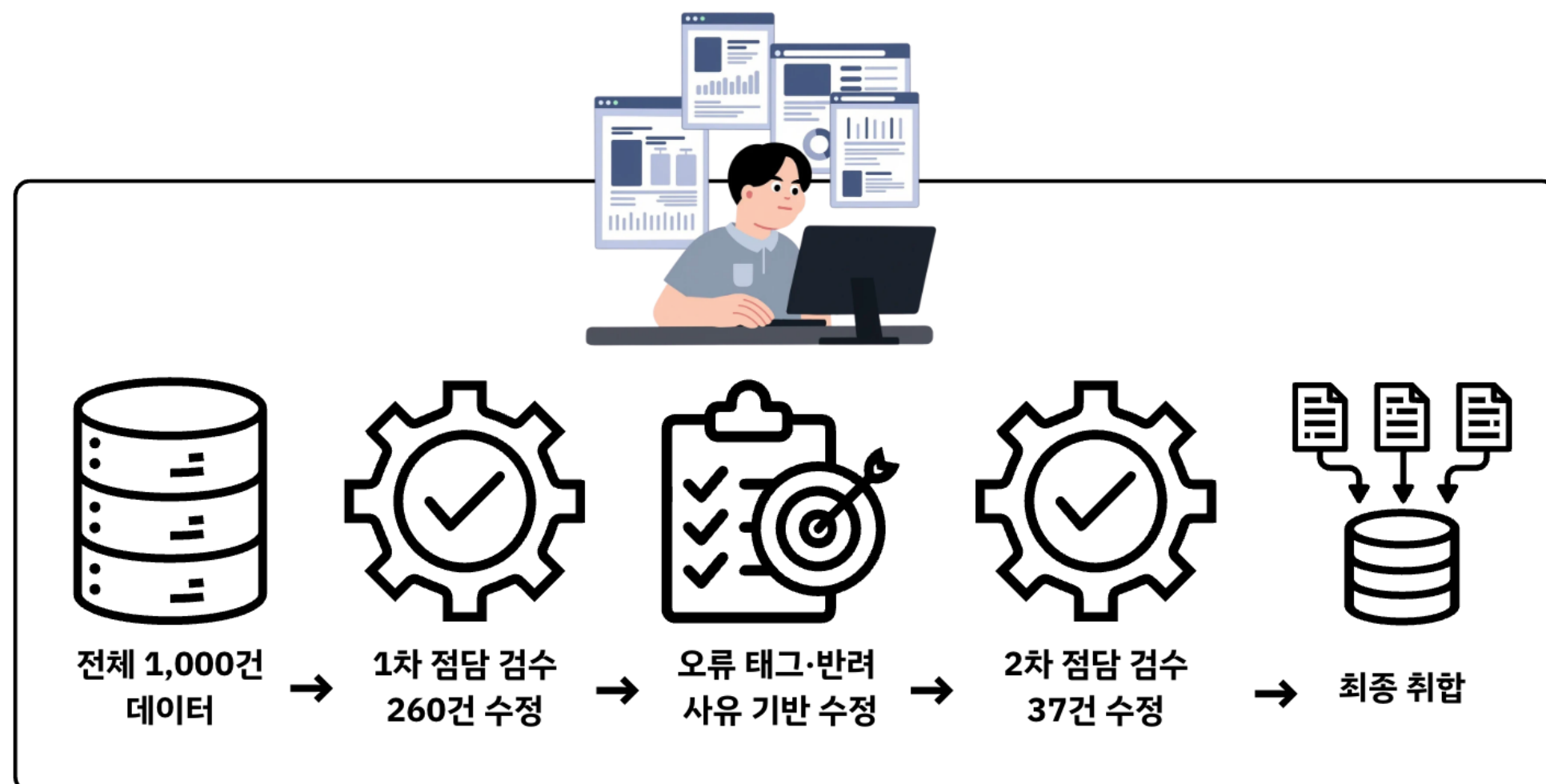
가족·친족	학교·학원	군대	직장·비즈니스	서비스업
200건	200건	200건	200건	200건
20%	20%	20%	20%	20%

난이도별 분포



일상적 존대부터 높임 일치·압존법, 복합 맥락까지 단계적으로 난이도를 구성

4. 품질 검수 결과



라벨별 1차 수정 문항

