개요: 기계독해 데이터셋이란?

자연어처리(NLP, Natural Language Processing) 분야에서 가장 도전적인 분야인 기계 독해(MRC: Machine Reading Comprehension) 기술 개발에 활용할 수 있는 학습 데이터셋으로 ㈜마인즈랩에서 구축했으며, 45만건의 한국어 질문과 대답으로 구성되어 있다.

기계 독해 기술은 질의응답 서비스에 주로 사용되며, 사람이 맥락을 이해하고 논리적으로 답을 찾는 것처럼 질의에 대한 답을 찾는 기술이며, AI 챗봇 상담, 방대한 전문지식에 대한 질의응답 및 시맨틱 검색 등에 활용하기 위해 활발한 연구 및 상용화를 진행하고 있다.

기계 독해의 사례 대해서는 아래를 참고할 수 있다.

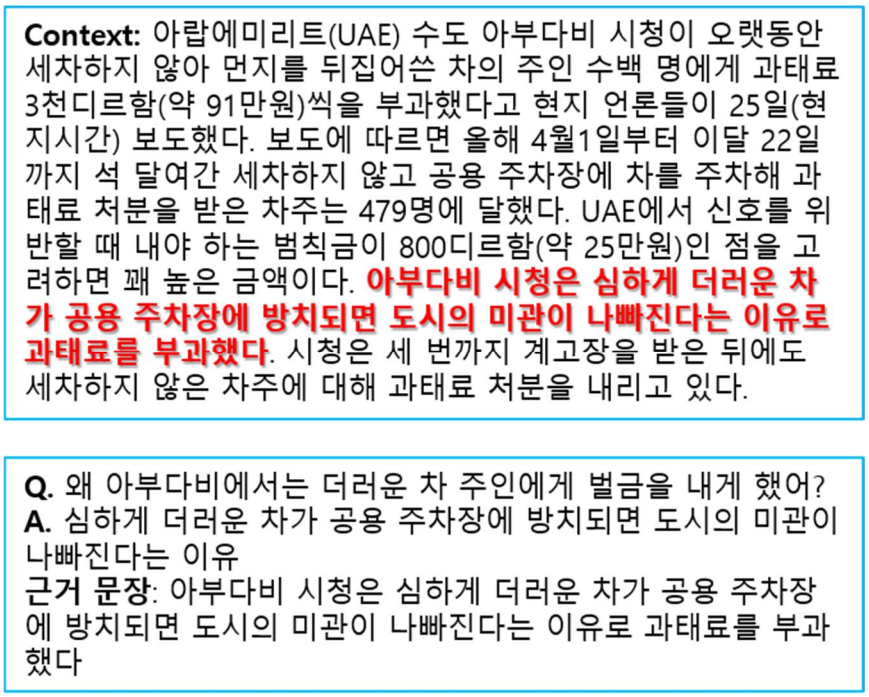

그림 1 기계 독해의 사례

데이터셋의 구성

본 데이터셋은 일반적으로 기계 독해 연구에 활용하는 질문-본문-정답 쌍의 표준 데이터셋 25만건과 질문-본문은 있으나 해당 본문에서는 답을 찾을 수 없는 정답 없는 데이터셋 10만 건, 그리고 질문에 대한 답 외에도 해당 답을 찾은 근거까지 레이블링한 설명가능한 데이터셋 10만건 등 총 45만건으로 구성되어 있다.

표준 데이터셋 25만건은 연구자가 일반적으로 연구를 진행하기에 충분한 양이며, 상용화 레벨에서는 강력한 사전학습모델을 만들 수 있는 양이다. 정답이 없는 데이터셋 10만건은 본문에 질문에 대한 답이 없을 경우 기계 독해 알고리듬이 정답이 없다고 대답할 수 있도록 학습하기 위해서 사용된다. 설명가능한 데이터셋은 이번 데이터셋을 구축하면서 추가된 형태로, 질문에 대한 답변 뿐 아니라 해당 답변을 도출하게 된 논리적 근거까지 제시할 수 있도록 하는 설명가능한 기계 독해 알고리듬을 개발할 수 있도록 했다.

데이터 종류	포함 내용	제공 방식
표준 데이터셋	질문과 답(25만 건)	JSON 포맷 파일
정답 없는 데이터셋	본문에서 답을 찾을 수 없는 질문(10만 건)	JSON 포맷 파일
설명 가능 데이터셋	질문과 답과 그 답을 선택한 단서(10만 건)	JSON 포맷 파일

데이터셋의 설계 기준과 분포

데이터셋을 설계할 때 가장 중요하게 고려했던 점은 데이터 밸런스이다. 본문과 질문, 정답을 구성할 때 적절한 분류기준을 만들었고, 해당 분류기준에 따라 골고루 데이터가 분포되도록 설계하여 학습 시 예상할 수 있는 데이터 편향성을 최소화하도록 했다.

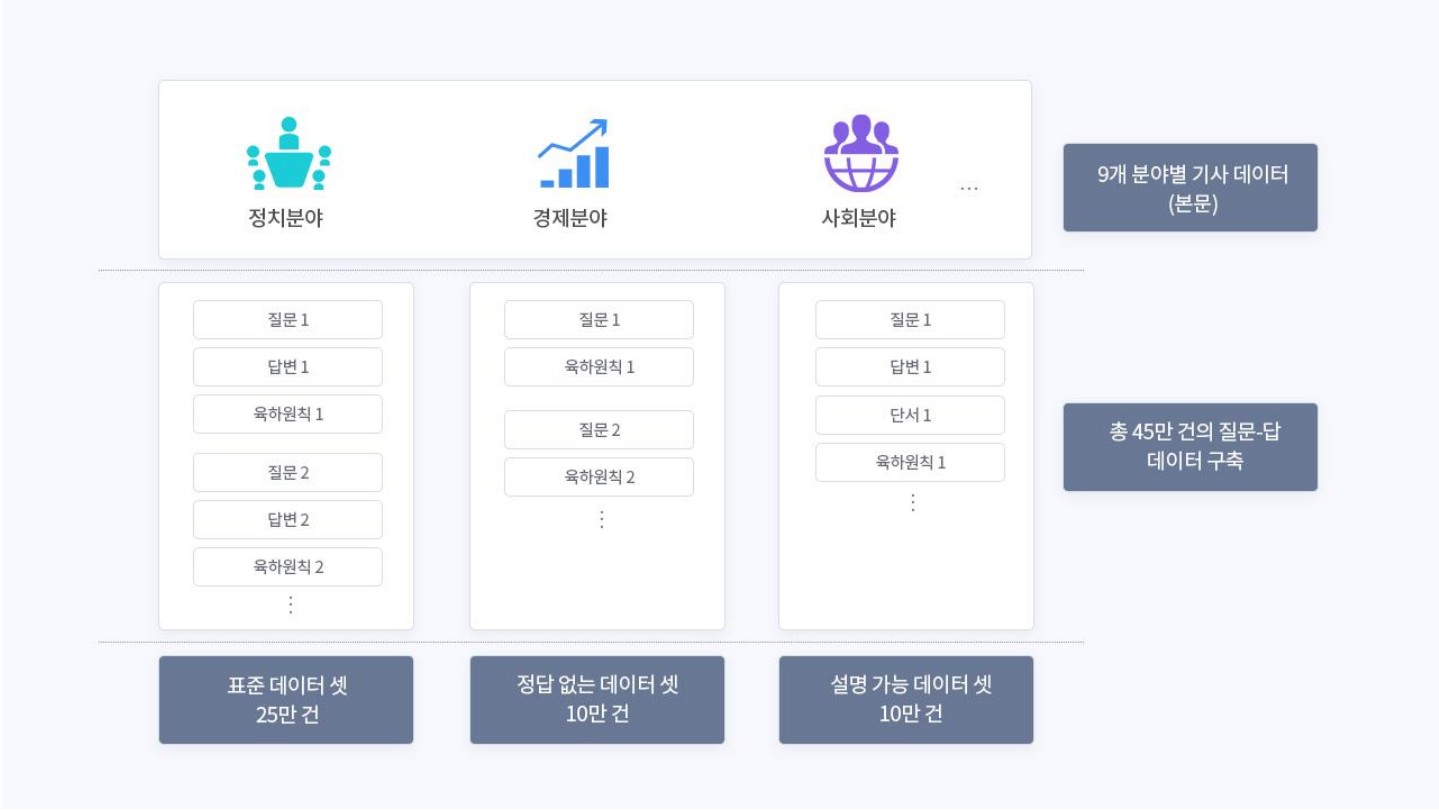

그림 2 데이터셋 구성 개요

본문, 질문, 답변의 구성 원칙과 주요 특징은 다음과 같다.

- 본문: 전체 100만 건 이상의 뉴스기사 본문에서 중복이나 유사 본문을 제거한 8만 건을 바탕으로, 크게 9가지 분야로 분류하여 각 분야 간 본문의 개수가 균등하도록 구축했다. 또한 짧은 본문(800자 이내), 중간 길이(800~2000자), 긴 본문(2000자 이상)이 고루 포함되도록 했다.

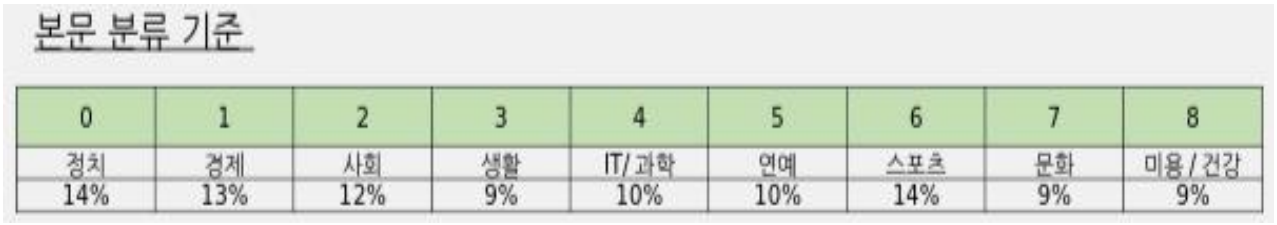

그림 3 본문 주제별 분류 분포

- 질문: 기존에 공개된 기계 독해 데이터셋들은 대부분 누가/언제/어디서 같은 단순 사실관계형 질의응답에 대한 데이터가 대다수의 비중을 차지하며, 어떻게/무엇을/왜 같이 상대적으로 어려운 질문은 적은 비중을 차지한다. 마인즈랩의 데이터셋은 누가/언제/어디서/어떻게/무엇을/왜 육하원칙에 따른 질문의 비중이 균등하도록 구축했다.
- 답변: 기계 독해 알고리듬이 답변할 수 있는 범위는 단어/구/절/문장/문단 등 다양하다. 기존에 공개된 기계 독해 데이터셋이 매우 단순하게 구성된 것과 달리, 이 데이터셋은 답변의 형식이 단어/구/절/문장/문단 등으로 다양하게 구성되어 있다.

데이터 구조

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

분류		표준 데이터셋	정답 없는 데이터셋	설명 가능 데이터셋
내용		본문, 질문, 답	본문, 질문	본문, 질문, 답, 근거
수량		25만	10만	10만
항목		포함여부		
제목	title	Y	Y	Y
분류	source	Y	Y	Y
본문 텍스트	context	Y	Y	Y
질문 번호	id	Y	Y	Y
육하원칙	classtype	Y	Y	Y

질문	question	Y	Y	Y
정답 시작 위치	answer_start	Y		Y
정답 텍스트	text	Y		Y
근거 시작 위치	clue_start			Y
근거 텍스트	clue_text			Y

데이터 예시

이 데이터는 설명 가능 데이터 기준이며, 표준 데이터셋, 정답 없는 데이터셋은 아래 예시에서 각각 clue, answers가 없는 구조를 가진다.


```
{"data": [{
        "source": 6,
        "paragraphs": [{
                "qas": [{
                        "question": "쎈 마이웨이 관련 기자간담회 누가 했어",
                        "id": "m4_278529-1",
                        "answers":[{
                                "answer_start": 0,
                                "text": "박영선"
                        }],
                        "clue":[{
                                "clue_start": 4,
                                "clue_text": "PD"
                        }],
                        "classtype": "work_who"
                }],
                "context": "박영선 PD는 18일 오후 서울 양천구 목동 SBS에서 모비딕의 토크 콘텐츠 쎈 마이웨이 관련 기자간담회를 열고 출연진에 신뢰를 드러냈다."
        }],
        "title": "1"
}, ...]}
```

※ 한 본문에 대해 qas(질문 - 답)가 여러 개일 수 있으며, 질문 번호(id) 생성규칙은 [제작자]_[질문번호]_[1]인 경우는 고유 질문이고, [2]의 경우는 유사 질문이다.

데이터 구축 과정

데이터 구축은 2018년 1월부터 12월까지 10개 뉴스 매체 뉴스 100만 건을 크롤링한 후 중복이나 유사 본문을 제거하는 필터링을 거쳐 8만 건을 기준으로 분야별 분류와 길이 별 균등 분포가 이루어지게 했다.

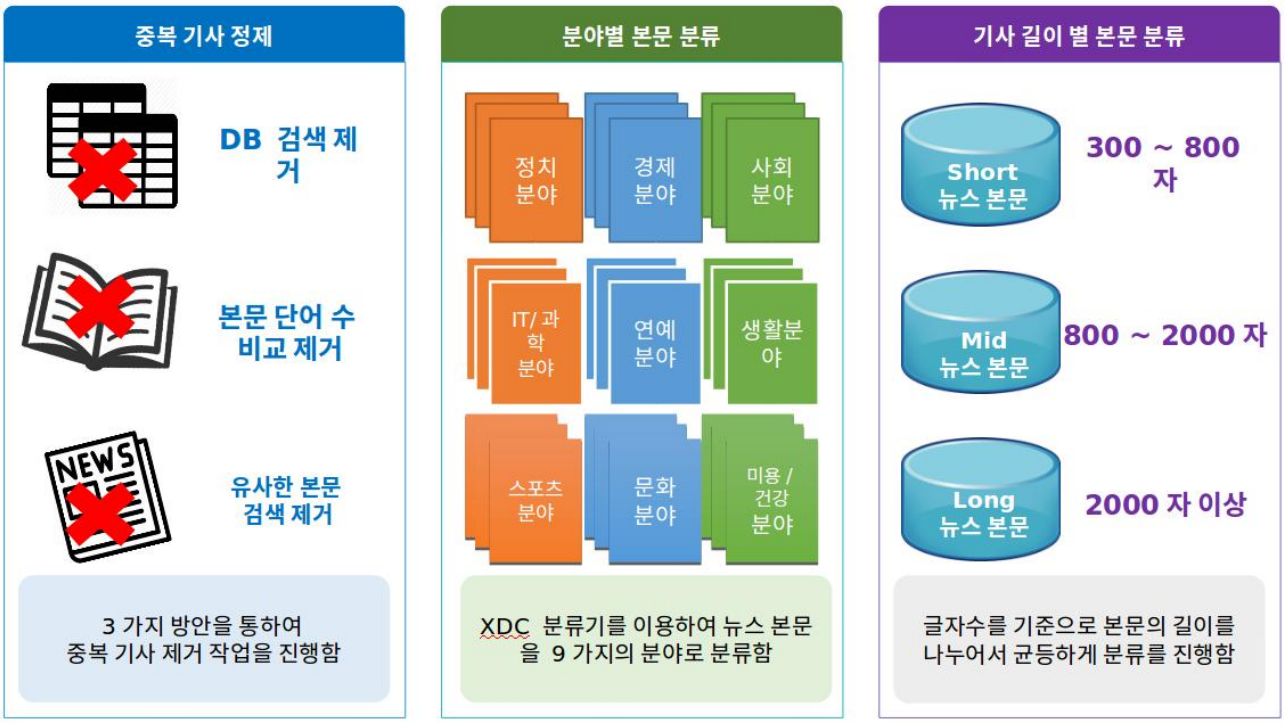

그림 4 데이터 정제와 분포 균등화

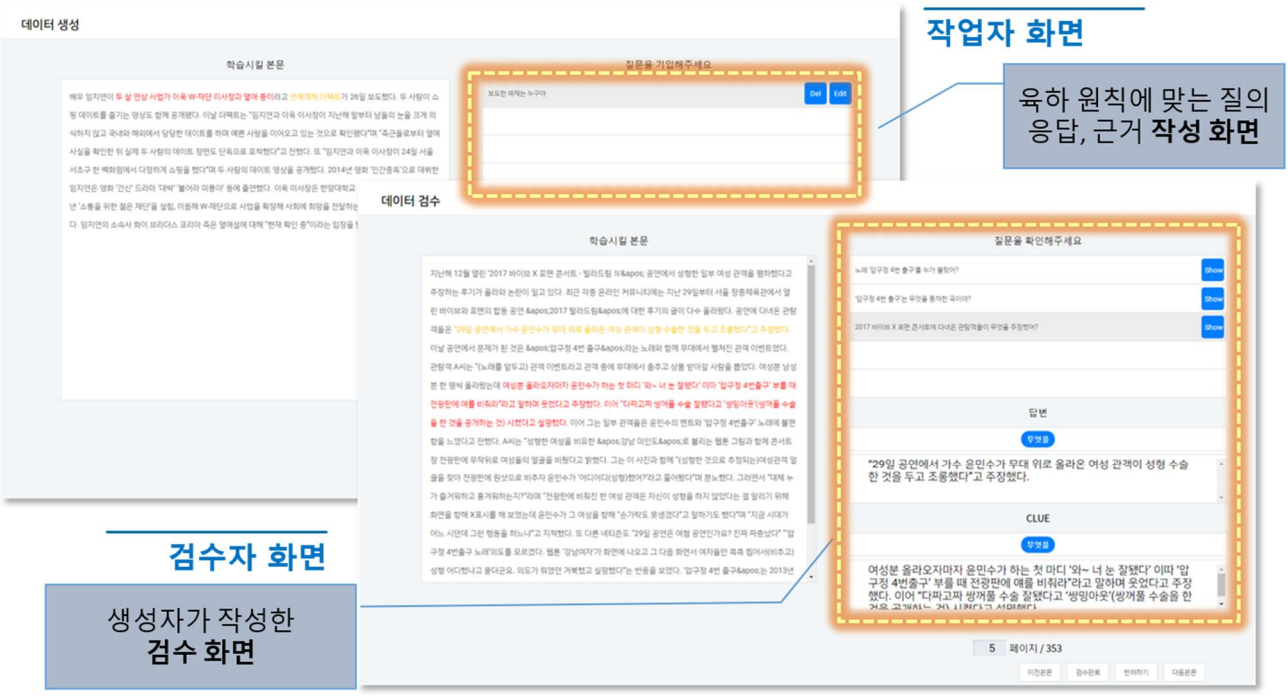

그림 5 효율적인 데이터 제작을 위한 웹 기반 툴

이렇게 일정한 기준에 따라 엄선된 8만 건의 기사에 대해서 기사 당 3개~7개의 질문과 답변 쌍을 생성하도록 작업하였다. 하나의 누가, 언제, 어디서 등 상대적으로 쉬운 질문은 한 본문 당 하나만 생성할 수 있도록 화면상에서 제한하였고, 기본적으로 본문 하나에서 육하원칙의 각 질문 유형이 하나씩 나오는 것을 원칙으로, 작업이 어려운 경우에만 작업불가로 넘길 수 있도록 하였기 때문에 결과적으로 한 본문 당 최소 3개에서 7개의 질의응답을 생성할 수 있었다. 어떻게, 왜, 무엇을 등 상대적으로 어려운 질문에 대해서는 보다 높은 작업단가를 책정했기에 육하원칙에 따른 질의응답 밸런스를 맞출 수 있었다.

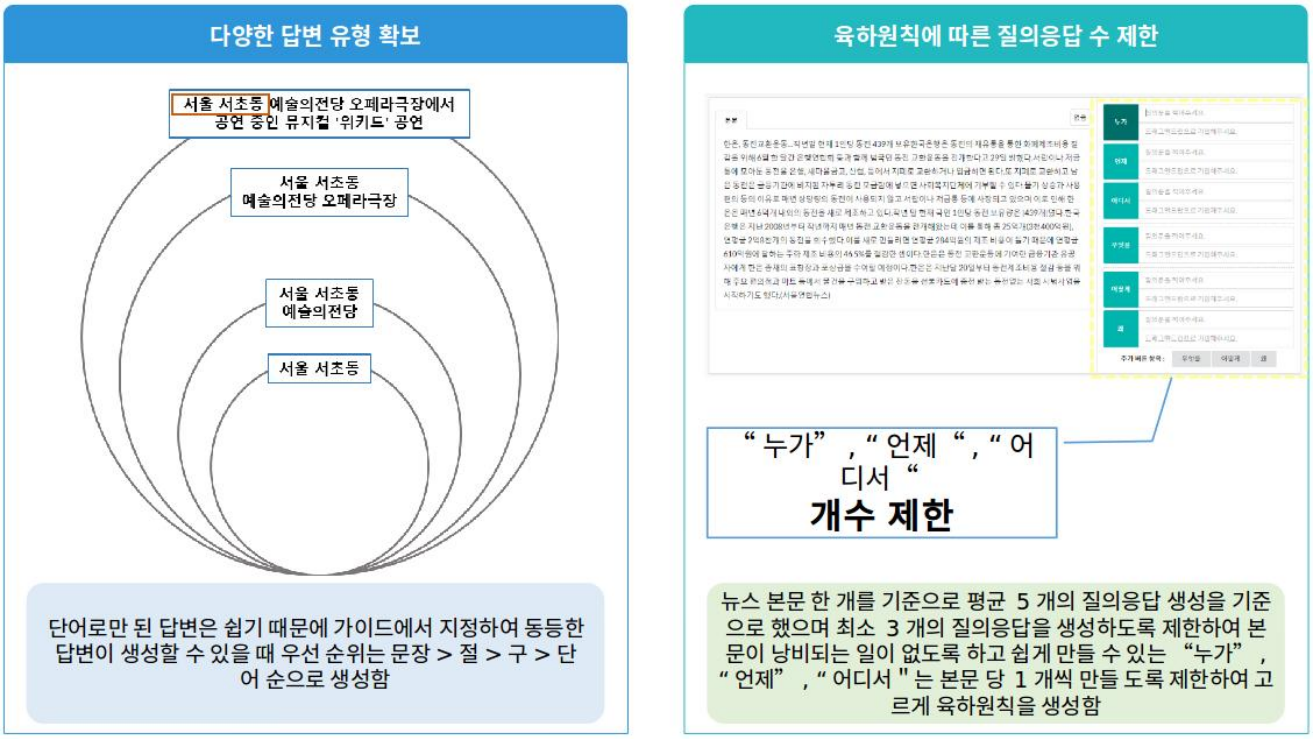

그림 6 답변 유형과 질문 생성

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 3단계 검수 체계를 구축했는데, 가장 하위 레벨에는 크라우드 워커들이 작업한 결과물을 가이드라인에서 제시한 형식에 맞는지(중복, 문장 규정 등) 체크하는 검수자가 있었고, 이들이 검수한 결과물에 대해서 내용적으로 유효한지 검수하는 재검수자가 팀을 이뤄 활동했다. 이렇게 만들어진 데이터셋을 전체적으로 들여다보며 데이터셋의 밸런스나 가이드라인의 적절성을 제시해주는 관리자는 마인즈랩의 직원으로 2년 이상 MRC 엔진을 직접 다뤄보며 학습 경험이 풍부한 인력을 배치하여 최종적인 데이터셋의 품질을 담보할 수 있었다.

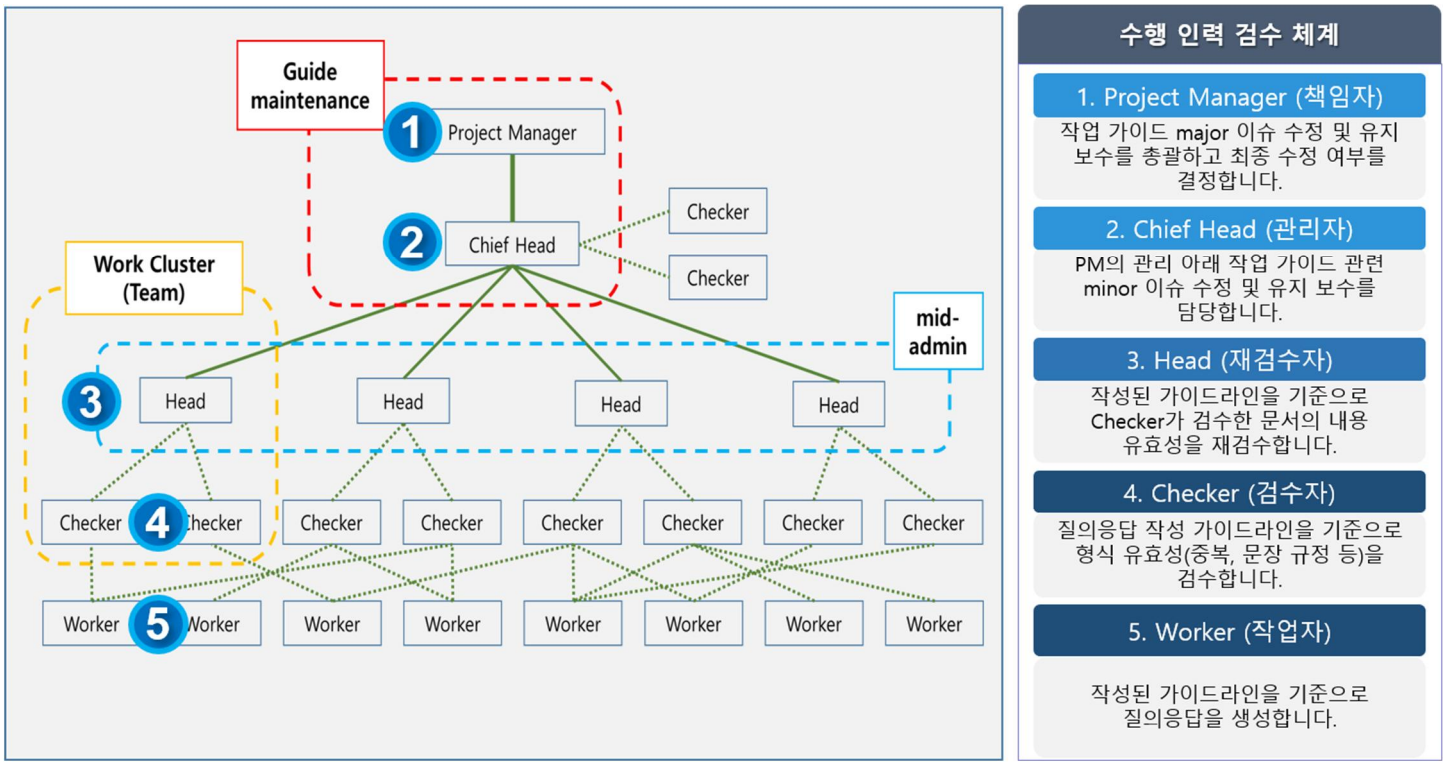

그림 7 품질 확보를 위한 3단계 품질 검수 체계

데이터 구축 담당자

수행기관(주관) : ㈜마인즈랩 (전화: 1661-3222), 이메일: hello@mindslab.ai