국립국어원 2022-01-15

발 간 등 록 번 호
11-1371028-000903-01

2022년 이야기 완성 평가 말뭉치 연구 분석

연구책임자
송 상 헌

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2022년 이야기 완성 평가 말뭉치 연구 분석’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2022년 5월 16일 ~ 2022년 12월 15일

2022년 12월 16일

연구책임자: 송상헌(고려대학교)

연구 기관: 고려대학교 산학협력단, (주) 플리토

연구책임자: 송상헌

공동연구원: 박진호 외 12명

보조연구원: 홍승혜 외 16명

보조원: 권은재 외 58명

<국문 요약>

2022년 이야기 완성 평가 말뭉치 연구 분석

본 사업은 디지털 언어자원 구축과 인공 지능 평가를 목적으로 '이야기 말뭉치'를 구축하였다. 이야기 말뭉치는 세 문장으로 구성된 담화를 1건으로 하여 15만 건의 이야기로 구성되어 있다. 개별 이야기는 시간의 흐름에 따라 일상생활과 관련된 사건을 기술한다. 관련 인공 지능 평가는 한국어 일반 상식에 부합하는 이야기 흐름을 기계가 포착하는지 정량적으로 평가하는 것을 목적에 둔다.

본 사업은 단계적으로 공정을 수행하였다. 첫째, 일상생활의 세부 주제를 포괄하는 이야기를 15만 건을 생성한다. 생성은 국어학에 대한 지식이 있는 크라우드 워커가 직접 작성한다. 작성 방식은 문형의 다양화와 인공 지능 평가 고도화를 위하여 (1) 자유 창작(creative writing), (2) 핵심어 기반 창작(keyword-based writing), (3) 그림 기반 창작(photo-based writing)을 택하였다. 일상생활의 세부 주제 선정은 한국어 학습자의 숙달도 평가를 위한 주제 항목을 일부 반영하였다.

둘째, 인공 지능 평가 목적으로 (1) 적절한 가설(plausible hypothesis) (2) 부적절한 가설(implausible hypothesis)을 생성한다. 가설은 일반 상식에 근거하여 유추할 수 있어야 하며, 일반 상식은 검색할 수 있는 백과사전적 지식보다 수용할 수 있는 사회적 통념이나 당연한 사실에 가깝다. 유추의 근거는 가설에 선행하는 문장이나 후행하는 문장이며, 각각 이야기의 첫 번째 문장과 세 번째 문장에 대응한다.

셋째, 담화 평정과 인공 지능 예비 평가를 수행한다. 담화 하나당 5인의 평정자가 두 가설 중 이야기의 흐름에 맞는 가설과 그렇지 않은 가설을 구별하여 선택한다. 평정자 간의 의견이 불합치하는 평정 미통과 사례는 그 비율이 매우 낮았다(0.33%). 평정이 표기된 가설과 이야기는 상식 추론 과제의 문제와 정답으로 활용된다. 인공 지능 예비 평가 결과, 한국어 인공 지능 모형 KR-BERT(49%)와 KLUE-RoBERTa-Base(52%)은 상식 추론을 매우 어려워했다.

본 사업은 이야기 말뭉치를 구축·변환하여 인공 지능 상식 추론 평가를 수행하였다. 또한, 한국어 인공 지능이 상식 추론에 취약할 수 있다는 점을 밝힘으로써, 구축된 이야기

말뭉치를 활용하여 인공 지능의 자연어 이해 수준을 한 층 더 향상시킬 수 있을 것으로 기대한다.

주요어: 인공 지능 언어능력평가, 이야기 완성, 자연어 이해, 상식 추론 평정

<Abstract>

2022 Research and Analysis of Story Cloze Task and Evaluation

This project investigates AI commonsense reasoning by building large-scale Korean Story Cloze datasets (henceforth, KSC). KSC consists of the three-sentence story corpus and (im)plausible sentence pairs. The entire KSC contains 150,176 instances (450,529 sentences) of annotated stories with a focus on the evaluation of AI commonsense reasoning. In addition to KSC, we created 150,275 pairs of (im)plausible hypothesis (300,550 sentences) with the total 751,375 human annotations (five annotators × 150,275 hypothesis sentence pairs). Specifically, the commonsense reasoning requires the ability to infer the casual and temporal relationships of sentences that describes an daily event such as cooking and baking.

To construct KSC datasets, we performed three human annotation tasks. First, we generated three-sentence stories using Korean WordNet (U-WIN) and Standard Word Categories for Korean learners (Kim et al., 2017). During the generation task, crowdworkers can refer to some clues borrowed from other sources including texts and images.

Second, we generated pairs of (im)plausible hypotheses inserted between the first and the last sentences of stories. The plausible hypothesis naturally completes the story, while the implausible hypothesis interrupts the story. This task is typically referred to as Story Cloze task because it requires the ability to infer the causal and temporal relations.

Lastly, we performed the evaluation task to annotate the stories. For each story, five crowdworkers are recruited to assess the plausibility of the given hypothesis.

Note that some stories are re-evaluated when crowdworkers fail to reach the agreement. These human assessments are annotated to Story Cloze tasks so as to measure the prediction accuracy of Korean neural language models.

In this project, we release a natural language understanding benchmark, KSC, for the systemic evaluation of AI. We found that KR-BERT and KLUE-RoBERTa struggle to infer the correct answer (Accuracy: 49% - 52%). We expect that Korean language models trained on KSC datasets can show more accurate commonsense reasoning.

Keywords: Artificial intelligence, Natural language understanding, Story cloze task, Commonsense reasoning

Project Director: Sanghoun Song (Korea University)

차 례

제1장 사업 개요

제2장 이야기 생성

제3장 가설 생성

차 례

제4장 가설 평정

제5장 검증 평가

제6장 맺음말

표 차례

그림 차례

제 1 장

사업 개요

1. 사업 목적

본 사업은 4차 산업혁명 대비 인공 지능 기술 개발, 활용을 위한 말뭉치를 구축하여 국어 자원의 활용도와 가치 제고를 목적으로 한다. 이를 위하여 국어 인공 지능 기술의 상식 추론 능력과 언어 생성 능력을 평가하는 말뭉치를 구축한다.

2. 사업 수행

1) 사업 수행 범위

본 사업은 인공 지능의 언어능력평가를 목적으로 15만 건 이상(총 45만 문장 이상)의 한국어 이야기 말뭉치를 구축하였고, 이를 변환하여 15만 건 이상(총 30만 문장 이상)의 상식 추론 평가 과제를 구축하였다. 이와 함께, 구축한 추론 평가 과제 15만 건은 각 5인의 평정(총 75만 건 평정 이상)을 통과하였다. 또한 2종의 말뭉치 구축 방법과 가이드라인을 개발하여 공정의 효율성과 일관성을 확보하기 위한 지침으로 삼았다.

2장에서는 인공 지능의 언어능력평가를 위하여 규모 15만 건 이상(총 45만 문장 이상)의 이야기 말뭉치를 구축하였다. 또한, 말뭉치 구축 지침을 수립하여 이야기 완성 평가 말뭉치의 방법론적 일관성과 이야기의 주제적 다양성을 확보하였다. 이와 함께, 양질의 이야기 말뭉치 구축을 위하여 본 사업 수행에서 채택한 방법과 공정률을 보고한다.

3장에서는 상식 추론 과제 말뭉치를 구축·변환하여 15만 건 이상(선택지 문장 30만 개 이상)의 평가 말뭉치를 구축한 내용을 다룬다. 이야기 말뭉치를 상식 추론 과제로 구축·변환하는 방법론 및 지침을 수립하여 인공 지능 평가의 고도화를 확보하였다. 향후, 양질의 상식 추론 과제 말뭉치 구축을 위하여 본 사업 수행에서 선별한 사례에 대하여 오류 분석을 시도하였다.

4장에서는 인공 지능의 언어능력평가라는 목적에 부합하는 품질 제고 계획을 수립·실행하기 위하여 15만 건 이상(평정 횟수 75만 개 이상)의 담화 평정을 수행하여 상식 추론 과제 정답 세트로 구축·변환하였다. 담화 평정은 상식 추론 과제에 정답을 부착하기 위함이며, 이를 활용하여 인공 지능의 예측 정확도를 평가할 수 있다. 인공 지능 평가는 구축된 문제·정답 세트를 활용하여 예비적으로 실행하였다.

본 사업은 상술한 수행 절차에 따라 공정을 완료하여 이야기 말뭉치와 상식 추론 과제 말뭉치 2종을 구축하였다.

2) 사업 수행 일정

본 사업은 2022년 5월 16일 착수하여 2022년 12월 15일까지 7개월간 수행되었다. 사업 수행은 매달 온라인 회의와 함께 월간 보고서를 작성하였으며, 구체적인 진행은 다음의 <표 1>에 제시하였다. 변동사항에 대한 보고와 공정 수행 결과에 대한 정보 교환은 발주기관과 수시로 진행하였다.

<표 1. 사업 공정 수행 경과>

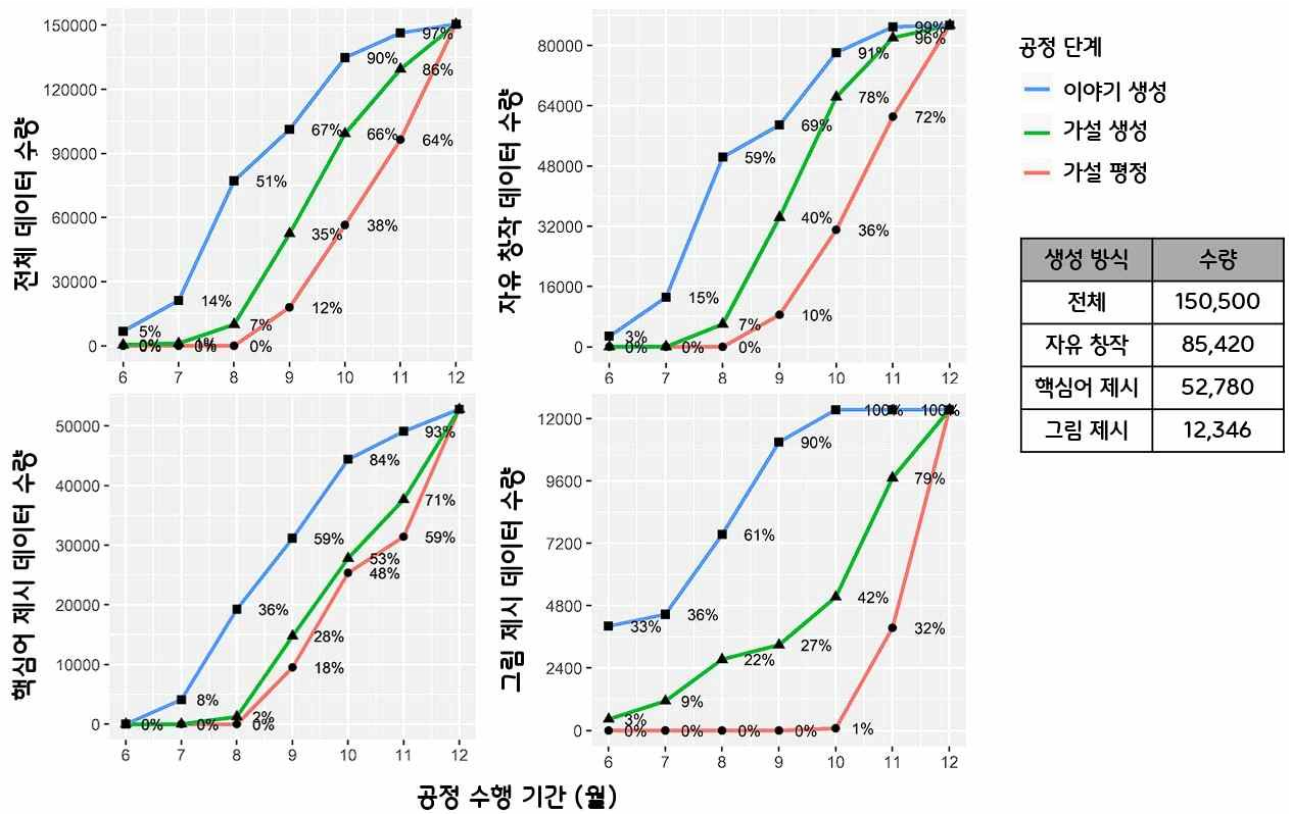
과업 구분	5월	6월	7월	8월	9월	10월	11월
사업 관리 계획 수립	■						
이야기 말뭉치 생성		■	■	■	■	■	■
가설 추론 과제 생성				■	■	■	■
이야기 완성 평정					■	■	■
인공 지능 예비 평가						■	■

3. 공정 수행의 난점과 해결 방안

1) 공정 수행의 난점

본 사업은 서로 다른 성격의 2종의 말뭉치 구축과 평정을 단계적으로 수행하였다. 특히, 이야기 말뭉치 구축이 완료되어야 상식 추론 과제로 변환이 가능하기 때문에 이야기 말뭉치 구축 공정은 주어진 사업 기간 내에 신속하게 공정을 완료하여야 했다. <그림 1>은 전체적인 공정 대비 생성 공정률 정보를 월별로 요약한 것이다.

<그림 1. 공정 수행률 요약>



생성 방식	수량
전체	150,500
자유 창작	85,420
핵심어 제시	52,780
그림 제시	12,346

이야기 생성 공정은 사업 기간 전체에 걸쳐 수행되었으며, 서로 다른 세 가지 생성 방식인 (1) 자유 창작 (2) 핵심어 제시 (3) 그림 제시를 활용하여 병렬적으로 수행되었다. 생성된 이야기 말뭉치를 상식 추론 과제로 변환하여 평정을 거쳐야 하는 사업의 특성에 비추어 보면, 이야기 생성과 검증을 모두 종료하고 상식 추론 과제로 변환하는 것이 타당해 보인다.

그러나 이와 같은 공정은 수개월 내에 이야기 말뭉치 구축을 종료해야 하므로, 사업 기간 내에 상식 추론 과제 변환과 평정까지 완료하기에는 매우 촉박하다. 이 때문에, 수십 명의 작업자와 평정자를 한꺼번에 운용해야 하는 사업 운영상의 부담에도 불구하고, 공정 수행 기간 내에 이야기 생성과 상식 추론 과제 변환, 그리고 평정을 병진적으로 진행하였다. 즉, 이야기 생성 15만 건이 모두 구축 종료되는 것을 기다리지 않고, 일정량의 이야기가 생성되면 즉시 검수하여 상식 추론 과제로 변환하고 평정하였다.

2) 세부 공정 과정

상술한 바와 같이, 대규모 작업자 운용이라는 운영상의 부담에도 불구하고 서로 다른 공정을 병진적으로 진행한 이유는 납기일이 도과하기 이전에 모든 공정을 성공적으로 수

행하기 위함이었다. 이와 관련하여, 공정 초기 과정 및 후기 과정의 공정률에 대한 정보를 아래 제시하였다(<표 2-7>).

<표 2. 6월 공정 수행률>

22.06.21

방식	목표량	**(A)생성**	(B)가설	(C)평정	계
자유 창작	87,400	**2,770 (3.17%)**	0 (0.00%)	0 (0.00%)	2,770 (1.01%)
핵심어	54,200	**0 (0.00%)**	0 (0.00%)	0 (0.00%)	0 (0.00%)
그림	12,000	**4,016 (34.47%)**	429 (0.4%)	0 (0.00%)	4,445 (12.35%)
계	153,600	**6,786 (4.32%)**	429 (0.28%)	0 (0.00%)	7,215 (4.47%)

<표 3. 7월 공정 수행률>

22.07.18

방식	목표량	**(A)생성**	(B)가설	(C)평정	계
자유 창작	87,400	**13,146 (15.04%)**	0 (0.00%)	0 (0.00%)	13,146 (5.01%)
핵심어	54,200	**4,097 (0.00%)**	0 (0.00%)	0 (0.00%)	4,097 (2.52%)
그림	12,000	**4,470 (34.47%)**	1,123 (0.4%)	0 (0.00%)	5,593 (15.54%)
계	153,600	**21,173 (14.14%)**	1,123 (0.28%)	0 (0.00%)	22,836 (14.86%)

<표 4. 8월 공정 수행률>

22.08.15

방식	목표량	**(A)생성**	(B)가설	(C)평정	계
자유 창작	87,400	**50,384 (57.65%)**	6,030 (6.90%)	0 (0.00%)	56,414 (21.52%)
핵심어	54,200	**19,261 (35.54%)**	1,237 (2.28%)	0 (0.00%)	20,498 (12.61%)
그림	12,000	**7,546 (62.88%)**	2,733 (22.4%)	0 (0.00%)	10,279 (28.55%)
계	153,600	**77,191 (50.25%)**	10,000 (06.51%)	0 (0.00%)	22,836 (18.92%)

공정 계획 수립 및 검토(5월) 이후, 공정 초기 단계(6-8월)에서 작업 성과는 최초 계획 대비 부진하였다. 이는 대규모 말뭉치 구축을 위하여 다수의 작업자를 고용하여야 한다는 점에서 사업 운영상의 어려움이 있었기 때문이다. 이를 해결하기 위하여 일부 이야기 말뭉치가 구축되면 상식 추론 과제로 변환하여 서로 다른 단계의 공정을 병진적으로 수행하였다. 7월에 수행된 이야기 말뭉치 구축에서 약 2만 건 내외의 이야기 생성이 완료되었고, 약 1천 건 내외의 상식 추론 과제로 먼저 변환하였다(<표 3>). 다만, 평정은 수행하지 않았다. 마찬가지로 8월에 전체 이야기 말뭉치 구축 요구량의 50% 이상을 수행하는 과정에서 약 1만 건 내외의 상식 추론 과제를 함께 병진적으로 구축하였다(<표 4>). 이러한 병렬적인 공정 수행을 통하여 사업 기간 내에 모든 공정을 완료할 수 있었다.

공정 후기 과정의 관건은 공정률의 향상과 평정을 성공적으로 수행하는 것이 중요했다.

<표 5. 9월 공정 수행률>

22.09.16

방식	목표량	(A)생성	(B)가설	(C)평정	계
자유 창작	85,420	**58,947 (69.01%)**	34,386 (40.26%)	8,506 (9.96%)	101,839 (39.74%)
핵심어	52,780	**31,155 (59.03%)**	14,779 (28.00%)	9,506 (18.01%)	45,934 (35.01%)
그림	12,300	**11,104 (90.28%)**	3,293 (26.77%)	0 (0%)	14,397 (39.02%)
계	150,500	**101,206 (67.25%)**	52,458 (34.86%)	18,012 (12.03%)	162,170 (35.92%)

<표 6. 10월 공정 수행률>

22.10.16

방식	목표량	(A)생성	(B)가설	(C)평정	계
자유 창작	85,420	**78,099 (91.43%)**	66,340 (77.66%)	31,040 36.34%)	175,479 (68.48%)
핵심어	52,780	**44,375 (84.08%)**	27,788 (52.65%)	25,336 (48.00%)	97,499 (61.58%)
그림	12,300	**12,346 (100.00%)**	5,134 (41.74%)	80 (0.65%)	17,560 (47.59%)
계	150,500	**134,820 (89.58%)**	99,262 (65.95%)	56,456 (12.03%)	290,538 (64.35%)

<표 7. 11월 공정 수행률>

22.11.16

방식	목표량	**(A)생성**	(B)가설	(C)평정	계
자유 창작	85,420	**84,948 (99.45%)**	82,065 (96.07%)	61,093 (36.34%)	228,106 (89.01%)
핵심어	52,780	**49,081 (92.99%)**	37,659 (71.35%)	31,401 (48.00%)	118,141 (74.61%)
그림	12,300	**12,346 (100%)**	9,724 (79.06%)	3,947 (0.65%)	26,017 (70.51%)
계	150,500	**146,375 (97.26%)**	129,448 (86.01%)	96,441 (64.08%)	372,264 (82.45%)

공정 후기 단계(9월-11월)에서는 후속 공정인 (B) 가설 (C) 평정 공정이 주로 수행되었으며, 매월 (B) 가설 공정은 20-30% 내외, (C) 평정 공정은 30%-50% 내외가 수행되었다.

공정률에 비추어 본 사업의 난점은 다음과 같다. 첫째, (A) 이야기 생성은 (B) 상식 추론 가설 생성과 (C) 담화 평정 공정에 선행하여 수행되어야 한다. 예를 들어, (A)의 공정률이 (B)의 공정률보다 낮으면 (B)의 공정이 더이상 진행되지 않아 공정 수행에 지연이 발생한다. 둘째, (A)의 공정 부담이 매우 크다. 45만 건의 문장이 생성되어야 하므로 수십 명의 작업자 관리가 요구되며, 공정 수행 일정을 (B)와 (C)와 함께 고려해야 하므로 일정 수립이 쉽지 않았다. 향후 유사 말뭉치 구축 사업에서는 위와 같은 난점을 미리 고려해야 할 것이다.

제 2 장

이야기 생성

1. 이야기 생성 대상과 범위

이야기의 사전적 정의는 '어떤 사물이나 사실, 현상에 대하여 일정한 줄거리를 가지고 하는 말이나 글'을 일컫는다(표준국어대사전, 국립국어원). 그렇기 때문에, 이야기는 현실 세계의 사건을 어떻게 인간이 언어적으로 개념화하는지와 연관이 깊다. 이를 바탕으로, Mostafazadeh et al. (2016)은 인공 지능 또는 언어 모형이 인간처럼 추론할 수 있는지를 간접적으로 탐구하기 위하여 ROC stories 말뭉치를 구축하여 언어 모형의 추론 능력을 평가하였다(<표 8>).

<표 8. ROC stories 예문>

Title	Five-sentence Story
The Test	Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Her teacher stated that the test is postponed for next week. Jennifer felt bittersweet about it.

ROC stories 말뭉치는 일상생활에서 흔히 볼 수 있는 사건들(예: 학교 가기, 옷 갈아입기, 빨래 하기)을 중심으로 5문장으로 구성된 짧은 이야기를 수집하였다. 인공 지능은 마지막 문장을 제외한 이야기를 읽고, 이야기의 흐름에 비추어 적절한 문장을 고르는 문제를 풀게 된다. Mostafazadeh et al. (2016)은 이를 이야기 클로즈 과제(Story Cloze Task)라고 불렀다.

ROC stories의 구축 내용과 목적에 착안하여, 본 과업 수행은 3문장으로 구성된 이야기를 다룬다(<표 9>).

<표 9. 국립국어원 이야기 말뭉치 예시>

제목	세 문장 이야기
초대	민서는 친구네 집에 초대를 받아 방문했다. 친구는 민서를 위해서 직접 음식을 해 주었다. 민서는 음식을 먹고 친구에게 엄지손가락을 들어 보였다.

본 과업에서 이야기는 (1) 일상생활과 관련된 주제 (2) 일반 상식에 근거한 시간적 흐름을 표상해야 한다. 또한 '인공 지능의 언어능력평가'라는 실용적 목적에 비추어 이야기의 문체와 어휘 수, 문장 길이 등이 통제된다. 구체적으로는 평서문과 과거시제이어야 하며, 어휘 수는 12단어 이내, 문장 길이는 3문장으로 한정된다. 이에 대하여, 이야기 생성 방법과 지침에서 상술한다.

2. 이야기 생성 방법과 지침

1) 이야기 생성 방법

이 절에서는 이야기 생성을 위한 작업 환경(작업자 선별 및 시스템 조성)과 이야기 생성을 위한 생성 방식 선정 및 과정에 대하여 설명한다. 작업자(또는 크라우드워커)는 아래와 같이 선발된다. 온라인 설문 조사지를 배부하여 자격 요건에 부합하는 지원자에게 간단한 퀴즈 풀이를 요청한다. 자격 요건은 국어학/언어학 및 유관 분야 전공 학부 3학년 이상이다. 이는 고품질 이야기 말뭉치 구축을 위한 언어학적 생성 지침을 정확하게 이해하고 수행할 수 있는지 선별하기 위함이다. 또한, 퀴즈는 한국어 문법에 대한 기초적인 10개 문제로 구성되어 있다.

선발된 지원자는 샘플 작업에 투입된다. 언어학적 지침이 포함된 안내문을 배부하고 10개 이야기를 제출하도록 하여, 생성된 이야기의 품질을 검수한다. 또한, 자기소개서(500자 내외)를 제출하도록 하여 말뭉치 구축 과제에 참여 사유와 개인적인 목표를 확인한다. 최종적인 지원자 과제 투입 여부는 참여연구원 3인의 동의를 요한다.

위와 같은 엄격한 선발 절차에도 불구하고, 과제의 규모와 난도로 인하여 참여 연구원의 적극적인 사후적 통제가 필요하였다. 사후적 통제는 (1) 주간 작업량 부진에 대한 서면

통지(이메일 포함) (2) 온오프라인 성과 관리 회의 참석 요청 (3) 박사급 참여연구원과 일대일 문제 점검 회의를 포함하였다. 그러나, 일부 (연구)보조원은 사후적 통제에 응하지 않았으며, 성과 부진이 심각한 보조원에게는 별도의 사후적 조치를 취하였다.

다음으로, 이야기 생성을 위한 생성 방식(자유 창작, 그림 기반, 키워드 기반)을 선정한 이유와 장점, 각 방식에서 단서 선정 과정 및 각 3가지 방식의 공통된 생성 방법의 특징들을 설명한다. 이야기 생성 공정을 수행하기 위하여 작업자에게 서로 다른 세 가지 방식을 부과하였다. 이 방식은 문체와 주제의 다양성과 인공 지능 평가의 고도화를 목적으로 도입되었다.

자유 창작 방식(Creative)은 주제에 해당하는 단어를 제시하고 이에 따른 문장 3개를 생성하는 방식이다. 자유 창작 방식은 문장을 완전히 새롭게 생성할 수 있으며 각 이야기 구조별 소주제를 분명하게 하기에 용이하다는 장점이 있다.

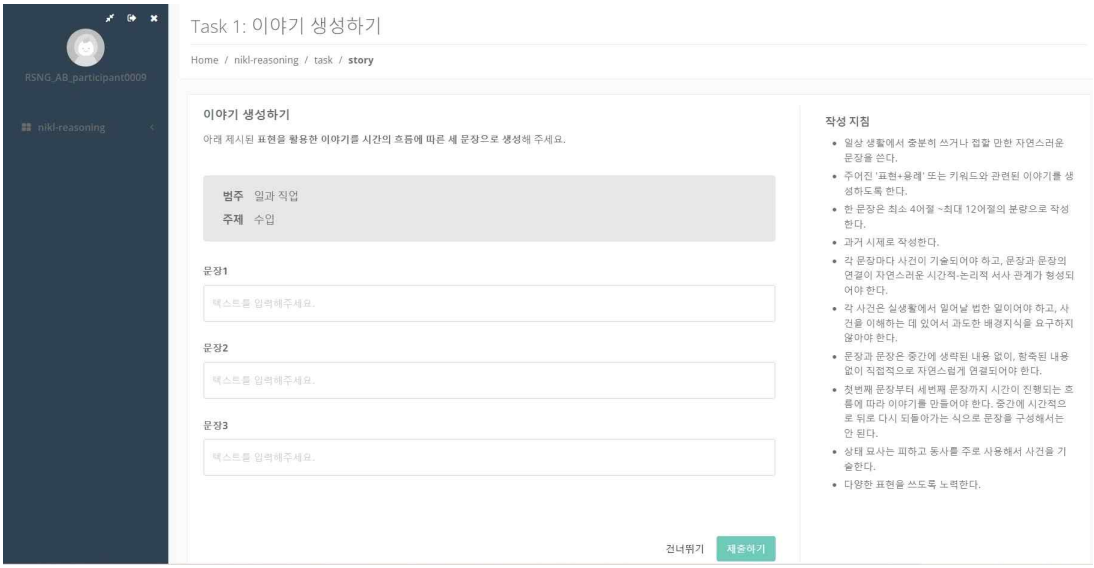
그림 기반 방식(Photo)은 주어진 사진 속 장면을 기술하는 문장을 활용하여 이야기를 구성하는 방식이다. 주어진 사진에 대한 문장에는 이미 대상, 배경, 사건 등이 명시되어 있어서 소주제의 선정이 상대적으로 쉽다. 또한, 사진을 참조하여 작업자가 이야기 구성을 자연스럽게 만들어 갈 수 있다는 장점이 있다.

핵심어 기반 방식(Keyword)은 제시된 대주제-소주제 핵심어를 활용하여 이야기를 구성하는 방식이다. 핵심어 활용 시에 임의의 핵심어를 자의적으로 활용하는 것이 아니라, 국립국어원의 연구 결과물인 김중섭 외(2017), <국제 통용 한국어 표준 교육과정 적용 연구>를 활용하였다. ‘사회’, ‘예술’, ‘전문 분야’ 세 범주를 제외하고 14 범주 72개 항목의 주제 목록을 ‘일상생활과 관련된 주제’로서 설정할 수 있으며, 일상생활에서의 원활한 한국어 사용을 목적으로 하는 다양한 주제를 활용한다는 장점이 있다.

핵심어 기반 방식에 해당하는 이야기 생성 작업 화면 예시를 아래 제시하였다(<그림 2>). 아래 크라우드소싱 플랫폼에서 작업자는 프롬프트에 제시된 범주와 주제에 따라 이야기를 생성한다. 이와 달리, 자유 창작 방식은 별도의 정보가 제공되지 않는다.

<그림 2. 이야기 생성 작업자 화면>

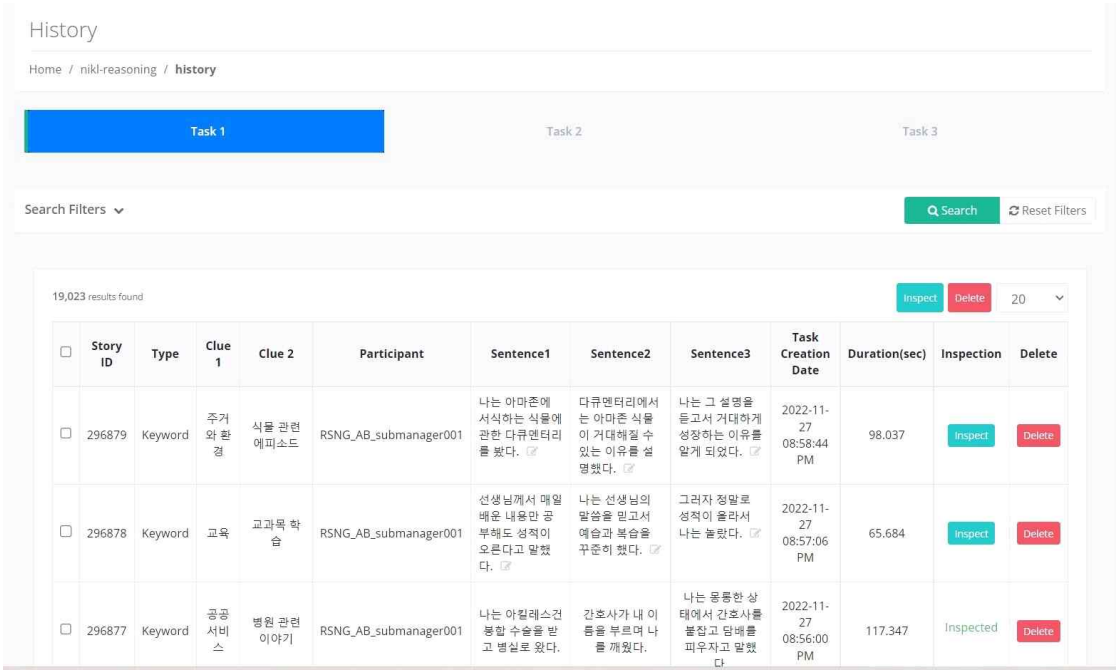
NO.	작업자 화면

1	 범주 - 총 15개 항목으로 구성됨 주제 - 총 139개 항목으로 구성됨 문장1 - 첫 번째 문장 문장2 - 두 번째 문장 문장3 - 세 번째 문장 건너뛰기 - 범주/주제가 부적절하여 문장 생성에 적합하지 않다고 판단될 때 넘어가는 용도이다. 제출하기 - 문장1, 문장2, 문장3의 생성이 완료된 경우에 결과물의 저장을 위한 용도이다. 작성지침 - 문장 생성 시에 개별 작업자가 참고 및 준수해야 하는 지침이다. 문장 생성 시에 주의해야 할 사항을 정리하여 작업자의 작업 안정성을 극대화하고 부적절한 문장의 생성이 최소화될 수 있도록 하였다.

범주/주제는 <국제 통용 한국어 표준 교육과정 적용 연구>에서 임의로 추출되었다. 건너뛰기는 임의로 주어지는 범주/주제를 조합하여 문장 생성이 적절하지 않다고 작업자가 판단할 때, 다른 범주/주제를 작업할 수 있도록 한다. 작성 지침은 작업 프롬프트 화면에 함께 출력하여 필요할 때마다 손쉽게 참조할 수 있도록 하였다.

이야기 생성에 대한 관리자 화면을 아래 제시하였다(<그림 3>). 크라우드소싱 플랫폼에서 관리자는 작업 결과를 검수하여 교정할 수 있다

<그림 3. 이야기 생성 관리자 화면>

NO.	관리자 화면
2	
	Story ID - 이야기 고유 식별 번호 Type - 다음의 세 가지 작업 방식 중 하나에 해당한다. (1) 자유 창작(creative) / (2) 핵심어(keyword) / (3) 그림(photo) Clue 1~2 - 첫 번째 ~ 두 번째단서 Participant - 작업자 고유 번호(개별 작업자의 식별이 가능하도록 하여 작업자별 문장 생성 진척 현황을 파악하기 위한 용도) Sentence1~3 - 첫 번째 ~ 세 번째 문장 Task Creation Date - 문장을 생성한 날짜를 지칭한다. Duration(sec) - 문장을 생성하는데 걸린 시간(초 단위)을 나타낸다. Inspection - 생성된 문장에 대하여 부적절 여부를 검수하여 통과 판정을 받은 경우에 해당한다. Delete - 부적절하거나 모호한 내용으로 인해 검수에 통과하지 못한 경우에는 삭제된다.

이야기 생성 관리자 화면은 작업의 내용을 검수할 수 있도록 할 뿐만 아니라, 각 작업자가 작업을 수행한 시간을 확인할 수 있도록 하였다. 이는 작업의 내용을 검수하여 최종 납품하는 절차를 준수하는 것과 함께, 작업자의 근태를 관리할 수 있는 기초 정보를 제공하기 위함이다. 예를 들어, 제출된 이야기의 작업 간격이 매우 짧다면 이야기의 품질을 검수하여 작업의 내용이 가이드라인에 알맞은지 확인하였으며, 반대로 작업 간격이 매우 길다면 작업자에게 별도로 연락하여 작업을 계속하고 있는지 확인하였다.

앞서 언급한 바와 같이, 이야기 말뭉치 구축 목표는 45만 개 이상의 문장을 제작하는 것이다. 구축 규모가 매우 크고 다수의 작업자가 투입되므로 세세한 지침이 아닌 대원칙을 제시하였다. 아래는 지침의 일부이다. 더욱 자세한 지침은 [붙임 1.]에 기술하였다.

(1) 먼저, 문장은 번역투를 피하고 자연스러운 한국어 문장으로 생성하는 것을 원칙으로 한다. 또한, 과거 시제를 사용하는 것을 원칙으로 한다.

- 오늘은 그가 고대하던 콘서트에 가는 **날이었다(*날이다)**. 그런데 갑자기 일이 많아져 휴가를 가지 못하게 **되었다(*된다)**. 안타까운 마음에 그에게 맛있는 점심을 사주며 위로했다.

다음으로, 자유 창작 방식을 통한 이야기 생성에서는 텍스트•이미지 단서(Clue)를 제공하였다. 자유 창작 방식에서 텍스트 단서는 한국어 어휘의미망인 울산대학교 U-WIN과 개방형 한국어 사전인 ‘우리말샘’의 용례를 활용하였다(<표 10>). 키워드 방식에서는 김중섭 외(2017), ‘국제 통용 한국어 표준 교육과정 적용 연구’를 활용하였다.

<표 10. 자료 유형과 텍스트 예시>

작업 방식	자료 유형	텍스트 단서 예시
자유 창작	U-WIN	(1) 주다(01_01_01) (상위어: 누리다 - 맛보다) (용례: 아이에게 용돈을 주다)
	우리말샘	가다듬다	설레는 마음을 가다듬고 약속 장소에 나갔다. 가다듬다	정신을 가다듬고 다시 한번 해 봐. 가두다	송아지를 우리에 가두었다. 가로-눕다	팔베개를 하고 가로누워서 텔레비전을 보았다.
키워드	한국어 표준 교육 과정	14개 범주 85개 항목 (1) 개인 신상: 이름, 전화번호, 가족, 국적, 고향, 성격, 외모, 연애, 결혼, 종교

한국어 U-WIN은 영어 WordNet에 대응하는 어휘의미망으로, 단어의 상위어-하위어 관계와 관련 용례 정보를 함께 제공한다. 우리말샘은 한국어 사용과 관련하여 풍부한 용례를 제공하므로 이야기 말뭉치 구축에 참조할 수 있다. 국제 통용 한국어 표준 교육과정 적용 연구는 한국 고유의 생활과 배경에 맞게 구성된 주제어 목록으로 일상생활 속의 사건이나 이야기를 부연하는 데 유용하므로 포함되었다.

그러나, 다양한 문형 확보와 작업 수월성 담보를 위한 참조 목적에서 관련 단서가 제공되므로 생성된 이야기가 반드시 제공된 단서와 일치하거나 단서를 포함해야 하는 것은 아니다.

<그림 4. 그림 기반 작업 방식>

Clue 1: 축구공/잡다
Clue 2: 아이가 운동장에서 축구공에 손을 올려놓았다.

<그림 4>은 그림 기반 작업 방식의 예시로 이미지 단서가 제공된다. 이미지는 직접적인 시각적 정보를 제공하므로 다른 자유 창작/핵심어 기반 방식과 구별된다. 단, 이미지 단서 이외에 다른 단서를 제공하지 않는 것이 아니라, 텍스트 단서도 함께 주어진다. 예를 들어, '축구공/잡다'와 함께 '아이가 운동장에서 축구공에 손을 올려놓았다'라는 문장이 제시될 수 있다.

그림 기반 방식은 기존에 구축된 유사 문장 생성 말뭉치를 활용한다. 유사 문장 생성 말뭉치는 그림 속의 인물, 행위, 장면을 묘사하는 캡션(caption)을 작업자가 직접 작성한 말뭉치이다. 이야기 완성 말뭉치와 달리 유사 문장 말뭉치는 가설을 생성하거나 별도의 가설 평정을 수행하지 않았고, 시간의 흐름이 나타나지 않는 장면을 기술하였다. 그러나, 유사 문장 말뭉치는 시각적 정보를 단서로 텍스트를 작성하였다는 점에서 자유 창작/핵심어 기반 방식을 보완하기 위하여 활용하였다.

말뭉치 구축 과정에서 맞춤법 오류와 오탈자는 검수하여 교정하되, 한국어 모국어 사용자의 언어 직관에 수용된다면 기계 학습의 견고성 측면에서 허용하였다. 다시 말해, 맞춤법 오류 및 오탈자를 검수하였지만 언중들이 주로 혹은 자연스럽게 많이 사용하는 오류 유형들은 기계 학습의 견고성을 위해 따로 수정하지는 않았다.

(2) [단서 1: 드러눕히다 / 단서 2: 유 선달은 그를 드러눕히고 아랫배를 만져 보았다.]

최 참판은 막걸리를 배 터지게 마셨다. 최 참판의 아내는 최 참판을 **마루에 들어눕혔**

다. 최 참판의 아내는 불룩 튀어나온 최 참판의 배를 꼬집었다.

위와 유사한 언어 사용은 인터넷 기사에서도 발견된다. '마루에 들어눕혔다'처럼 '들어 침상에 눕혔다'는 표현이 있다.

(3) 부처님은 병든 비구의 머리를 들고 아난은 두 다리를 **들어 침상에 눕혔다***.

다른 예로, '사실과 틀리다'는 어휘의 사전적 쓰임새와 맞지 않지만 SNS에나 웹 블로그 글에서는 (4)와 같은 표현들이 종종 발견된다.

(4) 가. (...)적어놓은 건 대부분 **사실과 틀리다**.
나. (...)팩트는 없다. 경제 또한 **사실과 틀리다**.

(3)과 (4)와 같은 표현이 종종 사용되며, 언중들이 자주 자연스럽게 사용하는 표현이다. 때문에 기계 학습의 견고성 측면에서 모두 교정하지 않고 부분적으로 허용하였다.

2) 이야기 생성 사례

또한, 단서의 유의어를 사용하여 단서에서 제시하는 쓰임의 사전적 의미와 다른 쓰임으로 이야기를 생성할 수 있다. 또한 다른 쓰임이 사전의 표제어 포함되지 않을 수 있다.

(5) [단서 1: 변동되다(→ 바뀌어 달라지게 되다) / 단서 2: 출발하기 하루 전에 계획이 변동되었다.]
나는 회의에 참석하기 위해 버스를 타고 가고 있었다. 가던 도중 상사로부터 회의 장소가 **변경되었다(→ 다르게 바뀌어 새롭게 고쳐지다)**는 소식을 들었다. 그래서 버스에서 하차한 뒤 변경된 장소로 향하는 지하철을 탔다.

(6) [단서 1: 뜨다 (→ 감았던 눈을 벌리다.) / 단서 2: 나는 아침에 눈을 뜨기가 무섭게 약수터로 달려갔다.]
노인이 헉헉대며 산 속 약수터에 도착했다. 노인은 쭈글쭈글한 손으로 물통 뚜껑을 땄다. 노인은 물통으로 약수를 **뜨기(-> 물속이나 지면 따위에서 가라앉거나 내려앉지 않**

* https://m.jejukcr.com/@/bbs/list/10123074

고 물 위나 공중에 있거나 위쪽으로 솟아오르다) 시작했다.

(7) [단서 1: 훌쩍대다. (→ 액체 따위를 남김없이 자꾸 들이마시다.) / 단서 2: 밥은 먹지도 않고 연방 국물만 훌쩍댄다.]
한 여성이 도서관에서 책을 읽다가 **훌쩍대기(-> 콧물을 자꾸 들이마시다)** 시작했다. 옆에 있던 사람들이 그녀를 쳐다보았다. 사람들의 시선을 느낀 그녀는 책으로 얼굴을 가렸다.

(8) [단서 1: 걸다 (→ 다른 사람을 향해 먼저 어떤 행동을 하다.) / 단서 2: 그는 사소한 일에 시비를 걸어 주먹을 휘두르곤 했다.]
효겸은 검은색 말에 10만원을 **걸었다(-> 돈 따위를 계약이나 내기의 담보로 삼다)**. 검은색 말은 초반부터 일등을 뺏기지 않고 결승선에 도착했다. 효겸은 자신이 건 돈의 3배를 땄다.

이와 함께, 이야기의 사건은 시간의 흐름을 표상하는 것을 원칙으로 한다. 이는 (1) 이야기의 흐름 속에서 장면의 전환이나 시간의 흐름이 뚜렷하게 드러나는 것을 말하며, (2) 장면 전환은 없더라도 행동의 연속이나 시간의 흐름이 언어 표현으로 명확하게 드러나지 않는 것은 지양한다. 다만, 시간의 흐름에 대한 작업자의 직관은 구체화하여 규정하기 어려우므로 작업자의 판단에 위임한다. 이야기의 장면 전환이 없는 짧은 순간에 발생한 사건이거나, 시간의 흐름이 언어 표현에서 명확하게 드러나지 않더라도 사건의 다양성 측면에서 허용한다.

(9) [단서 1: 어리다 / 단서 2: 앞들 무논 위에 아지랑이가 어리기 시작한다]
강하고 단단한 쇳조각으로 부싯돌을 쳤다. **한참을 치니** 돌에서 연기가 어리기 시작했다. 연기가 점점 자욱해지더니 마침내 불이 붙었다.

(10) [단서 1: 허비하다 / 단서 2: 노름판에서 세월을 허비하다]
그는 술에 세월을 허비했다. 그렇게 그는 알코올중독자가 되었다. **10년 후** 그는 결국 간암에 걸렸다.

예문 (9)는 “한참을 치니”라는 행위가 지속되었음을 나타내는 표현을 사용함으로써 이야기의 시간적 흐름을 나타냈다. 예문 (10)은 “10년 후”라는 시간 표현을 사용함으로써 행위의 결과를 적절히 표상하였다.

이와 달리 예문 (11)과 (12)는 행위의 지속성이나 결과를 직접적으로 나타내는 표현이

없지만, 연속된 행위가 하나의 사건을 구성하고 있으므로 허용한다.

(11) [단서 1: 공공서비스 / 단서 2: 전화/인터넷 등 통신 관련 이야기]
책상 위에 있던 진우의 핸드폰이 울렸다. 진우는 책상 앞으로 가서 핸드폰을 집어들었다. 진우는 핸드폰을 귓가에 대고 전화를 받았다.

(12) [단서 1: 듣다 / 단서 2: 두통에 잘 듣는 약]
나는 하루종일 복통에 시달리다가 약국에 갔다. 약사는 나에게 복통에 잘 드는 약을 주었다. 나는 약국에서 바로 그 약을 먹었다.

제 3 장

가설 생성

1. 가설 추론 대상과 범위

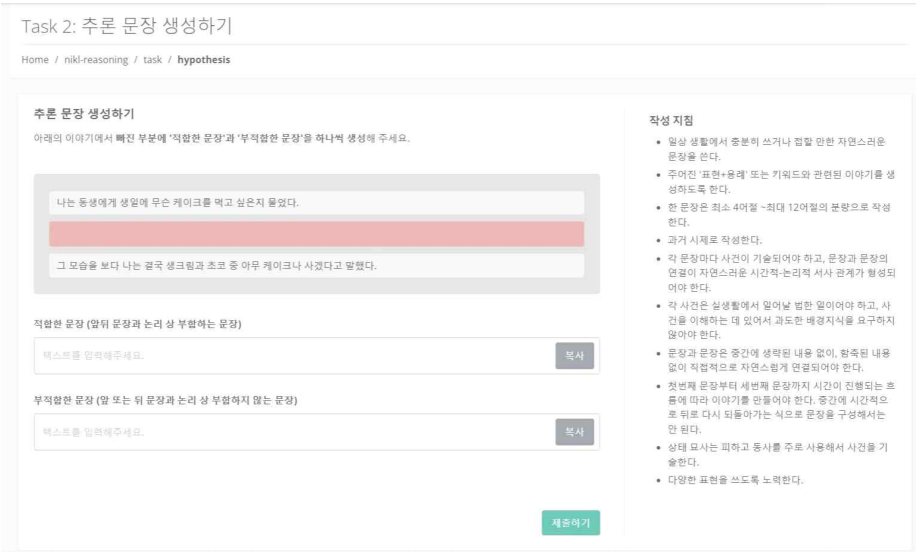
가설 추론은 주어진 가설의 유형과 성격에 따라 매우 다양한 양상을 보인다. 본 과제에서 가설 추론의 대상은 '이야기'이며, 이야기는 어떤 사건의 시간적 흐름과 인과적 연관성을 함축한다. 이러한 추론적 특성에 근거하여 기계가 상식 추론을 할 수 있는가를 평가하기 위한 데이터로 이야기를 활용하였다(Bhagavatula et al. 2020). 상식이란 넓은 의미에서 사회 상규(常規)와 통념(通念)을 일반 상식으로 볼 수도 있으며, 좁은 의미에서 백과사전식 지식이나 명문화된 윤리적 규정을 일반 상식으로 볼 수도 있다. 본 사업에서는 이야기 상식 추론 또는 가설 추론을 인공 지능 언어능력평가의 관점에서 가추적 추론(abductive reasoning)으로 한정한다. 가추적 추론은 논리적 추론의 하위 유형으로 연역적 추론이나 귀납적 추론과 유사한 성격이 있지만, 핵심적인 특징은 그 창발성(創發性 또는 emergent property)에 있다고 볼 수 있다(Peirce, 1965).

2. 가설 생성 방법과 사례

1) 가설 생성 방법

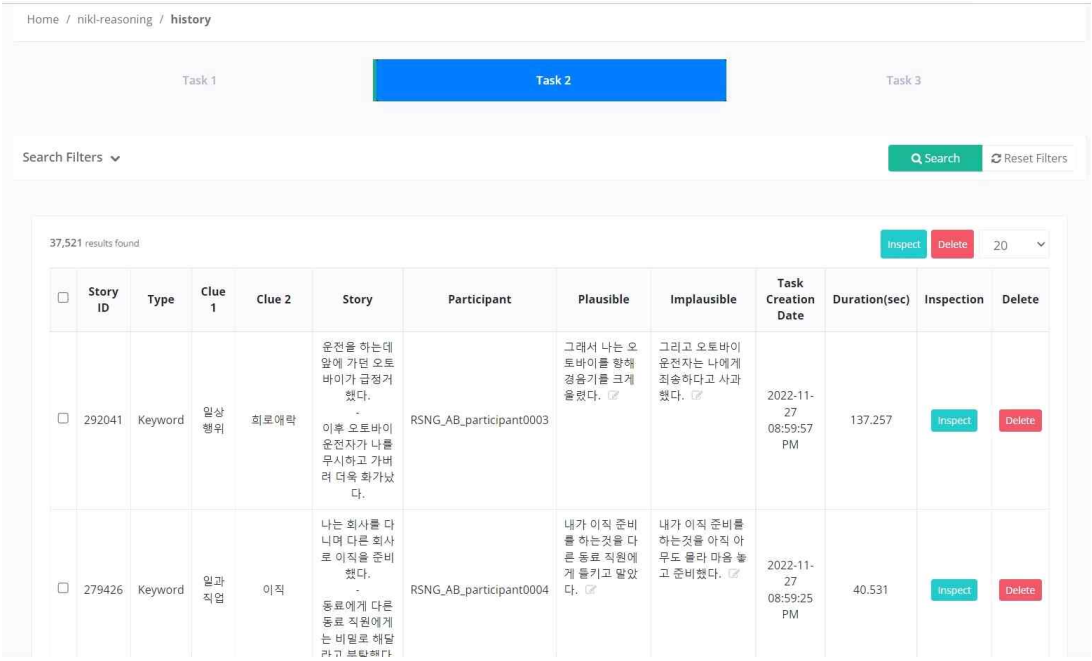
이야기 생성 과업(Task 1)과 동일한 플랫폼에서 가설 생성 과업(Task 2)을 수행할 수 있는 워크벤치를 탑재하였다. 아래 크라우드소싱 플랫폼에서 작업자는 작업 방식에 따라 가설 생성 화면의 구성을 달리한다(<그림 5>).

<그림 5. 가설 생성 작업 화면>

NO.	작업자 화면
1	 적합한 문장 - 적합한 문장은 이야기의 빠진 부분(작업자 화면 기준으로 분홍색 빈칸)에 들어갔을 때 앞뒤 문장과 논리상으로 부합하는 문장이다. 부적합한 문장 - 부적합한 문장은 이야기의 빠진 부분(작업자 화면 기준으로 분홍색 빈칸)에 들어갔을 때 앞뒤 문장과 논리상으로 부합하지 않는 문장이다. 작성 지침 - 문장 생성 시에 개별 작업자가 참고 및 준수해야 하는 지침이다. 문장 생성 시에 주의해야 할 사항을 정리하여 작업자의 작업 안정성을 높이고 부적절한 문장의 생성이 최소화될 수 있도록 하였다. 제출하기 - 문장 생성 종료 시에 결과물을 저장하고 다음 작업으로 넘어가기 위한 용도이다.

가설 생성에 대한 관리자 화면을 아래 제시하였다(<그림 6>). 아래 크라우드소싱 플랫폼에서 관리자는 작업 결과를 검수하여 교정할 수 있다.

<그림 6. 가설 생성 관리자 화면>

<table>
<tr><th>NO.</th><th>관리자 화면</th></tr>
<tr><td>2</td><td> 

검수 완료(Inspect) - 생성된 문장에 대하여 부적절 여부를 검수하여 통과 판정을 받은 경우에 해당한다.
검수 미통과 자료 삭제(Delete) - 부적절하거나 모호한 내용으로 인해 검수에 통과하지 못한 경우에는 삭제된다.</td></tr>
</table>

가설 생성 구축 목표는 30만 개 이상의 문장을 제작하는 것이다. 구체적인 지침은 붙임 1에 제시되어 있다.

2) 가설 생성 사례

아래는 가설 생성 과정에서 선별한 사례이다. 좋은 사례는 가추적 추론 능력을 평가한다는 본 사업의 목적에 부합하는 사례들이다. 가추적 추론은 백과사전적 지식과 달리 일반 상식에 비추어 부합하는 지식을 말한다. 즉, 명시적인 지식이 아니므로 이해하려면 묵시적인 지식이 필요하다. 나쁜 사례는 가추적 추론 능력 평가 목적에 비추어 적절하지 않은 사례들이다.

① 좋은 사례

첫 번째 사례의 경우, 일상생활에서 빈번하게 발생하는 상황과 관련된 추론이다. 인공지능이 사건의 흐름을 이해하는지 평가하기 위한 적절한 문제라고 할 수 있다. '휴지'가 떨어진 상황을 이해하고 휴지를 구매할 수 있는 장소가 '슈퍼'인지 추론할 수 있어야 한다.

NO.	좋은 상식 추론 과제 문장 세트 예시		
1	선행 문장	화장실 두루마리 **휴지**가 다 떨어졌다.	
	가설 추론	**적절**	하교하는 길에 **슈퍼**에 들렸다.
		부적절	하교하는 길에 **화장실**에 들렸다.
	후행 문장	두루마리 **휴지**를 구매하고 집으로 돌아왔다.	

두 번째 사례의 경우, 가설을 추론하기 위하여 표층적 정보인 '사이다'와 '콜라'에 의존하여서 답을 맞힐 수 없으며 이야기의 선후 관계를 이해해야 한다. 이런 추론은 표층적인 정보에 의존하여 학습한 인공 지능이 풀 수 없는 문제이므로 인공 지능의 평가 목적을 고려할 때 유용하다.

NO.	좋은 상식 추론 과제 문장 세트 예시		
2	선행 문장	레스토랑에서 **에이드**를 다 마셔서 점원에게 리필을 요청했다.	
	가설 추론	**적절**	점원은 **사이다**로 잔을 채워서 갖다주었다.
		부적절	점원은 **콜라**로 잔을 채워서 갖다주었다.
	후행 문장	나는 점원에게 **사이다말고 콜라**로 리필을 해달라고 말했다.	

세 번째 사례의 경우, ‘토마토’와 ‘양상추’는 소스가 아니라 채소에 해당한다는 사실을 인지할 수 있어야 한다. 그러나 해당 예시는 ‘토마토’와 ‘양상추’, ‘채소’의 어휘적 관계(하위어-상위어) 등의 정보로 추론할 수 있다. 관련된 추론은 어휘적 지식에 가깝지만, 사건의 흐름을 이해하는데 필요한 어휘적 지식이라는 점에서 인공 지능 평가 목적에 부합한다고 볼 수 있다.

NO.	좋은 상식 추론 과제 문장 세트 예시		
3	선행 문장	찬혁이는 샌드위치 가게에서 미트볼 샌드위치를 주문했다.	
	가설 추론	**적절**	직원은 찬혁이에게 어떤 **채소**를 빼드릴까요라고 질문했다.
		부적절	직원은 찬혁이에게 어떤 **소스**를 빼드릴까요라고 질문했다.
	후행 문장	찬혁이는 **토마토와 양상추**를 빼달라고 직원에게 요청했다.	

네 번째 사례의 경우, 대화 상황이 함축하는 내용을 어휘 의미에서 추론할 수 있는지 묻는 것이다. ‘후회하다’라는 표현에는 인공 지능이 제대로 기능하지 않는다는 것을 내포하며 이를 이해해야 추론할 수 있다.

NO.	좋은 상식 추론 과제 문장 세트 예시		
4	선행 문장	나는 **인공 지능 스피커에게 음악을 재생하라는 명령**을 내렸다.	
	가설 추론	**적절**	인공 지능 스피커는 **날씨를 알려주었다.**
		부적절	인공 지능 스피커는 **음악을 재생했다**.
	후행 문장	나는 이런 모습을 보며 인공 지능 스피커를 산 것을 **후회했다**.	

다섯 번째 사례의 경우, 끓는 물이 차가운 물에 영향을 줄 수 있음을 인지할 수 있어야 한다. 아래 문제는 일상생활에서 흔히 볼 수 있는 또는 일반 상식 선에서 알 수 있는 과학적인 현상을 이해하는지 묻고 있다는 점에서 유용하다.

NO.	좋은 상식 추론 과제 문장 세트 예시		
5	선행 문장	나는 주전자에 **물을 끓였다**.	
	가설 추론	**적절**	나는 **차가운 컵에 끓는 물**을 따랐다.
		부적절	나는 **차가운 컵에 차가운 물**을 따랐다.
	후행 문장	그러자 갑자기 **컵이 쨍 소리를 내며 갈라졌다**.	

여섯 번째 사례의 경우, 특별한 문맥이 제공되지 않는다면 일반 상식에 비추어 아이들이 담배보다 휴대폰 게임에 더 관심을 기울일 것이라는 사실을 추론할 수 있어야 한다. 예를 들어, 아이들이 비행 청소년이라거나 '학교에서 흡연을 했었다'라는 문맥적 단서가 제공되지 않는다면 일반적인 기대에 비추어 아이들이 '휴대폰 게임을 한다'는 것이 그럴듯한 가설이라고 볼 수 있다.

NO.	좋은 상식 추론 과제 문장 세트 예시		
6	선행 문장	여러 명의 아이들이 놀이터에서 뛰어놀고 있었다.	
	가설 추론	**적절**	나는 벤치에 앉아서 휴대폰 게임을 하고 있었다.
		부적절	나는 벤치에 앉아서 담배를 피우고 있었다.
	후행 문장	아이들은 벤치로 다가와 나에게 말을 걸었다.	

일곱 번째 사례의 경우, '빠르게 번지다'라는 표현이 비유적인 의미가 아니라 실제 불이 번진다는 의미로 사용되었을 때 적절하다는 것을 추론할 수 있어야 한다.

NO.	좋은 상식 추론 과제 문장 세트 예시		
7	선행 문장	담배꽁초로 인해 **큰 불이 났다**.	
	가설 추론	**적절**	산에 난 **불이 빠르게 번져** 걷잡을 수 없게 되었다.
		부적절	불이 났다는 **소문이 빠르게 번져** 걷잡을 수 없게 되었다.
	후행 문장	산이 거의 다 타고 나서야 불이 꺼졌다.	

여덟 번째 사례는 첫 번째 사례와 같이 유사한 사례라고 볼 수 있다. 시간의 흐름에 비추어 적절한 추론이 가능하지만. 그녀가 언니에게 반찬을 가져다 주었으므로 반찬을 배달시킬 필요가 없다는 사건의 일반적인 인과 관계를 추론할 수 있어야 한다.

NO.	좋은 상식 추론 과제 문장 세트 예시		
8	선행 문장	그녀는 자취하는 언니에게 반찬을 가져다주러 갔다.	
	가설 추론	**적절**	그녀는 언니네 집에서 밥상을 차렸다.
		부적절	그녀는 언니네 집에서 반찬을 배달시켰다.
	후행 문장	그녀는 가져간 반찬과 함께 언니와 밥을 먹었다.	

아홉 번째 사례의 경우, 초대를 받아 음식을 먹었다는 인과 관계가 성립하기 위해서는 음식에 대한 언급이 필요하다는 사실을 추론할 수 있어야 한다.

NO.	좋은 상식 추론 과제 문장 세트 예시		
9	선행 문장	민서는 친구네 집에 초대를 받아 방문했다.	
	가설 추론	**적절**	친구는 민서를 위해서 직접 음식을 해주었다.
		부적절	친구는 민서를 위해서 직접 네일아트를 해주었다.
	후행 문장	민서는 음식을 먹고 친구에게 엄지손가락을 들어 보였다.	

위 9개 가설 추론 과제는 인공 지능이 이야기의 맥락과 흐름을 이해하는지 평가하는 목적에 부합한다는 점에서 참조할 만하다.

② 나쁜 사례

이하 가설 추론 과제는 인공주석물에 의한 부정적인 영향이 있거나, 평정에 통과하지 못한 예문들이다. 잘못 만들어진 가설 추론 과제의 전반적인 특징은 이야기의 흐름이나 맥락에 비추어 두 가설 추론이 모두 적절하거나 모두 적절하지 않다는 점이다. 이 때문에 평정자의 직관에 따라 판단이 엇갈린다.

첫 번째 예문은 적절한 가설 추론 문장과 부적절한 문장이 모두 이야기의 흐름에 비추어 자연스럽기 때문에 평정자의 직관에 따라 모두 답이 될 수 있다.

NO.	나쁜 상식 추론 과제 문장 세트 예시		
1	선행 문장	돌부리에 걸려 선물받은 슬리퍼가 끊어져버렸다.	
	가설 추론	**적절**	그래서 나는 다시 집으로 돌아갔다.
		부적절	그래서 나는 급한 김에 슬리퍼의 끊어진 부분을 테이프로 붙였다.
	후행 문장	집에 와서 나는 끊어진 슬리퍼를 버렸다.	

두 번째 예문은 두 가설 추론 사이에 문형적인 차이가 존재하지만, 관련된 차이가 가

설의 적절함/부적절함을 구별하는 것에 도움이 되지 않는다.

NO.	나쁜 상식 추론 과제 문장 세트 예시		
2	선행 문장	그의 일을 도와주는 데 상당히 많은 시간이 소요되었다.	
	가설 추론	**적절**	그래서 그에게 더 이상 도와주기는 어려울 것 같다고 말했다.
		부적절	그래서 일을 다 끝내고 그에게 더 도울 일이 없냐고 물어보았다.
	후행 문장	그는 지금까지 도와준 것만으로도 고맙다고 했다.	

즉, 완곡한 거절과 도움을 제공하겠다는 친절이 모두 자연스럽기 때문에 두 추론 문장의 (부)적절성이 명확히 구별되지 않는다. 따라서 가설 생성 의도와 달리 잘못 만들어진 인공 지능 평가 문제이다.

세 번째 예문은 대명사 ‘그’, ‘그녀’가 반복적으로 사용되어 이야기를 이해하기가 매우 어렵다.

NO.	나쁜 상식 추론 과제 문장 세트 예시		
3	선행 문장	그녀는 그 사람의 정체가 무엇인지를 제멋대로 추측해 보았다.	
	가설 추론	**적절**	그 남자는 친구의 남자친구가 아니었다.
		부적절	그 남자는 친구의 남자친구였다.
	후행 문장	나는 그 모습을 보고 친구가 남자친구가 생겼다는 것을 추측했다.	

작업자는 대명사의 과도한 사용을 지양하고, 이야기의 흐름을 명확히 기술하여야 한다.

네 번째 예문은 어떤 척도에서 크다/작다를 비교하여 가설 추론 문장을 제작하는 방법이 가설의 적절함/부적절함을 구별하는 단서가 되기 어려운 사례이다.

NO.	나쁜 상식 추론 과제 문장 세트 예시		
4	선행 문장	지난 신체검사에서 나는 여동생과 키 차이가 3cm 났다.	
	가설 추론	**적절**	여동생의 키가 나보다 3cm 더 컸다.
		부적절	여동생의 키가 나보다 3cm 더 작았다.
	후행 문장	나는 더 크지말라고 장난으로 여동생의 머리를 눌렀다.	

이와 유사하게, 다섯 번째 예문은 어떤 척도에 시간적 흐름이 빠름/느림을 비교하여 가설의 적절함/부적절함을 구별하였으나 평정을 통과하지 못하였다.

NO.	나쁜 상식 추론 과제 문장 세트 예시		
5	선행 문장	그는 시리얼에 우유를 부었다.	
	가설 추론	**적절**	그는 전화를 받으러 거실로 나갔고 시간이 꽤 걸렸다.
		부적절	그는 시리얼을 가지러 주방에 갔고 금새 돌아왔다.
	후행 문장	다시 돌아오자 시리얼 건더기는 전부 바닥으로 가라앉아 있었다.	

이상과 같이 추론 과제 문장 세트를 제작하는 과정에서 참조할 만한 예문과 주의해야 할 예문을 정리하였다.

제 4 장

가설 평정

1. 가설 평정 대상과 범위

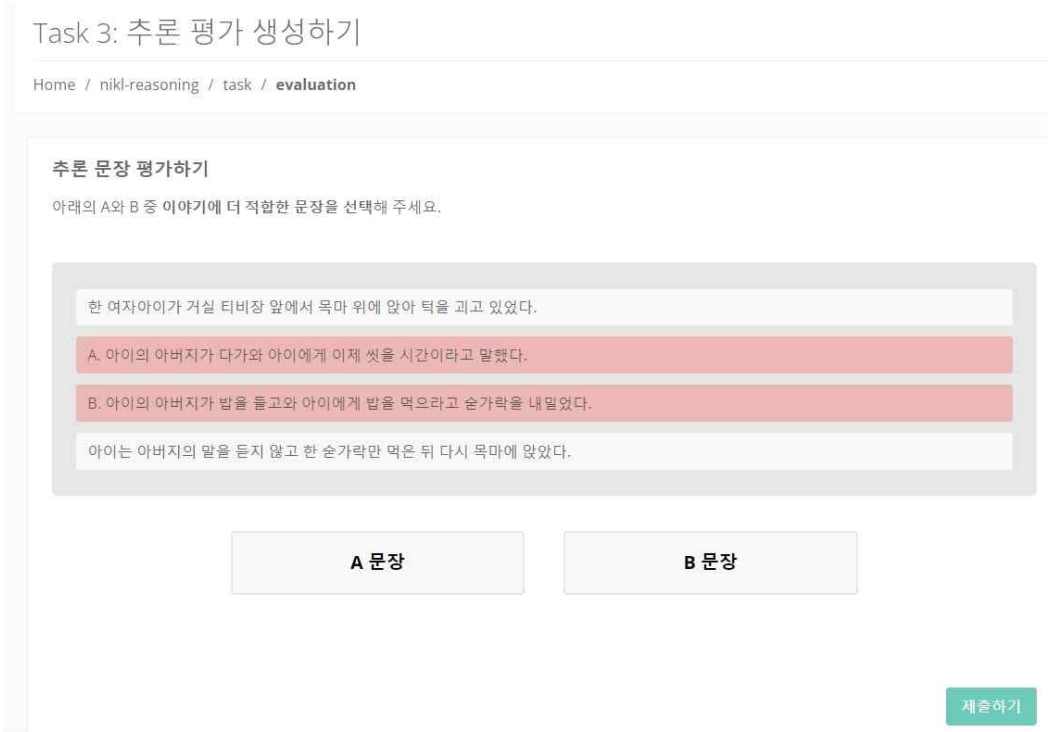
가설 평정은 문장 세트의 (부)적절한 추론 문장쌍이 제작된 의도에 부합하는지 평정하는 것이다. 평정 작업은 상식 추론 과제 문장 세트를 활용하여 이루어진다. 상식 추론 과제 문장 세트 제작에 참여하지 않은 작업자(또는 크라우드워커)를 각 문장 세트별로 5인을 배정하여 가설 평정을 수행하였다. 이때, 과반수(3명) 이상의 평정자가 판단하기에 기존 문장 세트의 적절한 추론 문장이 문맥에 비추어 부적절한 가설 추론으로 읽힌다면 그 문장 세트는 교정한다.

2. 가설 평정 방법과 사례

1) 가설 평정 방법

상식 추론 과제 문장 세트를 만드는 과정에서 작업자는 적절한 추론과 부적절한 추론을 구별하여 생성하였다. 그러나 가설 평정 과정에서는 해당 문장의 (부)적절성 여부는 노출되지 않는다. 즉, 평정자는 어떤 문장 쌍에 대하여 문맥적 흐름에 부합하는 문장을 고르는 것이다. 이를 수행하기 위한 가설 평정 작업 환경은 <그림 7>과 같다.

<그림 7> 가설 평정 작업자 화면

NO.	작업자 화면
1	
	적절 가설과 부적절 가설이 매 시도마다 문장 A와 문장 B의 위치에 무작위로 배치되었다.

2) 평정 미통과 사례와 통계 분포

향후 유사 사업 과제 수행을 돕기 위하여 평정 수행과정에서 발생한 오류 또는 평정 미통과 사례 또한 제시하고자 한다. 평정 미통과 사례로 분류하는 기준은 5명의 평정자 중에서 3인 미만의 평정자가 “부적절한 가설”로 생성된 문장을 적절하다고 혼동하는 것을 기준으로 한다. 예를 들어, 가설 생성 단계에서 적절한 가설 A와 부적절한 가설 B가 생성되었을 때, 어떤 것이 적절한 가설인지 모르는 상황에서 1-2인의 평정자가 적절한 가설 A가 부적절한 가설로 읽힌다고 판정하는 것이다.

평정 미통과 사례 중 일부는 최종 납품된 이야기 말뭉치와 상식 추론 과제에는 포함되지 않으므로 두 가지 유형으로 나누어 예시와 함께 평정을 통과하지 못한 이유를 설명한다. 다만, 아래 예시들은 상식 추론 과제의 평정이 종료된 이후에 의도된 정답과 평정이

엇갈리는 예시들이다. 즉, 작업자와 평정자의 직관이 일치하지 않는 사례이다. 따라서 평정을 통과하지 못하였더라도 다른 상위 검수자가 검토하여 수정하였다면 최종 납품 자료에 포함될 수 있다.

아래 유형은 평정자에게 가설의 적절성과 부적절성이 명확하지 않은 유형이다. 적절한 가설과 부적절한 가설 양자 모두가 이야기 문장을 자연스럽게 완성할 수 있으므로 평정자의 직관이 일치하지 않아 평정 미통과 사례로 분류되었다.

<table>
<tr><th>NO.</th><th colspan="3">평정 미통과 사례 1</th></tr>
<tr><td rowspan="4">1</td><td>선행 문장</td><td colspan="2">그래서 우리는 물에 빠져 허우적대고 있었다.</td></tr>
<tr><td rowspan="2">가설 추론</td><td>적절</td><td>그때 구조 헬기가 우리를 구하러 왔다.</td></tr>
<tr><td>부적절</td><td>우리는 구조 헬기가 오기도 전에 사망했다.</td></tr>
<tr><td>후행 문장</td><td colspan="2">우리는 구조 헬기가 오기를 간절히 기다렸다.</td></tr>
</table>

가설 추론 과제 제작 단계에서 의도된 부적절한 가설은 “우리는 구조 헬기가 오기도 전에 사망했다”였으나, 일부 평정자는 부적절한 가설 또한 이야기의 흐름에 비추어 자연스럽다고 생각했다. 후행 문장이 부적절한 가설이 진술하는 사건에 대한 감상으로 읽힐 수 있다는 것이다. 작업자나 평정자의 직관이 일치하지 않는 것은 자연스러운 일이지만, 인공 지능 평가용 문제로는 부합하지 않으므로 배제하였다.

위 사례와 동일한 유형으로, 적절한 가설과 부적절한 가설 양자 모두가 ‘거절하다’의 이유를 설명할 수 있으므로 평정자마다 직관이 일치하지 않아 평정 미통과 사례로 분류되었다.

NO.	평정 미통과 사례 2		
2	선행 문장	친구에게 조언을 해 주려다 화를 내고 말았다.	
	가설 추론	**적절**	결국 친구에게 사과를 했다.
		부적절	나는 친구의 고민을 잘 들어줄 자신이 없었다.
	후행 문장	친구의 고민을 들어줄 자신이 없어 거절했다.	

이처럼 이야기의 흐름을 완성할 수 있는 적절한 가설과 이야기의 맥락을 해치는 부적절한 가설이 평정자에게 명확하지 않는다면 해당 문장 세트는 배제하였다.

아래 예시의 유형은 생성한 이야기가 자연스러운 흐름에 어긋난 유형이다. 선행 문장과 후행 문장의 인과 관계를 알 수 없기 때문에 문장 간의 연결이 자연스럽지 않아 평정자의 직관이 어긋난 평정 미통과 사례로 분류되었다.

NO.	평정 미통과 사례 3		
3	선행 문장	전기 파리채를 이용하여 모기를 잡았다.	
	가설 추론	**적절**	모기가 또 들어올 것이 걱정되어 창문을 닫아 버렸다.
		부적절	모기를 잡는 사이 한 마리가 더 들어왔다.
	후행 문장	전기파리채로 모기 한 마리를 잡았다.	

마지막으로 적절 가설과 부적절 가설 양자 모두 이야기의 인과관계를 설명할 수 없으므로 평정 미통과 사례로 분류되었다.

NO.	평정 미통과 사례 4		
4	선행 문장	그러나 그는 정치를 하는 법은 몰랐다.	
	가설 추론	**적절**	결국 그는 간신들에 의해 처형이 되었다.
		부적절	그래서 그는 순식간에 이름을 날렸다.
	후행 문장	하지만 아무도 그의 성과를 알아주지 않았다.	

위와 같이, 적절한 가설이라고 볼 수 있는 가설이 없음에도 불구하고, 가설 추론 과제 단계에서 잘못 제작된 가설은 평정 미통과 사례로 분류되어 말뭉치에 포함시키지 않았다.

가설 추론 단계에서 제작되었으나 평정을 통과하지 못한 사례를 중간 관리하여 관련 통계적 분포를 확인하였다. 제작된 가설 추론 과제 138,694건에 대해서 다음과 같은 평정 통과 비율이 산출되었다(<표 11>, 사업 종료 기준).

<표 11. 평정 통과 비율 분포표>

적절 가설 선택 수	평정 통과 여부	가설 추론 과제 누계	비율	누적비율
0	미통과	92	0.06%	0.06%
1	미통과	183	0.12%	0.18%
2	미통과	380	0.25%	0.43%
3	통과	1,409	0.93%	1.36%
4	통과	11,550	7.58%	8.94%
5	통과	138,694	91.06%	100.00%

평정 미통과 사례는 전체 수량 대비 0.4%인 것으로 추정되며 평정 미통과 사례 555건은 상위 검수자가 (부)적절 가설을 교정하여 2차 평정을 수행하였다. 2차 평정은 1차 평정과

동일한 작업 환경에서 평정자를 교체하여 진행되었다.

제 5 장

검증 평가

1. 최종 결과물 산출

본 사업의 최종 결과물은 이야기 완성 말뭉치는 150,176건으로 모두 JSON 형식으로 변환하여 납품하였다. 또한 별도의 과제로 활용할 수 있게끔 이야기 완성과 상식 추론 과제를 분리하였다(<그림 8>과 <그림 9> 참조). 상식 추론 과제는 150,275건이 구축되었고 마찬가지로 JSON 형식으로 납품하였다.

본 사업의 과업 수행은 (1) 이야기 제작과 (2) 가설 생성, 그리고 (3) 가설 평정으로 구성되어 있으며, 해당 문서에서는 이러한 공정의 결과물을 “이야기 완성 평가 세트”라고 한다. 이때 이야기의 경우, 자연스러운 시간적 흐름과 사건의 인과관계에 따라 배열된 3개의 문장을 하나의 단위로 한다. 가설의 경우, (1)에서 제작된 이야기의 흐름을 자연스럽게 설명하는 적절한 가설과 흐름에 어색한 부적절한 가설을 포함한 2개의 문장을 단위로 한다. 평정 과제의 경우, (2)에서 생성된 두 가지 가설 중에서 평정자의 직관과 상식에 비추어 그럴듯한(plausible) 가설을 선택하는 것을 원칙으로 한다. 구체적으로는 이야기의 첫 번째 문장과 세 번째 문장을 먼저 제시하고, 공란으로 제시된 두 번째 문장의 자리에 올 수 있는 가설을 선택하는 것이다.

위와 같은 과업 수행 절차에 따라서 구축된 이야기 완성 평가 세트 1건은 아래와 같이 JSON 형식으로 표현된다 (<그림 8>).

<그림 8> 이야기 완성 말뭉치 JSON 형식

NO.	이야기 완성 말뭉치 JSON 형식
1	

이야기 완성 말뭉치 JSON 형식의 경우, 최상위의 “id”는 이야기 완성 말뭉치 세트에서 이야기의 식별 번호를 부여하는 기능을 한다. “metadata”는 “제목(title)”, “생성자(creator)”, “배포자(distributor)”, “년도(year)”, “범주(category)”와 같은 국어원 말뭉치 공통 형식에 관한 정보를 포함한다. 이야기 완성 말뭉치의 경우, “제목(title)”, “생성 방식(type)”을 포함하며 이야기를 구성하는 문장 3개 모두(sentence 1, sentence 2, sentence 3)가 JSON 형식에 포함되어 있다.

가설 추론 과제 세트 1건 또한 JSON 형식으로 표현된다 (<그림 9>).

<그림 9> 가설 추론 과제 JSON 형식

NO.	가설 추론 과제 JSON 형식
2	

가설 추론 과제 JSON 형식의 경우, 최상위의 “id”는 가설 추론 과제 세트에서 이야기의 식별 번호를 부여하는 기능을 한다. “metadata”는 “제목(title)”, “생성자(creator)”, “배포자(distributor)”, “년도(year)”, “범주(category)”와 같은 국어원 말뭉치 공통 형식에 관한 정보를 포함한다. 가설 추론 과제는 제목/생성 방식을 포함하고, 이야기의 첫 번째 문장(sentence 1)과 마지막 문장(sentence 3)이 포함되어 있다. 가설 추론 과제의 정답 세트에는 (부)적절 가설쌍((im)plausible)과 평정(count)이 포함되어 있다.

2. 인공 지능 예비 평가 결과

1) 인공 지능 평가 개요

본 사업이 구축한 이야기 완성 말뭉치와 상식 추론 과제는 인공 지능 학습용 말뭉치 또는 인공 지능 언어 능력 평가용 말뭉치로 활용된다. 이에 적합하게 JSON 형식으로 가공하여 기계 학습에 활용할 수 있도록 하였다. 또한, 가설 평정 점수는 사람 수행(Human performance) 점수로 활용될 수 있다. 사람 수행 점수는 인공 지능 예측의 정확도를 산출하는 준거 기준이 된다.

2) 언어 모형의 학습과 평가

이야기 완성 말뭉치 데이터를 활용하여 구축한 상식 추론 과제가 인공 지능 평가 목적에 부합하는지 검증하기 위하여 인공 지능 평가를 진행했다. 이를 위해 대표적인 한국어 기반 BERT 모형인 KR-BERT와 KLUE-BERT를 이용하였다. 이 두 언어 모형은 사전 학습이 완료된 모형이다.

사전학습 완료된 언어 모형의 학습 매개 변수를 가설 추론 과제에 적합하게 미세조정(fine-tuning)하여 인공 지능 모델 평가를 진행하였다. 공정 수행 일정을 고려하여 이야기 완성 말뭉치 구축 공정이 90% 이상 완료된 시점에서 인공 지능 평가를 수행하였다. 구축 목표 150,000건에서 1차 검수가 완료된 138,761건의 데이터를 활용하였다. 80%(111,008건)는 훈련 데이터로, 10%(13,876건)는 검증 데이터로, 나머지 10%(13,877건)은 평가 데이터로 활용하였다.

미세조정이 완료된 이후에는 정답 라벨(label)에 대한 BERT 언어 모형의 예측 정확도를 측정하였다. 평가 데이터 10%는 그림 기반 방식을 통해 만들어진 데이터 13,876건으로 미세조정에 포함되지 않은 아웃 도메인(out-domain) 데이터이다. 아웃 도메인 데이터는 사전학습과 미세조정 단계에서 언어 모형이 사람처럼 새로운 사실에 대하여 합리적인 유추를 할 수 있는지 평가하기 위하여 선별된다. 기계학습에서는 일반적으로 인 도메인(in-domain) 데이터를 활용하여 과제 훈련과 평가를 수행한다. 예를 들어, 감성 분석 정확도를 평가하기 위하여 새로운 감성 분석 말뭉치를 구축하는 것이 아니라, 동일한 감성 분석 말뭉치를 임의로 분할하여 훈련과 평가를 수행한다. 이와 달리, 아웃 도메인 데이터는 서로 다른 성격의 말뭉치를 2종 이상 구축하여, 훈련과 평가에 서로 다른 데이터를 활용하는 것이다. 이러한 방법은 성능이 크게 향상된 언어 모형에게 이전에 비하여 어려운 평가 과제를 부여함과 동시에, 인간이 변화하는 환경 속에서 자신의 지식과 기술을 조금씩 조정해야 하는 상황을 모방한 것이라고 볼 수 있다.

사전학습을 진행하기 이전 평가 데이터에 대한 언어 모형의 정확도는 KR-BERT에서 0.48, KLUE-BERT에서 0.54로 매우 낮은 수준으로 나타났다. BERT를 기반으로 한 모델 중 높은 성능을 보이는 두 모델이 50% 수준의 정확도를 보이는 것은 이야기 완성 말뭉치를 활용한 가설 추론 과제가 매우 어렵다는 것을 보여준다.

이는 Mostafazadeh et al. (2016)에 이미 지적된 바와 같이, 언어 모형이 매우 큰 텍스트 말뭉치를 학습하더라도 이야기의 자연스러운 흐름을 이해하기 어렵다는 것과 연관이 있다. 즉, 다른 주변 단어의 분포로 단어의 의미를 유추하는 언어 모형의 학습이 비명시적으로 출현한 암묵적인 지식을 학습하기가 어렵다는 사실과 관련이 깊다. 따라서 이야기 완성 말뭉치를 추가로 학습하여 관련 취약점을 보완할 수 있다.

미세조정을 진행하여 아래의 체크 포인트를 평가에 활용할 모형으로 선정했다. 학습 과정에서 더 높은 검증 정확도와 낮은 손실값을 보이는 구간이 있었지만, 데이터에 대한 과적합을 방지하기 위하여 손실값을 고려하여 아래의 체크 포인트에서 평가 결과를 보고한다 (<표 12>).

<표 12. 인공 지능 예비 평가 결과>

모델명	학습 단계	손실값	학습률	검증 정확도	평가 정확도
KR-BERT	27,000	0.2878	2e-5	0.87	**0.84**
KLUE-BERT	27,000	0.2371	2e-5	0.89	**0.81**

미세조정 이후 평가 데이터를 통해 두 모델의 평가 정확도를 구한 결과, KR-BERT는 0.84, KLUE-BERT에서 0.81이라는 정확도를 보여주어 향상된 수행 점수를 보여주었다. 이는 구축한 상식 추론 과제 데이터가 인공 지능 평가 목적으로 활용될 수 있음을 검증한 것이다. 또한, 미세조정을 진행하기 이전의 정확도보다 각각 75%, 50%의 향상 폭을 보인 것이다.

이후, 아웃 도메인 데이터를 활용하여 평가를 진행했다. 이 데이터의 경우 미세조정 학습에 사용된 데이터와는 다른 방식으로 구축된 데이터로, 미세조정 학습 과정에서 전혀 활용되지 않았다. 그렇기 때문에, 이 데이터를 통한 평가는 이번 연구를 통해 구축된 이야기 완성 데이터가 더 넓은 다른 영역의 데이터에도 충분한 성능을 보일 정도로 견고하게 구축된 데이터인지를 보여줄 수 있다. 평가 결과 KR-BERT는 0.90, KLUE-BERT에서 0.86이라는 성능을 보여주었다. 이러한 결과는 학습에 사용된 인 도메인 데이터에 대한 정확도보다 높은 수준의 성능을 보여주는 것으로, 미세조정에 사용된 데이터의 품질과 견고성을 보여준다고 할 수 있다.

3) 인공주석물(Annotation Artifact)

다음에서는 인공 지능의 언어능력평가 과정에서 잠재적으로 인공주석물의 영향을 받을 수 있는 대표적인 예시를 제시하였다. 인공주석물은 의도된 추론 과정을 거쳐서 정답을 맞히는 것과, 휴리스틱(heuristic)으로 정답을 맞히는 것과의 차이가 나지 않도록 만드는 부정적인 단서를 말한다. 이 때문에, 인공주석물은 인공 지능 평가 결과를 부풀리거나 왜곡하는 요인으로 지적되었다.

이상적으로는 인공주석물이 모두 제거되는 것이 바람직하지만 실제 평가 데이터 구축 과정에서 모두 제거하기에는 어렵다. 이는 인공주석물이 작업자 또는 데이터 라벨러(data labeler)가 주석 처리하는 과정에서 작업의 부담을 덜기 위하여 활용하는 습관이나 패턴 전체를 일컫는 의미로 이해되기 때문이다. 예를 들어, 함의 분석에서 작업자가 전제 문장과 논리적으로 양립할 수 없는 가설 문장을 제작하기 위하여 반복적으로 부정어(negation)를 삽입하는 것이 대표적인 인공주석물의 사례이다. 또한, 중립(neutral) 문장을 제작하기 위하여 유정물 주어(animate subject)나 대명사를 임의의 다른 무정물 주어(inanimate subject)로 바꾸는 문장도 인공주석물의 사례라고 볼 수 있다.

이러한 패턴이 평가 데이터에서 반복된다면, 부정어 삽입이나 무정물 주어 사용이 발견되면 인공 지능이 평가의 의도와 관계없이 모순(contradiction)이나 중립 라벨을 기계적으로 선택하여 평가 결과를 왜곡할 수 있다. 특히, 인공주석물이 통제되지 않을수록 평가 점수가 상승하는 경향이 있다고 알려져 있으므로, 인공 지능의 실제 자연어 이해 수준보다 평가 점수가 부풀려질 수 있다.

따라서 본 사업에서는 인공주석물에 관한 선행연구를 준용하여 관련 내용을 작업자 교육과 주석 가이드라인에 일부 포함시켰다. 또한 인공주석물이 작업자의 주석 패턴이나 반복적인 어휘선택 등 습관에서 비롯될 수 있으므로 최대한 다양한 문장을 제작하도록 작업자를 교육하였다. 다만, 인공주석물의 비율을 작업자 교육을 통하여 최대한 통제하는 것이 목적이므로 구체적인 비율을 산정하기 어렵고, 인공주석물과 '유의미한 학습 패턴' 간의 경계가 뚜렷한 것은 아니므로 작업자가 인공주석물의 부정적인 영향을 인지하고 이를 통제하여 다양한 이야기를 제작할 수 있도록 독려하는데 초점을 두었다.

첫 번째 유형의 인공주석물은 이야기의 인과 관계에 대한 이해가 없어도 선행 문장과 후행 문장에 주어진 동일한 단어 '축구공'이 포함된 가설을 선택하여 정답을 맞힐 수 있도록 하는 부정적인 단서이다.

NO.	인공주석물 사례 1		
1	선행 문장	나는 **축구공**을 잠시 내려놓았다.	
	가설 추론	**적절**	그러자 **축구공**이 밑으로 데굴데굴 굴러갔다.
		부적절	그러자 **야구공**이 밑으로 데굴데굴 굴러갔다.
	후행 문장	나는 **축구공**을 주우러 갔다.	

위와 같은 인공주석물을 작업자가 인지하여 적절/부적절 가설 제작 단계에서 피하는 것이 바람직하다.

두 번째 사례의 경우, 성별에 따른 구분인 ‘아내’와 ‘남편’을 구별하면 이야기의 인과 관계에 대한 이해가 없어도 정답을 맞힐 수 있다.

NO.	인공주석물 사례 2		
2	선행 문장	병에 걸려 미각을 잃은 나에게 **남편이** 밥을 해 달라고 했다.	
	가설 추론	**적절**	그래서 나는 **남편에게** 국을 해 주었다.
		부적절	그래서 나는 **아내에게** 밥을 해 주었다.
	후행 문장	**남편은** 국이 달다며 나에게 소리를 지르며 화를 냈다.	

위와 같은 인공주석물은 단순 단어 일치 또는 일부 절의 일치로 인한 인공주석물에 비하여 부정적인 영향이 덜하다고 볼 수 있다. 이야기의 맥락과 흐름에 깊이 고민하는 것에 비하여, 여성 명사와 남성 명사를 구별하는 휴리스틱이 생각의 부담이 덜하기 때문이다. 이러한 휴리스틱 사용은 가설 추론과 관련이 없다는 점에서 인공주석물이다.

세 번째 사례의 경우, 이야기의 인과 관계에 대한 이해가 없어도 ‘돌아보다’라는 술어의 다의적인 의미를 포착하면 정답을 맞힐 수 있다.

NO.	인공주석물 사례 3		
3	선행 문장	뒤에서 나를 부르는 소리를 들었다.	
	가설 추론	**적절**	그래서 나는 **뒤를 돌아봤다**.
		부적절	그래서 나는 내 **삶을 돌아봤다**.
	후행 문장	뒤에는 나의 친구가 나에게 손을 흔들고 있었다.	

이처럼 이야기의 흐름과 상관없이 어휘의 다의적인 의미를 포착하면 정답을 맞힐 수도 있다. 이야기의 흐름을 이해하기 위한 추론이 생략될 수 있다는 점에서 가설 추론 과제 작업에서 주의해야 할 인공주석물이라고 볼 수 있다.

네 번째 사례의 경우, 성별을 표상하는 '남자친구'라는 단어는 여성 작업자 비율이 높은 작업자 그룹에서 고빈도로 등장하였다.

NO.	인공주석물 사례 4		
4	선행 문장	유선이는 **남자 친구의** 우직한 모습을 보고 결혼을 결심했다.	
	가설 추론	**적절**	**남자 친구에게** 결혼을 하자고 프로포즈를 할 생각이었다.
		부적절	유선이는 결혼을 하자고 프로포즈를 해주기를 바랐다.
	후행 문장	적당한 날 유선이는 **남자 친구에게** 결혼하자고 프로포즈를 했다.	

위 사례는 이야기 이해와 관련된 추론이나 어휘적인 지식, 휴리스틱과 관련 없이 작업자의 특성과 관련이 있다. 예를 들어, 작업자의 구성이 대부분 여성이라면 특정 이야기 주제(예: 네일샵)나 인칭 표현(예: 남자친구)이 빈번하게 등장한다. 이러한 요인은 의도치 않은 인공주석물 또는 데이터 편향이 발생하는 대표적인 요인이다. 따라서 사전에 작업자의 성별, 연령, 학력 등을 고려해야 할 필요가 있다.

제 6 장

맺음말

1. 결론

본 사업은 인공 지능의 언어능력평가를 목적으로 (i) 이야기 완성 말뭉치와 (ii) 가설 추론 과제를 구축하였다. 말뭉치와 과제는 15만 건의 데이터를 포함하였다. 이야기 완성 말뭉치는 150,176건(450,528문장)이 구축되어 목표 구축량인 15만 건을 초과 달성하였다. 가설 추론 과제는 150,275건(300,550문장)이 구축되어 목표 구축량인 15만 건을 초과 달성하였다. 가설 추론 과제는 751,375건의 평정 정보(작업자 5인 × 150,275건)를 활용하여 구축되었다. 공정은 단계적으로 수행되었으며, 이야기 생성, 가설 생성, 그리고 가설 평정의 3단계를 거쳤다. 이를 통하여 본 사업은 고품질 한국어 말뭉치를 구성하였으며, 이를 가공하여 인공 지능 평가 과제로 제안하는 사업의 목적을 성공적으로 수행하였다.

본 사업의 결론은 다음과 같다. 첫째, 말뭉치와 추론 과제의 검증을 목적으로 수행한 인공 지능 예비 평가 결과, 한국어 인공 지능은 상식 추론 능력이 취약하였다. 이는 본 사업에서 구축한 평가 과제의 난도가 높다는 점을 보여준다. 둘째, 이야기와 추론의 범주를 명확하게 정의하기 어려웠다. 이야기는 일상 사건을 기반으로 문장들 사이에 인과 관계가 존재한다. 그러나 인과 관계는 백과사전적 지식이 아닌 한국어 일반 상식이므로 명시적인 형태나 구체적인 사례를 기술하여 명확한 지침을 기술하기 어렵다. 이러한 이유로, 세세한 지침을 제시하지 않고 다양한 문장을 포괄하도록 구체적인 서술적 제한을 두지 않았다.

이를 통하여 본 사업은 인공 지능 평가를 고도화하는 목적을 수행하였으며, 향후 한국어 인공 지능의 학습 말뭉치에 포함되어 인공 지능의 인과 관계 추론 능력 향상에 이바지하리라고 믿는다.

2. 제언

이야기의 사전적 정의는 ‘어떤 사물이나 사실, 현상에 대하여 일정한 줄거리를 가지고 하는 말이나 글’을 일컫는다. 그러나 어떤 내용이나 주제를 명확히 기술할 것인지, 또는 줄거리의 구조를 어떻게 기술할 것인지에 따라 이야기의 성격이 매우 달라진다. 따라서 이야기의 성격을 명확히 규정하고 이를 어떻게 기술할 것인지 세세하고 구체적인 지침이 필요하다. 다만, 본 사업은 한국어 이야기 말뭉치를 구축하고 가공하는 선도적인 사업으로서, 15만 건이라는 대량의 한국어 데이터를 수집하는 것에 중점을 두었다. 향후 이야기 말뭉치와 추론 과제를 보다 면밀히 검토하고 지침을 세분화하여 추론 과제 평가를 고도화하여야 한다. 또한 보다 실용적인 인공 지능 언어 능력 평가가 가능하도록 해야 할 것이다.

<붙임 1. 이야기 완성 말뭉치 2022 구축 통합 지침>

1. 이야기 완성

<공통 작업 지침>

- 제시된 키워드 또는 동사와 관련된 일상생활 관련 이야기를 3문장으로 만든다.
- 3문장은 시간 순서에 따른 상황 변화를 보여 주는 것이어야 한다.
- 각 문장마다 동사를 주로 사용해서 사건을 기술하고, 각 문장이 연결되어 자연스러운 서사 관계를 형성하도록 한다.
- 각 사건은 실생활에서 일어날 법한 일이어야 하고, 일상생활에서 충분히 쓸 법할 만한 자연스러운 문장으로 표현한다.
- 각 문장은 과거 시제로 작성한다.
- 한 문장은 최소 4단어~최대 12단어 정도의 분량으로 작성한다.
- 문장 마지막에는 마침표가 들어가야 한다. 물음표나 느낌표로 끝나는 문장은 쓰지 않는다.

1.1. 키워드(keyword) 기반 방식

<작업 개요 및 지침>

- 제시된 키워드와 관련된 일상생활 관련 이야기를 3문장으로 만든다.
- 이야기는 키워드에 부합하도록 하며, 예컨대 키워드가 '일상행위-설명'으로 제시된 경우, 오른쪽 키워드, 즉 '설명'에 대한 이야기를 만든다.
- 키워드로는 아래 표에서 제시한 139개 항목이 제시된다.
- 키워드와 관련하여 전형적으로 떠오르는 상황뿐 아니라 키워드와 관련된 다양한 상황을 자유롭게 떠올려 이야기를 만든다 (<표 1>).

표 1. 한국어 학습자 교육과정을 활용한 주제어 목록

범주(15범주)	항목(139개)
주거와 환경	집 꾸미기/집 정리, 이사, 방 꾸미기/방 정리, 가구.침구, 주거비, 생활 편의 시설, 거주 지역에서의 활동, 동물 관련 에피소드, 식물 관련 에피소드

일상생활	가정에서의 에피소드, 학교에서의 에피소드, 외모.복장 관련 이야기, 종교 활동, 생일 등 각종 기념일, 하루 일과
쇼핑	쇼핑 시설 이용, 식품, 의복, 가정용품
식음료	음식, 음료, 배달, 외식
공공 서비스	공공기관 이용 관련 이야기, 우편 관련 이야기, 전화/인터넷 등 통신 관련 이야기, 은행 관련 이야기, 병원 관련 이야기, 약국 관련 이야기, 경찰서/소방서 관련 이야기
여가와 오락	휴일, 취미 생활, 라디오, 텔레비전, 영화, 공연, 전시회, 박물관, 독서, 스포츠, 방학, 휴가
일과 직업	진로, 취업, 면접, 직장 생활, 수입, 이직
대인 관계	친구 관계 관련 이야기, 동료 관계 관련 이야기, 선후배 관계 관련 이야기, 가족 관계 관련 이야기, 초대, 대접, 방문, 약속, 모임, 클럽/커뮤니티 활동, 연애, 결혼
건강	위생, 질병, 치료, 운동, 다이어트, 운동/다이어트 외의 건강 관리
기후	날씨, 봄, 여름, 가을, 겨울
여행	여행 계획, 여행 준비, 여행지에서의 이야기, 고향 방문
교통	길 찾아가기, 교통수단 이용, 운송, 택배
교육	교과목 학습, 자기 계발, 진로 탐색, 진학, 유학, 시험
뉴스, 시사 문제	뉴스에서 접할 수 있는 유형의 사건과 사고, 재해, 환경 문제
일상행위	설명, 진술, 보고, 묘사, 서술, 기술(記述), 확인, 비교, 대조, 수정, 문답, 제안, 권유, 요청, 경고, 충고, 조언, 허락, 명령, 금지, 주의 주기, 지시, 동의, 반대, 부인, 추측, 문제 제기, 의도, 바람.희망.기대, 가능/불가능, 능력, 의무, 사과, 거절, 만족/불만족, 걱정, 고민, 위로, 불평.불만, 후회, 안도, 놀람, 선호, 희로애락, 인사, 소개, 감사, 축하, 칭찬, 환영, 호칭

<예시>

- 키워드: 쇼핑 / 의복 (‘쇼핑’ 중에서도 ‘의복’ 쇼핑과 관련한 이야기를 만들어야 함)
S1 >> 친구와 함께 백화점에 쇼핑을 가서 옷을 사려고 했다.
S2 >> 6층 매장에 올라가니 마네킹에 걸린 파란 옷이 바로 눈에 띄었다.
S3 >> 가격표를 보고 놀랐지만 옷이 너무 마음에 들어서 할부 결제로 구매했다.

- 키워드: 주거와 환경 / 생활 편의 시설
S1 >> 오랜만에 동네 목욕탕에 갔다.
S2 >> 탕 속에서 반신욕을 하고 세신사에게 세신도 받았다.
S3 >> 개운함을 느끼며 집으로 돌아왔다.

1.2. 자유 창작(creative) 방식

<작업 개요 및 지침>

- 제시된 동사를 세 개의 문장 중 하나에서 사용하여 이야기를 만든다.
- 주어진 동사를 그대로 쓰지 않고 변형을 해도 무방하다. (예: 능동사 '자르다'가 주어졌을 때 피동사 '잘리다'를 쓰는 것도 가능. '기억하다'가 주어졌을 때 명사 '기억'을 쓰는 것도 가능.)
- 동사 및 용례는 국립국어원 <우리말샘>에서 추출된 것이다.

<예시>

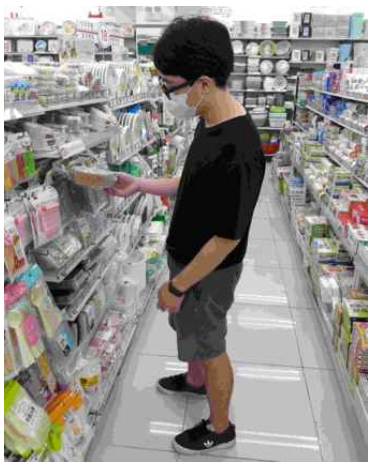
- 동사: 돕다 / 용례: 아버지의 일을 돕다 S1 >> 아버지께서 저녁 준비를 하고 계신 어머니를 도우려고 부엌으로 가셨다. S2 >> 어머니는 혼자서 준비하는 게 편하다고 하셨다. S3 >> 실망한 아버지는 소파로 돌아와 텔레비전을 켜셨다.

1.3 사진(photo) 기반 방식

<작업 개요 및 지침 >

- 사진을 묘사하는 한 개의 문장을 보고 이어지는 두 개의 문장을 만들어 이야기를 완성한다.

<예시>



문장 ① (주어진 문장) : 남성이 매장 안에 있는 물건을 고르고 있었다.

문장 ②: 곁에 있던 친구가 다른 곳을 가자고 말했다.

문장 ③: 남성은 물건을 내려놓고 가게를 나왔다.

- 사진을 묘사하는 문장① 의 시제는 기계적으로 과거시제로 변경하였으며 고유명사 및 불필요한 설명은 삭제하였다. 이어지는 문장도 모두 과거시제로 작성한다.
- 과업에 사용한 사진은 총 1만 개이다. 각 사진마다 상황을 묘사하는 서로 다른 5개의 문장이 존재한다.
- 총 5만 개(1만 개의 사진x5개의 문장)의 사진과 설명문 쌍(pair)을 가지고 이야기를 생성하는 단서(clue)로 활용한다.
- 같은 사진과 설명문이 주어지는 경우에도 이야기를 생성하는 작업자가 다르기 때문에 다른 이야기가 생성될 수 있다.

<예시>

사진 묘사 문장 1 : 수영복을 입은 세 사람이 바다를 향해 걸어가고 있었다.
사진 묘사 문장 2 : 한 남성이 튜브를 들고 바닷가를 걷고 있었다.

- 자연스러운 이야기를 생성하지 못할 것 같은 사진과 설명문은 사용하지 않고 자연스러운 이야기를 생성할 수 있는 사진과 설명문만을 사용하여 작업한다.
- 사진 기반 문장생성 방법의 장단점은 아래와 같다.
- 장점: 자연스러운 일상을 포착한 사진을 사용하였기 때문에 인위적이지 않고 자연스러운 이야기 구조를 만들 수 있다.
- 단점: 밥을 먹고 있는 사진, 사진을 찍고 있는 사진 등 대상 및 배경이 비슷한 사진이 많기 때문에 생성할 수 있는 이야기가 제한적일 수 있다.

- 사진 기반 방식으로 작업자가 생성한 이야기의 편향(bias)을 알아봄으로써 과업의 신뢰도(reliability)를 측정하는 목적으로 사용한다.
- 이 방식은 Out-domain 데이터 구축 방식으로, 기존의 In-domain(키워드 등) 방식과 유사하지만 다른 방법으로 이야기를 만든다.
- 기존 In-domain 방식의 이야기 완성 작업을 해본 적이 없는 작업자들이 분리된 주석 작업자로 이야기를 생성한다.
- 이러한 작업방식을 통해 언어 모형이 동일한 상황에서 미묘하지만 다른 가설의 차이를 판별하는 능력을 평가한다.

<작업 지침>

- 문장은 너무 길게 쓰지 않는다. 최소 4어절에서 최대 12어절 길이의 문장을 만든다.
- 문장은 과거시제로 작성한다. 사진을 묘사하는 첫 번째 문장은 기계적으로 과거시제로 변경하였지만, 만약 변경이 되지 않았을 경우 직접 과거시제로 수정하여 작성한다.

<예시>

수정 전: 한 여성이 바다 옆길을 걷는다.
수정 후: 한 여성이 바다 옆길을 걸었다.

문장에 주어인 '한 여성', '한 남자' 등은 고유이름, 직업명으로 바꿔서 작업해도 된다.

<예시>

수정 전: 한 남성이 식당에 앉아서 음료를 마시고 있었다.
수정 후: 준호가 식당에 앉아서 음료를 마시고 있었다.

- 문장과 문장은 중간에 생략된 내용 없이 직접적으로 자연스럽게 연결되어야 한다.
- 두 번째 문장이 첫 번째 문장과 세 번째 문장을 자연스럽게 이어야 된다.
- 첫 번째 문장에서 세 번째 문장으로 바로 넘어가는 것이 자연스러운 경우에도, 두 번째 문장이 있어서 더욱 자연스러운 이야기가 되는 것이 이상적이다.

<예시>

문장 ①: 영철이가 옷걸이에 걸린 티셔츠를 들고 있었다.
문장 ②: 영철이는 티셔츠를 들고 계산대로 다가갔다.
문장 ③: 영철이는 고양이가 그려진 셔츠를 한 벌 구매했다.

- 첫 번째 문장부터 세 번째 문장까지 시간이 진행되는 흐름에 따라 이야기를 만들어야 한다.
- 중간에 시간적으로 뒤로 다시 되돌아가는 식으로 문장을 구성해서는 안 된다.

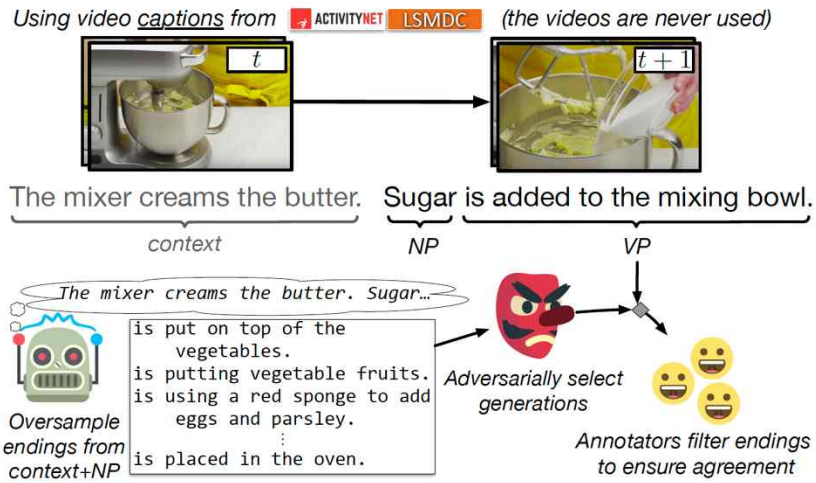
<*예시>

문장 ①: 할아버지가 회갑연에서 노래를 부르고 있었다.
문장 ②: 회갑연에 온 다른 사람들이 신나서 의자에 앉아 들썩거렸다. → (o)
문장 ②: 할아버지가 노래를 부르기 전 다른 사람들은 음식을 먹었다. → (x)
문장 ③: 사회자는 다른 분들도 다 같이 춤추셔도 좋다고 했다.

- 문장의 난이도는 초등학교 저학년 학생도 이해할 수 있는 정도의 수준이어야 한다. 사건을 이해하는데 있어서 과도한 배경 지식을 요구하지 않아야 한다.
- 너무 독특한 상황은 피하고 일상생활에서 일어날 법한 일이어야 한다.
- 상태 묘사는 피하고 동사를 주로 사용해서 사건을 기술한다. 동적인 이야기를 구성 할 수 없는 경우에는 작업하지 않고 다음 이야기로 넘어간다.
- 이미 주어진 문장으로 다음 두 문장이 자연스럽게 이어질 수 있는 이야기만 생성한다.
- 영상을 설명하는 문장 하나가 있고, 그다음 문장을 이어갈 수 있는 명사와 동사구가 여러 개 주어진다. 언어 모형은 명사와 동사구로 만들어진 문장 중 더욱 자연스러운 이야기를 만드는 문장을 선택한다(<그림 1>).

* 해당 예시는 개인정보보호를 위하여 모자이크 처리함.

그림 1. SWAG 상식 추론 문장 세트 예시



- 주어진 문장을 약간 변경하였을 때 더욱 자연스러운 이야기가 만들어질 수 있는 경우, 주어진 사진과 관련된 내용의 범위 안에서 수정을 해도 된다.
- 최대한 다양한 표현을 쓰도록 노력한다. 같은 단어 혹은 상황이 반복되지 않도록 한다.
- 주어진 상황으로 만들 수 있는 이야기가 생각이 나지 않거나 비슷한 이야기가 반복될 것 같을 때에는 다음 이야기로 넘어가도록 한다.

② 가설 부착

- 부적절 가설은 첫 번째 문장과 세 번째 문장을 자연스럽게 잇는 문장이 아니어야 한다.
- 부적절 가설이 이야기를 자연스럽게 잇는 문장인 경우에도, 적절 가설이 문맥상 더욱 자연스러운 문장이어야 된다.
- 부적절 가설은 적절 가설을 약간만 변형하도록 한다.
- 문장 길이에 따라 2단어에서 4단어 정도의 차이만 있도록 한다. 이는 인공주석물(annotation artifacts)로 작용할 수 있는 요소들을 최소한으로 제한하기 위함이다.

<예시>

문장 ①: 모자를 쓴 남자가 자신의 곁에 강아지를 두고 돌 의자에 앉아 있었다.
문장 ② 적절 가설: 강아지가 갑자기 의자에서 내려와 앞으로 달리기 시작했다.
문장 ② 부적절 가설: 강아지가 갑자기 의자에서 내려와 앞발로 머리를 긁었다.
문장 ③: 남자도 강아지를 따라 일어나서 뛰기 시작했다.

- 언어 모형이 실제로 문장의 인과관계를 이해하는 것이 아니라 문체의 특징(stylistic feature)에 의존하여 이어지는 문장을 선택하였다는 것이 선행연구에서 밝혀졌다 (Sharma et al., 2018).
- 문체의 특징을 최소한으로 포함하여 더욱 정확한 언어모델의 성능 평가를 하기 위한 지침을 세운다.

    1) 문장은 같은 주제 안에 있어야 된다.
    2) 비슷한 어조와 감성을 띄고 있어야 된다.
    3) 2-3글자 이상 차이가 나지 않도록 한다.

- 적절 가설에서 단순히 not(~않았다)로 바꾸는 방식으로 부적절 가설을 만들지 않도록 한다.

<예시>

문장 ①: 영수가 욕실에서 강아지를 샤워시키고 있었다.
문장 ② 적절 가설: 강아지가 욕실에서 나가기 위해 발버둥을 쳤다.
문장 ② 부적절 가설: 강아지가 욕실에서 나가기 위해 발버둥을 안쳤다. → X
문장 ③: 영수는 아랑곳하지 않고 샤워기를 가져다 댔다.

- 상황에 맞지 않는 뜬금없는 대명사가 나오지 않도록 한다. 이는 인공주석물로 작용할 수 있을 뿐만 아니라 데이터의 신뢰도(Reliability)에 영향을 주기 때문이다.

<예시>

문장 ①: 재현이가 도로에서 자전거를 타고 있었다.
문장 ② 적절 가설: 재현이의 친구가 뒤에서 다가오며 소리를 질렀다.
문장 ② 부적절 가설: 재현이의 친구가 뒤에서 다가오며 햄버거를 던졌다. → X
문장 ③: 그 바람에 재현이는 놀라 자전거에서 떨어졌다.

- 동물과 식물 등 인간이 아닌 대상에 의인화는 배제한다. 추론 작업에서 문장의 참 거짓을 판단하는 데 방해 요소로 작용할 수도 있기 때문이다.

2. 가설 추론

<작업 개요>

- 다른 작업자가 만든 이야기 완성 결과물 중 문장 2가 삭제되고 문장 1과 문장 3만 제시된 것을 보고, 문장 1과 문장 3을 자연스럽게 이을 수 있는 '적절 가설' 문장과, 그와 유사한 부분이 있지만 문장 1과 문장 3을 잇기는 어려운 '부적절 가설' 문장을 작성한다.

<작업 지침>

- 부적절 가설은 문장 자체는 일상적으로 일어날 법한 일을 담고 있지만, 이야기의 흐름상 문장 1과 문장 3 사이에 놓이기 어려운 것이어야 한다.
- 적절 가설과 부적절 가설은 2~3단어를 공유하는 등 유사성이 있어야 한다.
- 적절 가설을 단순히 부정하는 것으로 부적절 가설을 만드는 것은 되도록 지양한다.
- 적절 가설과 부적절 가설의 문장 길이는 크게 차이 나지 않도록 한다.

<예시>

S1 >> 나는 밥을 먹으러 나갔다가 오랜만에 고등학교 동창을 만났다.
S2 >> _____

S3 >> 나는 동창과 기분 좋게 헤어졌다.

적절 가설 >> 고등학교 동창이 저녁을 사 주었다.
부적절 가설 >> 고등학교 동창이 내 뒤통수를 때렸다.

S1 >> 옷이 맞는지 확인하기 위해 탈의실로 갔다.
S2 >> _____
S3 >> 옷을 더럽힌 나는 어쩔 수 없이 그 옷을 살 수밖에 없었다.

적절 가설 >> 옷을 입다가 립스틱이 옷에 묻어 버렸다.
부적절 가설 >> 옷을 입다가 목이 늘어나 버렸다.

참고문헌

김중섭 외(2017), 국제 통용 한국어 표준 교육과정 적용 연구, 국립국어원.

Chandra Bhagavatula et al. (2020). Abductive Commonsense Reasoning. Paper presented at *International Conference for Learning Representation (ICLR)*.

Charles Sanders Peirce. *Collected papers of Charles Sanders Peirce*, volume 5. Harvard University Press, 1965.

Nasrin Mostafazadeh et al. (2016). A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories. *arXiv preprint arXiv:1604.01696*.

Pereira, L., Liu, X., Cheng, H., Poon, H., Gao, J., & Kobayashi, I. (2021). Targeted adversarial training for natural language understanding. *arXiv preprint* arXiv:2104.05847.

Sharma, R., Allen, J., Bakhshandeh, O., & Mostafazadeh, N. (2018). Tackling the Story Ending. Biases in The Story Cloze Test. ACL.

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7, 625-641.

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). SWAG: A Large-Scale Adversarial Dataset. for Grounded Commonsense Inference. EMNLP.

<기획·연구>

국립국어원 강미영 언어정보과장

국립국어원 유희정 학예연구사

국립국어원 박미은 연구원

국립국어원 이민주 연구원

국립국어원 정영은 연구원

<연구 참여자>

연구 책임자 송상헌(고려대학교 언어학과)

공동 연구원 이도길(고려대학교)

박현아(연세대학교)

최윤지(인하대학교)

장하연(부산외국어대학교)

김일환(성신여자대학교)

박진호(서울대학교)

정연주(홍익대학교)

김태우(부산대학교)

정슬아(성신여자대학교)

이영진((주) 플리토)

김도균((주) 플리토)

안상욱((주) 플리토)

전한길((주) 플리토)

연구 보조원 김푸른솔(서울대학교)

최진(서울대학교)

이강혁(서울대학교)

연구 보조원 김다미(서울대학교)

신운섭(고려대학교)

이정현(고려대학교)

박하율(고려대학교)

박권식(고려대학교)

홍승혜(고려대학교)

이규민(고려대학교)

추민영(고려대학교)

강예림(부산대학교)

반순웅(고려대학교)

신유경(고려대학교)

윤인경(고려대학교)

지해인(고려대학교)

정민화(고려대학교)

보조원 김서연(고려대학교)

박수빈(고려대학교)

권다회(고려대학교)

박준영(고려대학교)

이조은(고려대학교)

이찬영(고려대학교)

임가희(고려대학교)

김혜정(부산대학교)

석민정(부산대학교)

보조원 안시현(부산대학교)
이하영(부산대학교)
정창영(부산대학교)
강소미(부산외국어대학교)
홍채원(고려대학교)
이소곤(고려대학교)
김나영(부산외국대학교)
김효겸(부산외국어대학교)
노희진(부산외국어대학교)
한선아(고려대학교)
권은재(부산대학교)
김두연(고려대학교)
노하영(고려대학교)
도수안(고려대학교)
민정연(고려대학교)
방성은(고려대학교)
배진훈(고려대학교)
이다은(인하대학교)
이민형(건국대학교)
이세림(서강대학교)
조수민(고려대학교)
김가은(부산대학교)
윤서영(부산대학교)

보조원 권수연(성신여자대학교)
김수경(성신여자대학교)
김정언(성신여자대학교)
김유연(성신여자대학교)
김이준(성신여자대학교)
박제인(성신여자대학교)
박혜민(성신여자대학교)
신혜연(성신여자대학교)
이승은(성신여자대학교)
이예린(성신여자대학교)
이용현(성신여자대학교)
김연재(성신여자대학교)
정미리(성신여자대학교)
정서연(성신여자대학교)
정유정(성신여자대학교)
한이지(성신여자대학교)
김나은(인하대학교)
이수현(인하대학교)
인주연(인하대학교)
장수진(인하대학교)
조아라(인하대학교)
권어진(부산대학교)
조민영(중앙대학교)
홍정빈(부산대학교)

보조원 최서윤(부산외국어대학교)
김수인(고려대학교)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2022년 12월 16일
발행일: 2022년 12월 16일
인 쇄: 현대문화사

※ 이 보고서는 국립국어원의 용역비로 수행한 '2022년 이야기 완성 평가 말뭉치 연구 분석' 사업의 결과물을 발간한 것입니다.