국립국어원 2023-01-01

발간등록번호
11-1371028-000930-01

2022년 말뭉치 함의 분석 및 연구

연구책임자

김일환

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2022년 말뭉치 함의 분석 및 연구’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2022년 6월 ~ 2023년 1월

2023년 1월 19일

연구책임자: 김일환(성신여대)

연구 기관 성신여자대학교 연구산학협력단
고려대학교 산학협력단
충남대학교 산학협력단
㈜나라지식정보

연구 책임자 김일환
공동 연구원 강아름, 김태우, 박진호, 박현아, 송상헌, 송영숙, 이도길, 이지은, 장하연, 정슬아, 정연주, 조경찬, 최윤지

국문 요약

2022년 말뭉치 함의 분석 및 연구

이 사업은 한국어 인공지능의 과업 수행 능력 일반과 높은 상관관계를 갖고 있는 언어 추론인 '함의 분석'을 위해 신문 텍스트로부터 선행 담화를 추출하고, 이를 기반으로 적대적 사례를 생성하여 인공지능 평가용 벤치마크로 가공함으로써 자연어 이해 벤치마크의 취약성 문제를 보완하기 위한 방안을 마련하는 데 목적이 있다.

이를 위해 2021년 국립국어원 신문 말뭉치로부터 7차례의 라운드를 통해 모두 630,000건의 선행 담화를 추출하고 이를 대상으로 적대적 가설을 생성하였다. 적대적 가설 문장을 생성하기 위해서는 선행 연구들을 참조하여 6가지의 가설 생성 전략을 기반으로 한 구축 지침을 수립하여 적용하였고, 작업의 효율성을 최대한으로 높이기 위해서 작업용 벤치마크를 개발, 활용하였다. 이를 통해 모두 53,214개의 적대적 가설을 생성하였다. 한편 생성된 적대적 가설은 추론 판단의 정확성을 위해 2명의 작업자의 검수를 거치도록 하였다. 결과적으로 3인의 작업자가 추론 판정에 일치한 문장만이 다음 단계에 진입할 수 있게 되었다. 이렇게 작업자 간 검수까지 통과한 적대적 가설은 모두 38,018개에 해당하였다.

적대적 가설 데이터 세트가 인공지능 언어 모델의 강건성을 확보하는 데 유효한 것인지를 판단하기 위해 언어 모델의 속이기(fooling) 실험을 수행하였다. 최종적으로 KLUE-RoBERTa base와 KRBERT 두 개의 언어 모델에 대한 속이기 실험을 통과한 적대적 가설만으로 데이터 세트를 구성하였다. 이 데이터 세트는 모두 20,052개의 문장으로 구성되었다. 마지막으로 5개의 언어 모델을 활용하여 본 사업을 통해 구축된 데이터 세트가 충분히 '적절한' 데이터인지를 검증하였고, 실험 결과 본

사업을 통해 구축된 적대적 가설 데이터가 언어 모델의 성능 개선에 유의미한 결과를 보인다는 점을 확인하였다.

주요어: 언어 모델, 추론, 자연어 이해, 속이기(fooling), 적대적 가설, 함의, 모순, 중립, 말뭉치

Abstract

2022 Research and Analysis on Korean Corpus of Adversarial Natural Language Inference

This project aims to build natural language understanding (NLU) benchmark for the evaluation of Korean artificial intelligence (AI) with a focus on its ability to infer the truth or falsity of presupposed contents in sentences. This linguistic inference is strongly correlated with the Korean AI's general capacity to perform many downstream tasks. Using the corpus of newspaper texts, we extracted discourses and sentences containing some presupposed contents and generated adversarial examples that complement the fragility or vulnerability of current NLU benchmarks.

More specifically, we collected 630,000 discourses from the newspaper corpus of 2021 National Institute of Korean Language, up to seven rounds of data extraction procedures. We then employed six different strategies of generating adversarial hypothesis from newspaper discourses based on the previous research on NLU benchmarks. In addition, we developed crowdsourcing platforms to help generating adversarial hypothesis with manual annotation guidelines. With this approach, we created 53,214 adversarial hypothesis.

For all the generated hypothesis, two crowd-workers manually inspected whether annotated data is accurately created without fallacy of the inferential judgment of annotators. As a result, annotated data is not passed into the next step of procedures without the agreement of three crowd-workers that the created data meets the annotation guidelines. The total number of annotated adversarial hypothesis with the agreement of crowd workers is 38,018.

In the procedure of examining the usability of adversarial hypothesis, we performed language model 'fooling' that tests whether the model is robust. We tested two Korean language models called KLUE-RoBERTa-base and KR-BERT. Crucially, adversarial hypothesis that fails to pass the fooling test (or those models successfully defended the confusion of adversarial hypothesis) are discarded. The final version of adversarial hypothesis data consists of 21,300 examples. In the last procedure of our project, we used five different language models for assessing the reliability and usability of the 20,052 examples. We found that adversarial hypothesis data is 'proper' to standard criteria of NLU benchmarks since it meaningfully boosts the performance of tested language models.

Key words: language model, Inference, NLU(Natural Language Understanding), fooling, adversarial hypothesis, entailment, contradiction, neutral, corpus

차 례

제1장 서론

제2장 선행 담화의 추출

제3장 적대적 가설 문장의 생성

차 례

제4장 언어 모델 속이기(fooling) 실험

제5장 검증

제6장 결론

표 차 례

표 차 례

그림 차례

제 1 장

서론

1. 사업 개요

1.1. 사업의 배경과 필요성

인공지능 언어 모델의 일반적인 언어 이해 능력을 평가하는 GLUE (General Language Understanding Evaluation, Wang et al., 2018)은 총 9개의 과제로 이루어져 있으며 이 과제들은 크게 (1) 단일문장 과제, (2) 유사성과 어휘 변용 과제, (3) 추론 과제 등으로 분류될 수 있다. 이는 최신 자연어처리 모델인 BERT, GPT-3의 주요 성능 평가 지표로 기능해 왔지만, 최근 언어 모델들의 발전 속도가 빨라짐에 따라 SuperGLUE (A Stickier Benchmark for General-Purpose Language Understanding Systems, Wang et al., 2020)가 제안되기도 하였다. 그러나 GLUE와 SuperGLUE의 경우, 모두 한국어 언어 모델의 성능을 평가하기에는 적절하지 못한 면이 있다. 특히, 자연어처리 과업(task) 중 하나인 자연 언어 추론(Natural Language Inference, NLI)에서 문장 사이의 함의 관계를 추론하는 능력을 평가하는 한국어 데이터 세트가 충분하지 않다는 점이 중요하다.

기존의 한국어 NLI 데이터 세트는 영어 데이터를 번역하여 사용한다거나 (KorNLI, Ham et al., 2020) 한국어 언어 모델의 성능 평가를 위해 한국어 데이터를 가지고 구축된 데이터이지만, SNLI, MNLI와 같은 기존 NLI 데이터와 비슷한 방법론을 통해 구축되었다는(KLUE-NLI, Park et al., 2021) 한계가 있다.

이러한 상황 속에서 발전하고 있는 한국어 인공지능 언어 모델의 자연어 추론 능력을 평가하고 그 성능을 높일 수 있도록 학습시키기 위한 한국어 함의 분석 데이터 세트가 필요하며, 이를 위해 '적대적' 방식을 사용한 함의 분석 말뭉치를 구축할 필요성이 크게 대두되고 있다.

이때 ‘적대적’ 방식이라 함은 언어 모델이 쉽게 결과를 추론해 내지 못하도록 의도적으로 추론 관계를 설계한 방식을 뜻한다. 즉 데이터 세트를 구축하는 데 있어서 인간과 언어 모델이 다르게 추론할 수 있는 방식으로 데이터 세트를 구성함으로 궁극적으로는 언어 모델의 추론 능력을 인간과 유사하거나 높은 수준으로 훈련시키고자 데이터를 구축하는 방식을 ‘적대적’ 방식이라 하는 것이다.

적대적 함의 방식을 적용하여 구축된 데이터 세트의 필요성을 정리하면 다음과 같다.

- SuperGLUE의 9가지 평가 과제에서 4개 과제(MNLI, QNLI, RTE, WNLI)가 자연어추론 벤치마크임
- 높은 자연어추론 성능은 구체적인 하위과제인 질의응답, 관계추출에서의 높은 성능과 긍정적인 상관관계를 가지고 있음
- 주석자가 어려운 가설을 생성하여 인공지능의 잘못된 예측을 유도하는 방법은 인간과 인공지능의 수행능력 차이(performance gap)가 좁혀지는 상황에서 매우 유용한 접근임
- 적대적 프로세스는 함의 분석과 논리적 구조를 달리하는 새로운 한국어 벤치마크에도 적용이 가능함
- 초거대 인공지능이 공개된 바 있으나, 두 살 아이의 추론 능력보다 부족하다는 비판이 있었음
- 자연어추론은 보통 사람이라면 누구나 수용하지만, 인공지능은 쉽게 풀 수 없는 문제를 제시함으로써 인공지능의 향후 발전 방향을 구체적으로 제시함
- 현대 한국인의 언어 사용 양상을 수집한 말뭉치를 인공지능 산업에 활용가능한 디지털자원으로 가공한다는 점에서 의의가 큼. 영미권 말뭉치를 번역한 어

색한 한국어가 아닌 자연스러운 한국어를 학습한 인공지능 평가가 가능함

1.2. 사업의 목적

본 사업은 한국어 인공지능의 과업 수행 능력 일반과 높은 상관관계를 갖고 있는 언어 추론인 '함의 분석'을 위해 신문 텍스트로부터 선행 담화를 추출하고, 이를 기반으로 적대적 사례를 생성하여 인공지능 평가용 벤치마크로 가공함으로써 자연어 이해 벤치마크의 취약성 문제를 보완하기 위한 방안을 마련하는 데 목적이 있다. 특히 인공지능의 급속한 발전으로 인하여 인간과 인공지능의 과제 수행 능력 차이가 매우 좁혀지면서 자연어 이해를 위한 벤치마크의 짧은 수명(longevity)이 시급히 보완해야 할 문제로 부각되고 있다.

그동안 개발된 일부 벤치마크들을 살펴보자면, RTE는 고품질의 수작업 말뭉치이지만, 규모가 작아 기계학습에 적합하지 않고, 전제-가설에 대한 평정이 체계적이지 못한다는 한계가 있다. SNLI는 데이터 규모를 100배 이상으로 확장하고, 2,500명의 크라우드 워커를 활용하여 구축되었는데 전제-가설 1건당 5인의 주석 평정을 수행하고, 이후 다수의 의견을 따름으로써 경험적으로 타당한(empirically valid) 직관을 라벨로 부착하기도 하였다. MNLI는 단일 장르(이미지 캡션)가 아닌, 구어 및 문어 10종의 다양한 장르에서 전제를 추출하여 자연어추론의 도메인 전이(domain-transfer)를 강화하였다. SCITAIL은 주석자가 인위적인 가설을 생성하는 것이 아닌, 질의응답 말뭉치를 전제-가설로 변환하는 방법론을 제시하기도 하였는데 단 자연스러운 담화에서는 모순 관계가 매우 희소함을 보여주었다. ANLI는 모델이 예측을 틀릴 때까지 주석자가 문제를 반복적으로 출제하는 적대적(adversarial) 접근법을 취하고 있다는 점에서 본 과업과 추구하는 방향이 일치한

다.

이와 같은 자연어추론 벤치마크의 주요 개발 내용 특히, ANLI를 중심으로 한 함의 분석을 자연어이해 벤치마크의 견고성(robustness) 강화의 주요 축으로 발전시키는 것이 본 사업의 목적이라고 할 수 있다.

1.3. 사업 결과의 활용과 의의

적대적 사례 기반으로 구축된 함의 분석 말뭉치는 우선 한국어 인공지능 모델의 성능 향상에 기여할 수 있다. 앞에서도 지적한 바와 같이 적대적 사례에 기반하지 않은 기존 벤치마크 데이터를 활용한 모델들은 견고성이 떨어진다(Szegedy et al., 2014). 반면 적대적 사례를 기반으로 한 벤치마크 데이터들은 모델의 견고성을 높이고 과업 수행의 측면에서도 높은 성능을 보여주는 것으로 알려져 있다. 이와 같이 '적대성'을 적극 활용할 수 있는 기반을 마련함으로써 한국어 인공지능 모델의 성능 향상에 크게 기여할 수 있다.

또한 적대적 사례 기반의 데이터는 비교적 장기간 '벤치마크'로서 기능할 수 있다. 언어 모델의 발전 속도가 빨라지면서, 벤치마크의 개발 속도가 이를 따라잡지 못하고 있음이 지적되기도 하였다(Nie et al., 2020). 결과적으로 인공지능 모델의 약점을 공략하는 적대적 사례들을 통해 모델들이 벤치마크 데이터 세트의 통계적 편향을 부적절하게 학습하거나 활용하는 것을 방지하고, 이를 통해 데이터 세트 자체가 평가를 위한 데이터 세트로서 더 오래 유지될 수 있다.

한편 한국어를 대상으로 한 적대적 사례 기반의 함의 분석 말뭉치를 통해 얻을 수 있는 효과를 좀 더 구체적으로 논의하면 다음과 같다.

우선 적대적 사례 학습을 통한 자연어처리 모델의 정확도를 향상할 수 있다. 기

존의 연구 결과인 WordCNN, WordLSTM, BERT와 같은 여러 종류의 자연어처리 모델에서도 모두 적대적 사례를 학습 데이터로 사용한 후에 정확성이 높아졌음이 보고된 바 있다(Jin et al., 2020).

또한 한국어 적대적 사례 함의 분석에 대한 언어 모델들의 성능 평가가 가능해진다. 지금까지는 적대적 NLI를 한국어에 적용한 연구가 거의 불가능했으며, 김민호 외(2020)의 연구는 한국어를 포함한 다국어로 이루어진 적대적 말뭉치 'PAWS-X'를 활용한 것이었다. 그러나 해당 연구는 '다시쓰기'(paraphrasing) 탐지에 초점을 두고 있을 뿐 이를 NLI 연구로 보기에는 어려움이 있으며, 해당 말뭉치 또한 대응되는 외국어 자료의 번역을 통해 구축되었다는 한계가 있다.

따라서 본 사업을 통해 한국어로 구축된 최초의 적대적 함의 분석 말뭉치를 확보하게 된다. 이 말뭉치는 한국어 적대적 함의 분석 말뭉치의 표준으로서의 지위를 가지게 되며, 한국어의 적대적 말뭉치 구축에 있어 선도적인 역할을 수행함으로써 후속 연구들을 촉발하게 될 것이다.

2. 사업의 범위와 진행 과정

2.1. 사업의 범위

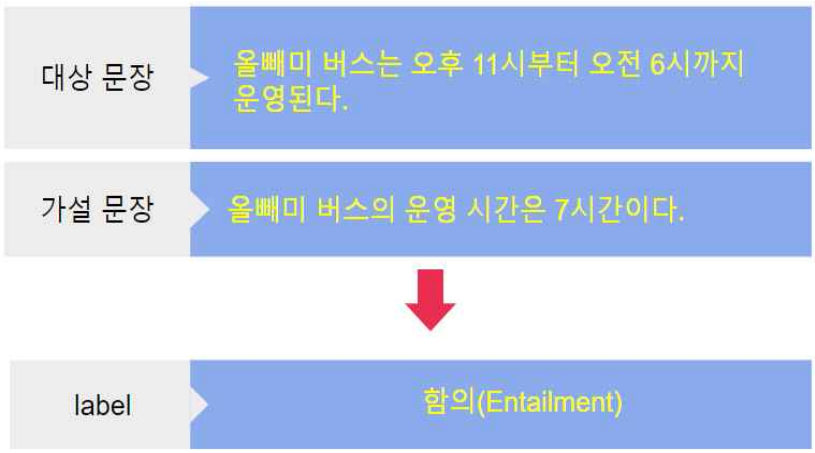
일반적으로 함의 분석(Recognizing Textual Entailment/RTE)이란 주어진 한 쌍의 문장(선행 담화와 가설 문장)에 대하여 두 문장 간의 관계를 추론하는 과제(Fyodorov et al. 2000; Condoravdi et al. 2003; Bos and Markert 2005; Dagan et al. 2005; MacCartney and Manning 2009)를 말한다. 이번 과업에서

는 주어진 선행 담화와 가설 문장의 쌍에 대하여 두 문장 간의 관계가 함의(entailment), 중립(neutral), 모순(contradiction)의 세 가지 라벨(label) 중 어떤 경우에 해당하는지를 추론하여 적합한 라벨을 부착하는 것이 핵심 내용이 된다. 이때 함의, 중립, 모순은 <표 1>과 같이 정의된다.

라벨(label)	정의
함의 (entailment)	전제 혹은 문맥이 되는 대상 문장이 가설 문장을 함의하는 관계 (definitely correct)
중립 (neutral)	전제 혹은 문맥이 되는 대상 문장의 정보만으로는 가설 문장이 대상 문장에 대하여 함의인지 모순인지 확정할 수 없는 관계 (neither definitely correct nor definitely incorrect)
모순 (contradiction)	전제 혹은 문맥이 되는 대상 문장이 가설 문장과 모순되는 관계 (definitely incorrect)

<표 1> 함의/모순/중립 라벨의 정의

즉 본 과업은 전제 혹은 문맥으로 기능하는 선행 담화와 가설 문장의 쌍이 주어질 때 두 문장 간의 관계를 추론해서 함의/중립/모순 중 적절한 라벨을 하나 선택하여 부여하는 것으로 다음의 <그림 1>은 그 예시이다.



<그림 1> 대상 문장과 가설 문장 간의 관계 추론 예시

이러한 내용을 기본적인 전제로 하여, 2022년도 과업에서는 다음과 같은 내용까지 포함하는 것으로 과업의 범위가 설정되었다.

[1] 적대적 함의 분석 말뭉치 구축 방법론 및 지침 수립

[2] 적대적 가설 문장 생성

- 적대적 가설 문장 생성을 위한 선행 담화 추출
  - 선행 담화는 국립국어원 신문 말뭉치 2021에서 추출함
- 모델 속이기(fooling)를 위하여 3만 건의 가설 문장 생성
  - 최종 결과물은 모델 속이기에 성공한 총 2만 건의 적대적 가설 문장으로 구성된 말뭉치
  - 모델 속이기에 실패한 중간 산출물도 포함해서 제출

[3] 모델 검증

- 언어 모델을 사용하여 적대적 가설 말뭉치의 강건성 검증

한편 적대적 가설을 생성하기 위한 선행 담화를 추출할 국립국어원 신문 말뭉치 2021의 구성은 <그림 2>와 같다.

매체 종류	매체 이름
전국종합일간 (5개)	국민일보, 서울신문, 조선일보, 한겨레, 한국일보
지역종합일간 (16개)	강원도민일보, 경기일보, 경남도민일보, 경북일보, 남도일보, 대구신문, 대전일보, 매일신문, 부산일보, 인천일보, 전남일보, 전북도민일보, 중도일보, 중부일보, 충청일보, 충청투데이
경제일간 (8개)	e대한경제, 머니투데이, 서울경제, 아시아경제, 아주경제, 파이낸셜뉴스, 한국경제, 헤럴드경제
스포츠일간 (1개)	스포츠서울
전문일간 (2개)	전자신문, 환경일보
인터넷신문 (3개)	EBN산업뉴스, 노컷뉴스, 뉴스핌

<그림 2> 국립국어원 신문 말뭉치 2021의 구성

2.2. 사업의 진행 과정

적대적 함의 분석의 하향식(Top-down) 생성 공정은 크게 (1) 선행 담화 추출 단계, (2) 적대적 사례 생성 단계, (3) 작업자간 검수 단계, (4) 모델 속이기(fooling) 단계의 네 단계로 구성된다. 이 과정을 도식화하면 <표 2>와 같다.

(1) 선행 담화 추출	(2) 적대적 가설 생성	(3) 작업자 간 검수	(4) 모델 fooling
신문 말뭉치 2021에서 선행 문장 + 대상 문장 추출	대상 문장을 바탕으로 한 적대적 사례 생성 및 라벨 부착	적대적 사례의 라벨에 대해 원 작업자 1인과 타 작업자 2인의 의견이 모두 일치하는 경우를 수합	작업자 3인의 의견이 모두 일치하는 사례를 모델에 입력하여 fooling이 성공하는 사례만을 수합

<표 2> 적대적 함의 분석 말뭉치 구축 과정

3. 관련 연구 동향

영미권에서 자연어추론 벤치마크는 인공지능이 구체적인 과업 일반의 수행 능력과 상관관계가 높은 추상적인 수준의 논리적 관계를 이해하고 있는지 평가하기 위하여 개발되었다. 이러한 자연어추론 과제는 데이터의 규모화, 장르의 다양화, 주석 공정의 정교화를 주요 축으로 하여 발전하였는데 이를 정리하면 <표 3>과 같다.

벤치마크	데이터 규모	전제 문장 장르	가설 문장 주석 공정
텍스트 함의 인식 (RTE-1)	1,337 건 (dev + test)	주석 말뭉치 (질의 응답) 웹 자료 (뉴스 기사)	전제 문장에 대하여 참과 거짓 가설(True/False hypothesis)을 생성함
스탠포드 자연어추론 (SNLI)	570,152 건 (train + dev +test)	이미지 캡션 (Flickr 30k)	주석 작업자는 3개의 전제와 함의(entailment), 모순(contradiction), 중립(neutral) 가설을 생성하고, 5인의 평정 점수를 수집하여 라벨로 가공함
다종장르 자연어추론 (MNLI)	433,000 건 (train + dev + test)	10종의 구어/문어 말뭉치	단일 장르인 SNLI와 달리, 음성 전사, 1:1대화, 연설문, 편지, 유명 소설에서 전제를 추출함
과학 질의응답 자연어추론 (SCITAIL)	27,000 건	중고등학교 과학 분야 문제집	질문과 4지 선다형 객관식 문제를 함께 1건으로 하여, 질문을 가설 문장으로 간주하고 선지를 10개의 전제 문장으로 변형함
적대적 자연어추론 (ANLI)	169,000 건 (train + dev + test)	기존 말뭉치/벤치마크	주석자를 '화이트해커'로 간주하고 모델이 취약한 적대적 자연어 추론 사례를 만드는 역할을 부여함

<표 3> 영미권 자연어추론 벤치마크의 사례

여기서 RTE는 참과 거짓 가설의 분포를 50% 대 50%로 통제하고, 논리적 관계에 대한 주석자의 판단이 일치하지 않는 전제-가설 (20%)을 제외하였으며, SNLI는 전제와 달리 가설의 길이가 대부분 7단어 길이였다. 또한 5인의 평정자 중에서 3인

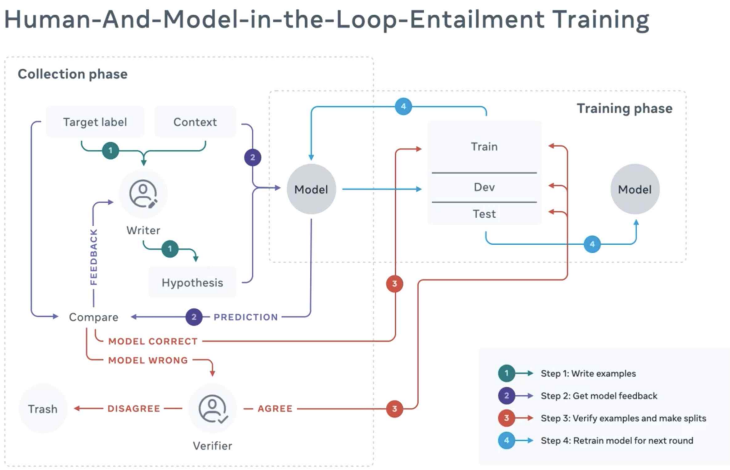
이상의 직관이 합치하는 라벨이 포함된 전제-가설을 최종 평가세트로 공개하고 있다. 이 SNLI는 2% 미만의 전제-가설에서 직관 불일치 문제가 발생하였다. SCITAIL은 5명의 평정자가 전제를 읽고 전제가 가설을 뒷받침하는지를 기준으로 삼분지 주석 정보를 부착하였다. 이때 전제가 가설과 상충하는 모순은 매우 희소하여 제외되었다. 마지막으로 ANLI는 Round 1~3에서 주석자가 가설을 생성하고 BERT, RoBERTa모형이 정답 라벨을 반복적으로 예측하도록 수행되었다. 이 ANLI는 신경망 모형이 예측을 틀리면 주석자의 임무는 완수되고(일종의 모델 속이기 실험), Round를 거듭할수록 더 다양한 장르의 말뭉치에서 보다 적대적인 사례를 제작하였다.

이 밖에 함의 분석을 활용한 자연어추론(NLI) 영역에서의 대표적인 적대적 사례 연구로 ANLI & HAMLET(Nie et al., 2020)을 들 수 있다. 이 연구에서는 기존 NLI 데이터 세트의 문제점을 지적하였는데, 즉 인공지능 모델의 발전 속도를 기존의 자연어이해 벤치마크가 따라가지 못하고 있다는 것이다. 기존의 NLU 벤치마크로는 보통 SNLI, SQuAD, SentEval, GLUE 등을 들 수 있는데 이들이 인간의 이해 능력에 도달하는 데에는 길게는 15년(MNIST), 짧게는 7년(ImageNet) 정도의 기간이 소요되었다. 그 후 2018년 BERT의 등장 이후에는, 이러한 AI 모델들이 GLUE와 같은 벤치마크에서 인간 수준의 이해도에 도달하는 속도가 매우 빨라지게 되었고 이러한 이유로 SuperGLUE라고 하는 새로운 벤치마크가 등장하게 되었다.

그러나 과연 현재의 최신(state-of-the-art) 모델들이 실제로 벤치마크가 평가를 받는 것만큼의 좋은 성능을 내고 있는지는 재고의 여지가 있다. 즉 기존의 벤치마크에서 높은 점수를 받고 있는 모델들이 인간이 의미를 이해하는 것처럼 어떤 일반화가 가능한 방식으로 의미를 이해하는 것이 아니라, 벤치마크 데이터 세트의 의사적인 통계 패턴만을 활용해서 높은 점수를 받고 있다는 의문이 타당성 있게 제기되

었기 때문이다.

ANLI 데이터 세트와 HAMLET의 핵심은 난이도가 높고, 오랜 기간 동안 인공지능 성능 평가에 유효한 벤치마크를 구축하는 것이다. 여기서는 적대적인 데이터 세트를 구축하기 위한 구체적인 방법론을 제시하고 있을 뿐 아니라 특정 데이터 세트에만 유용하게 맞추어진 통계 패턴을 악용하는 것을 방지하고, 실제 인공지능 언어 모델의 언어 추론 능력을 벤치마크를 통해 확인하고, 인공지능 모델 성능을 전반적으로 향상시킬 수 있도록 노력하였다. 참고로 HAMLET의 작업 공정은 <그림 3>과 같다(Nie et al., 2020).



<그림 3> HAMLET 작업 공정

또한 적대적 데이터를 통해 훈련을 시킬수록 모델의 견고성이 향상된다는 점을 보여주었으며, ANLI를 통해 학습시킨 모델이 논문 출판 당시를 기준으로 가장 좋은 성능을 보였다. 이 연구에서 실험한 ANLI 데이터 세트를 활용한 인공지능 언어 모델의 평가 결과는 <그림 4>와 같다(Nie et al., 2020).

Model	Training Data	A1	A2	A3	ANLI	ANLI-E	SNLI	MNLI-m/-mm
BERT	S,M*1	00.0	28.9	28.8	19.8	19.9	91.3	86.7 / 86.4
	+A1	44.2	32.6	29.3	35.0	34.2	91.3	86.3 / 86.5
	+A1+A2	57.3	45.2	33.4	44.6	43.2	90.9	86.3 / 86.3
	+A1+A2+A3	57.2	49.0	46.1	50.5	46.3	90.9	85.6 / 85.4
	S,M,F,ANLI	57.4	48.3	43.5	49.3	44.2	90.4	86.0 / 85.8
XLNet	S,M,F,ANLI	67.6	50.7	48.3	55.1	52.0	91.8	89.6 / 89.4
RoBERTa	S,M	47.6	25.4	22.1	31.1	31.4	92.6	90.8 / 90.6
	+F	54.0	24.2	22.4	32.8	33.7	92.7	90.6 / 90.5
	+F+A1*2	68.7	19.3	22.0	35.8	36.8	92.8	90.9 / 90.7
	+F+A1+A2*3	71.2	44.3	20.4	43.7	41.4	92.9	91.0 / 90.7
	S,M,F,ANLI	73.8	48.9	44.4	53.7	49.7	92.6	91.0 / 90.6

<그림 4> ANLI 데이터 세트를 통한 인공지능 언어 모델 평가 결과

마지막으로 참고할 만한 연구로 Jin et al.(2020)을 들 수 있다. 이 연구에서도 적대적 사례 데이터를 자연어처리(NLP) 모델에 학습하여 모델의 성능 개선을 시도하였다. 이때 적대적 사례 데이터를 학습할 자연어처리 모델로는 WordCNN, WordLSTM, standard InferSent, ESIM, 그리고 BERT(Bidirectional Encoder Representations from Transformers) 등이 포함되었다. BERT를 비롯한 여러 자연어처리 모델의 텍스트 분류(classification)와 함의 분석 과제(RTE) 수행에 있어서 여러 가지 데이터 세트를 활용하여 적대적 사례를 학습시켰는데, 함의 분석을 수행하는 과정에서 적대적 사례를 활용하였을 때 standard InferSent, ESIM, BERT 세 모델의 정확도(accuracy)가 크게 떨어진 것으로 나타났다. 또한 소스 데이터만을 훈련시켰을 때보다 적대적 사례를 훈련시켰을 때 MR과 SNLI 두 과제의 수행에 있어 BERT 모델의 견고성(robustness)이 향상되는 것으로 드러났다. 이 연구에서 다루고 있는 데이터 세트는 <표 4>와 같다.

과업(Task)	데이터 세트 Dataset	훈련 규모 Train	테스트 규모 Test	평균길이 Average Length
분류 Classification	AG's News	30K	1.9K	43
	Fake News	18.8K	2K	885
	MR	9K	1K	20
	IMDB	25K	25K	215
	Yelp	560K	38K	152
함의 Entailment	SNLI	570K	3K	8
	MultiNLI	433K	10K	11

<표 4> 데이터 세트의 개요

또한 <표 5>는 적대적 사례와 비교 결과를 정리한 것이고, <표 6>은 적대적 사례를 학습한 결과와 그렇지 않은 경우 BERT의 성능 차이를 나타낸 것이다.

	InferSent 언어모델		ESIM 언어모델		BERT 언어모델	
	SNLI	Multi NLI (m/mm)	SNLI	Multi NLI (m/mm)	SNLI	Multi NLI (m/mm)
최초 정확도 Original Accuracy	84.3.	70.9/ 69.6	86.5	77.6/75.8	89.4	85.1/ 82.1
적대적 이후 정확도 After-Attack Accuracy	3.5	6.7/6.9	5.1	7.7/7.3	4.0	9.6/8.3
% Perturbed Words	18.0	13.8/14.6	18.1	14.5/14.6	18.5	15.2/14.6
의미적 유사성 Semantic Similarity	0.50	0.61/0.59	0.47	0.59/0.59	0.45	0.57/0.58
질문 수 Query Number	57	70/83	58	72/87	60	78/86
평균 문자 길이 Average Text Length	8	11/12	8	11/12	8	11/12

<표 5> 적대적 사례의 비교 결과

	MR 언어모델		SNLI 언어모델	
	After Accuracy	Perturbed Words	After Accuracy	Perturbed Words
Original	11.5	16.7	4.0	18.5
+ Adv. Training	18.7	21.0	8.3	20.1

<표 6> 적대적 사례를 학습한 경우와 그렇지 않은 경우 BERT의 성능 비교

요컨대 AI의 발달은 자연어처리 분야에서 대규모 자연어 벤치마크의 발달에 의해 촉진되어 왔으나 AI의 급속한 발달 속도로 인해, 자연어이해 벤치마크들은 앞서 나가는 모델의 성능 향상을 따라잡기 어려워지는 상황이 되었다. 이러한 상황은 장기간 사용 가능한 대규모의 벤치마크 데이터 세트를 수집할 수 있을지에 대한 의문으로 이어졌으며, 현재의 자연어처리 모델이 실제 벤치마크에서 보이는 성능만큼 실세계에서 좋은 성능을 보일 수 있는지에 대한 의문도 지속되는 상황이다. 또한 현재 존재하는 최신의 모델(state-of-the-art)들은 인간이 하는 것처럼 유연하고 일반화된 의미 이해에 입각하여 문제를 해결하는 것이 아니라, 표면적인 통계적 패턴들로만 문제를 해결한다는 증거들이 다수 발견되고 있다. 이러한 시각에 입각해서 Nie et al.(2020)은 벤치마크의 지속 가능성과 강건성을 위한 새로운 자연어추론 데이터 세트를 제시한 것으로 평가되며, 해당 데이터 세트는 인간 주석자를 활용하여 적대적 사례를 구축하는 것에 초점이 맞춰져 있다. 즉 작업자로 하여금 문맥에 맞는 가설을 생성하도록 하고, 라벨과 다른 대답을 내놓도록 모델을 속임(fooling)으로써 새로운 데이터 세트를 만드는 것이 주요한 목적이다. 이를 통해 모델의 취약성을 꾸준히 발견하고, 지속적으로 훈련할 수 있게 되었다. 한편 모델을 속이기 위한 의미화용적 추론 관계로는 수와 양(numerical&quant), 지시와 명칭(reference&names), 표준(standard), 어휘(lexical), 속임수(tricky), 추론 및 사실(reasoning&fact)을 사용하였는데, 이러한 방식은 본 과제에서도 적극 반영되었다.

제 2 장

선행 담화의 추출

1. 선행 담화 추출 대상

적대적 가설을 생성하기 위한 전제가 되는 '선행 담화'는 국립국어원의 신문 말뭉치 2021에서 추출하였다. 이 신문 말뭉치는 모두 729,280건의 기사를 포함하고 있으며 총 2.95GB 분량의 데이터이다. 신문 말뭉치의 구조는 <그림 5>와 같다.

<그림 5> 국립국어원 신문말뭉치 2021의 JSON 구조

선행 담화의 추출 대상이 신문이라는 단일한 장르로 제한되어 있다는 점은 올해 함의 분석의 중요한 특징 중 하나이다. 먼저 신문기사는 문형과 어휘가 비교적 정

형화되어 있다는 점을 고려할 필요가 있다. 따라서 선행 담화 추출을 균형 있고 분산되게 수행해야 하며, 내용적으로는 특정 집단에 대한 편견, 윤리적으로 부적절한 표현 등이 나타나지 않아야 한다.

국립국어원의 신문 말뭉치 형식인 JSON은 데이터 내용과 구조를 함께 저장하는 파일 형식으로, Python 등의 프로그래밍 언어로 해당 파일로부터 데이터를 구조까지 모두 불러오기에 용이하게 되어 있다. 특히 신문 말뭉치의 JSON은 파일에 대한 메타데이터와 기사에 대한 메타데이터, 그리고 기사를 구성하는 문단들로 구성되어 있으며, 각 문단과 기사, 문서 등에 id가 부여되어 있어 선행 담화를 추출하는 과정에서 이들을 적극 참조하였다.

2. 선행 담화 추출 기준

신문 말뭉치로부터 선행 담화를 추출하기 위해서는 몇 가지 기준과 원칙이 필요하다. 앞에서도 지적한 바와 같이 신문이라는 동일한 장르 내에서 가급적 다양한 문형과 어휘가 포함된 선행 담화를 추출하기 위해서는 균형 잡힌 텍스트 선별이 수행되어야 하기 때문이다.

먼저 선행 담화 추출은 자동 임의 추출을 원칙으로 하였다.

이미 전산적으로 처리하기 쉬운 형태로 제공된 신문 말뭉치의 장점을 활용해 담화 추출의 전 과정을 스크립트를 통해 자동 추출하며, 다양한 영역의 기사 문장을 살펴보면서 작업할 수 있도록 임의로 추출하였다.

다음으로 “paragraph”>“form”을 중심으로 담화를 추출하였다. 궁극적으로 작업자는 제시된 선행 담화를 보고 적대적 가설 문장을 생성하여야 하기 때문에 이 문장 정보를 가진 paragraph 항목의 form을 중심으로 하여 최소한의 내용을 추출하

였다.

세 번째, id를 포함하여 추출해서 추적이 가능하도록 하였다. id 정보는 작업에 직접적인 정보를 제공하지 않지만, 중복 추출 방지 및 추후 오류 추적 등의 이유로 그 효용이 인정되어, 문단 id 정보를 함께 추출해 내부적으로 저장하였다.

마지막으로 메타데이터(metadata)를 포함하여 추출함으로써 다양한 활용이 가능하게 하였다. 언론사 정보 등을 담은 메타데이터는 id와 마찬가지로 작업에 직접적인 정보를 제공하지 않으나, 이를 추출해 내부적으로 저장해 두어 차후 메타데이터 내용을 기반으로 한 성과물 활용(주제에 따른 가설 문장 필터링 등)을 도모할 수 있다.

3. 선행 담화 추출 과정 및 결과

이제 구체적인 담화 추출 과정과 결과를 살펴보기로 한다. 자료 추출 과정의 개략적인 순서는 아래와 같다 :

[1] JSON으로 공유된 말뭉치 데이터 파싱 및 통합
[2] 각 말뭉치 파일에서 ‘document’의 항목 추출
[3] ‘document>paragraph’ 중에서 임의 추출된 대상에 대해,
  [3-1] 같은 ‘document’ 안에서, 추출 대상에 선행하는 ‘document>paragraph’가 있는지 확인
  [3-2] 선행 ‘document>paragraph’가 있을 경우, ‘document>paragraph>id’가 기추출 id 목록에 포함되었는지 확인
  [3-3] 미포함된 경우 ‘document>paragraph>id’를 기추출 id 목록에 추가

[3-4] 대상 ‘document>paragraph>form’을 동일한 ‘document>paragraph’ 안에서 선행하는 항목과 함께 ‘document>metadata>topic’과 ‘document>paragraph>form’을 추출된 담화 목록에 추가

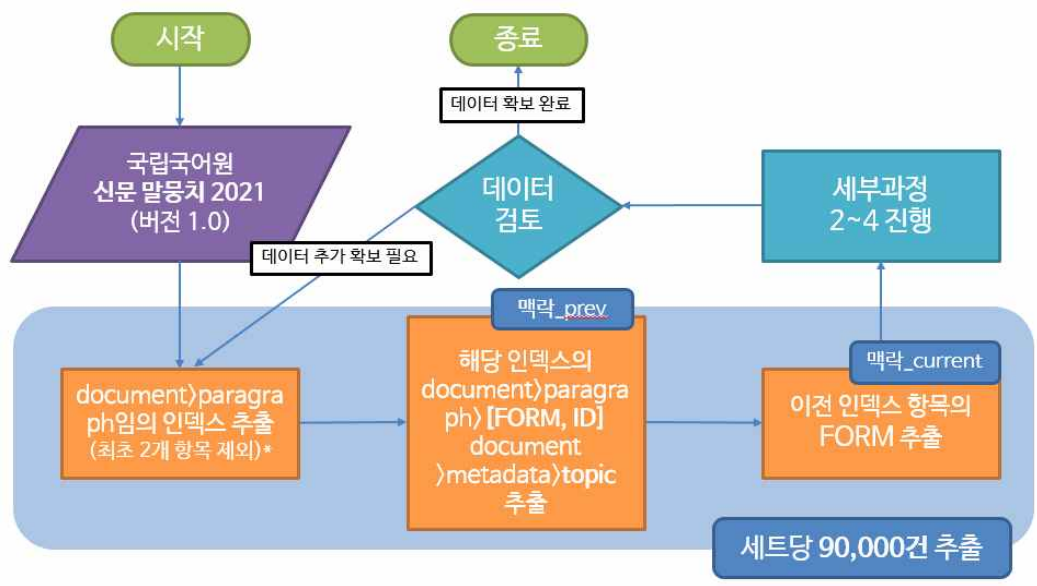
[4] 추출된 담화 목록의 ‘document>paragraph>form’ 내용을 작업자가 읽기 쉬운 형태로 통합

[5] 통합된 내용을 ‘document>metatdata>form’ 정보와 함께 웹상의 MySQL 데이터베이스에 업로드

[6] 업로드 된 대상에 대해 웹 기반 워크벤치를 통해 함의 사례 생성 작업 진행

다만, 기존 데이터의 ‘document>paragraph’는 문서 전반에 걸쳐 각 항목의 단위 및 길이에 차이가 있었다. 해당 항목의 상당수가 하나의 문장으로 볼 수 있는 단위였기에 해당 항목 두 개를 불러옴으로써 두 문장 정도 길이의 데이터를 구축할 수 있었으나, 길이가 너무 길어지는 것을 방지하기 위하여 임의 추출된 대상과 선행하는 대상에 대해 길이와 문장 종결 부호 등을 활용하여 휴리스틱하게 작업 문장 단위를 제어하였고, 이후 작업자에게 작업 문장이 올바르지 않게 주어진 경우 작업하지 않도록 지시하였다. 이외에도 ‘document>paragraph’에 기사의 제목 및 헤드라인 등 연결되는 문장이라고 보기 어려운 사례가 있는 경우가 있었기에, 주로 제목으로 구성된 ‘document>paragraph’의 첫 항목이 실제 문장에 포함되지 않도록 조절하였고, 결과에 대해서도 위와 같은 휴리스틱을 통해 두 개의 연속된 문장으로 구축되도록 유도하였다.

이러한 과정을 도식화하면 <그림 6>과 같다.



<그림 6> 선행 담화의 추출 과정

이러한 과정을 거쳐 모두 7개의 라운드를 수행하여 630,000건의 선행 담화 후보를 성공적으로 추출하였다(라운드별 90,000건). 추출한 결과물의 샘플은 <그림 7>을 참조할 수 있다.

NPRW2100000005.49656.2	정치	도널드 트럼프 미국 대통령이 4일(현지시간) 새벽 백악관에서 긴급히 가진 연설을 통해 "사실상 우리가 승리를 거뒀다"며 승리를 확신했다.
NLRW2100000005.2494.8	사회	민갑룡 경찰청장이 'n번방' 사건을 계기로 디지털 성범죄 영상 제작자와 조력자, 가담자 전원에 대한 처벌 의지를 밝힌 가운데 디지털 성범죄 특별
NPRW2100000003.27167.2	경제	문재인 대통령이 역점을 두고 추진 중인 '한국판 뉴딜' 정책에 금융권도 협력해달라는 정부의 요청에 신한금융그룹도 팔을 걷어붙였다.
NWRW2100000003.8728.5	사회	본지 취재를 종합하면, 대검은 13일 A4 용지 10여쪽 분량의 의견서에서 추미애 법무부 장관이 밀어붙인 이번 직제개편안을 정면 반박한 것으로 알려졌
NLRW2100000008.13181.6	정치	민주당은 국민의힘 불참 속에 지난 24일 열린 법사위 법안소위에서 중대재해법안에 대한 일부 수정이 불가피하다는 데 의견을 모았고, 정부가 이를 반
NPRW2100000005.59891.3	사회	싱가포르의 첫 접종 대상자로 선정된 사라 림(46) 국립전염병 센터 수석간호사는 이날 화이자-바이오엔테크의 코로나19 백신을 30여명의 센터직원들과
NPRW2100000009.41360.4	경제	재임 기간에 운용 성과가 나쁘지 않다는 것이 가장 큰 이유다. 기금 운용과 관련해 특별한 실책이 없는 CIO를 신종 코로나바
NPRW2100000007.1871.6	생활	'루이 책상/책장'은 셀프 인테리어 트렌드에 맞는 깔끔하고 세련된 디자인이 특징이다. 책상과 탈부착이 가능한 스크린보드를 세트로 구성해 큰 공
NPRW2100000002.5050.20	생활	이에따라 조 회장 측이 앞으로 경영권 분쟁에서 경영권을 계속 행사해야 한다는 명분 측면에서 우위를 점할 것이라는 관측이다.
NPRW2100000002.2257.7	정치	정 총리는 공직자들이 일하는 방식을 바꿔야 한다고 주문했다. 정 총리는 "(공직자들이) 이일을 어떻게 'NO'라고 할까를
NPRW2100000009.32619.3	연예	올 하반기 첫 방송 예정인 tvN 새 드라마 '여신강림'은 외모 콤플렉스를 가지고 있다가 '화장'을 통해 여신이 된 주경과 남모를 상처를 간직한 수
NWRW2100000002.30643.7	경제	공급은 한국토지주택공사(LH)나 서울주택도시공사(SH) 등 공적 기관이 주도한다. 김현미 장관은 9일 국회 예결위원회에서 "전세 임대는 이미
NPRW2100000009.23645.8	사회	한편, 에르난데스 대통령은 지난 12일 문재인 대통령과의 전화통화에서 코로나19 및 포스트 코로나 시대에서 양국의 협력 방안 등에 논의한 바 있다.
NPRW2100000007.3933.5	사회	알마덴디자인리서치 컨설팅 역량과 경기혁신센터 창업지원 인프라를 결합, 아이디어 개발, 비즈니스 모델 수립, 린 스타트업으로 이어지는 합작 모델
NLRW2100000008.126.10	생활	특히 자율차의 경우 미래형자율주행차 연구개발(R&D) 기반조성 및 부품산업 육성, 자율차 시범운행지구 지정, 자율차 부품평가 인증센터 구축 등 3개
NLRW2100000005.3636.7	미용/건강	특히 탈부착이 가능한 체인 핸들이 함께 구성돼 룩의 완성도를 높여주는 포인트 아이템이다. 캔디 케이스에서 영감을 받아
NLRW2100000016.844.10	사회	가족과 동거인이 있는 경우 더욱 신경 써야한다. 대화 등 접촉을 하지 않는 것이 가장 좋으나 불가피하게 접촉해야 한다면 얼굴을 맞대지
NPRW2100000010.32418.6	사회	이는 지난 8월 '8·15 집회' 때처럼 대규모로 참가자들이 모이는 집회는 하지 않겠다는 이야기로, 차량 시위나 1인 시위는 당초 준비한 대로 진행할
NPRW2100000010.29993.4	정치	당정이 결정한 2차 지원금의 지급 방식에 사실상 반기를 든 셈이다. 이 지사가 보편 복지를 늘 주장한 것을 감안하더라도 당의
NLRW2100000010.1219.5	사회	광명시보건소는 광명시 18개 동 전역을 3개 동씩 묶어 6개 지역으로 나누고 전문 방역업체를 통해 방역을 하고 있다. 전문 방역업체는 휴대용 분무
NPRW2100000004.6131.3	생활	현대자동차는 25일(수)부터 전국 영업점을 통해 사전계약에 들어간 '올 뉴 아반떼'의 첫 날 계약대수가1만58대를 기록했다고 밝혔다.
NWRW2100000001.5526.8	미용/건강	드라이브 스루 선별진료소를 이용하면 문진표 작성부터 의사진료, 검체 채취까지 모든 것이 차량 안에서 이루어지기 때문에 검사시간이 1인당 종전 1

<그림 7> 신문으로부터 추출한 선행 담화 예시

제 3 장

적대적 가설 문장의 생성

1. 적대적 가설 문장 생성을 위한 전략

선행 담화가 적절히 추출되었다면 그다음 단계는 적대적 가설 문장을 생성하는 단계이다. 적대적 가설 문장은 검수자 간의 검토를 거쳐 궁극적으로 언어 모델을 속이는 데(fooling) 성공해야 하기 때문에 가설 문장을 생성하는 단계는 작업자에게 많은 부담을 줄 수 있다. 따라서 적대적 가설 문장 생성을 효율적으로 진행할 수 있는 다양한 책략과 작업 환경을 마련해 주는 것이 매우 중요하다. 이 사업에서는 Nie et al.(2020)에서 제시한 적대적 가설 문장 생성을 위한 논리적 관계를 하향식으로 적용하여 가설 문장을 다양하게 생성할 수 있도록 작업자들을 교육하였으며, 동시에 작업자가 쉽게 사용할 수 있는 작업용 워크벤치를 개발하여 활용하였다.

(1) 수량 양화 표현 Numerical & Quant.

▸가설 도출 전략 1: 숫자와 공기하는 술어를 변환하여 숫자/날짜/연령의 해석을 활용하여 속이기(fooling)

[선행 담화] 넷플릭스가 19일(현지시간) 발표한 1분기 실적보고서에 따르면 지난 1~3월 넷플릭스 가입자는 20만명 감소했다
[가설] 지난 1~3월에 넷플릭스에 가입한 사람은 20만명이다.
[함의 관계] 모순(Contradict)
[추론] 함의(Entailment)

▸가설 도출 전략 2: 문맥에서 제시된 집합-부분집합-여집합 간의 관계를 활용하여 속이기

[선행 담화] 넷플릭스는 미국과 캐나다에서만 3천만 가구가 계정 공유를 통해 콘텐츠에 접근하는 등 전 세계적으로 1억이 넘는 가구가 다른 유료 회원의 계정을 공유하는 것으로 추산하고 있습니다.
[가설] 미국과 캐나다를 제외한 전 세계에서는 7천만 가구가 계정 공유를 통해 콘텐츠에 접근하고 있다.
[함의 관계] 함의(Entailment)
[추론] 모순(Contradict)

(2) 지시와 명칭 Reference & names

▸가설 도출 전략 3: 일반명사 혹은 고유명사와 상호참조하는 대명사로 수정하여 속이기

[선행 담화] 독자 여러분께서 언론윤리와 언론법제의 상호보충성과 상호독립성에 더 많은 관심이 있으시다면 저자가 이 책과 함께 짝을 이루는 저술로 박영사(博英社)에서 이미 '언론법제론'을 출간해, 전정판(제3판/2007)까지 펴냈으므로 그 책을 참고하여 읽어 주시기를 부탁드린다. 책의 부피를 조금이라도 줄이기 위해 참고문헌은 초판에서처럼 각주로 대체하였고 '찾아보기' 부분도 인명의 경우는 외국인도 '가나다'순으로 배열한 초판을 그대로 두었고 사항의 경우에 개정판 작업의 일환으로 일부만 고쳤다.
[가설] 그는 이 책과 함께 박영사의 '언론법제론'을 과거에 펴 냈다.
[함의 관계] 함의(Entailment)
[추론] 중립(Neutral)

(3) 표준적인 방법 Standard

▸가설 도출 전략 4: 인과적 논리 관계를 이용하여 속이기

[선행 담화] 택시기사들도 코로나 여파로 배달이나 택배 등으로 업종을 전환했다. 그 결과 택시 수가 대폭 감소했다.
[가설] 코로나 여파 이전에는 배달이나 택배 등으로의 업종 전환이 없었다.
[함의 관계] 중립(Neutral)
[추론] 함의(Entailment)

▸가설 도출 전략 5: 부정 관계를 fooling

[선행 담화] 구름카페 문학상 수상집을 엮으며, 나의 문학 현주소를 읽는다. 수상집에 실린 1부는 삶의 반경에 든 가슴을 울린 대상을 의미화한 글이다. 최근 수필전문지와 신문에 연재한 신작이다.
[가설] 구름카페 문학상 수상집은 아직 수필전문지나 신문에는 공개되지 않은 새로운 작품들을 엮은 것이다.
[함의 관계] 모순(Contradict)
[추론] 중립(Neutral)

▸가설 도출 전략 6: 비교 관계를 이용한 속이기

[선행 담화] 그래서 젊은이들은 더 나은 직장을 찾는다. 그러나 이것은 함께 극복해

야 하는 동시에 받아들여야 할 현실이다. 사회적으로 인정받는 좋은 직장만이 자존감을 높이는 길이 아니다.
[가설] 요즘 젊은이들은 자신의 직장을 잘 받아들이고 좋은 직장으로 생각한다.
[함의 관계] 모순(Contradict)
[추론] 중립(Neutral)

(4) 어휘적인(Lexical) 방법

▸가설 도출 전략 7: 동의어 혹은 반의어 관계 간의 어휘를 수정하여 fooling

예시 1.
[선행 담화] 월스트리트저널(WSJ)는 17일(현지시간) 지난달 미 남부 국경을 통해 입국을 시도한 이민자수가 22만 1303명이라고 보도했다. 이는 22년만에 경신된 최대치 기록이다.
[가설] 미 남부 국경을 통해 입국을 완료한 이민자수가 22만 1303명이다.
[함의 관계] 중립(Neutral)
[추론] 모순(Contradict)

예시 2.
[선행 담화] 차 지부장은 고려인들의 잃어버린 문화와 언어를 되찾아주는 것이 중요하다고 강조했다. 한국 사회에서 융화하며 살아가기 위해 언어와 문화는 가장 필요한 부분이라는 생각에서다. 그러기 위해서 고려인 문화원의 역할이 중요하지만, 현재 인천 고려인문화원은 별도의 공간 없이 교회 건물을 빌려 사용하고 있다

[가설] 고려인들이 한국 사회에서 함께 살아가기 위해서 우리 문화를 이해하고 한국어를 배우는 것은 매우 중요하다.
[함의 관계] 함의(Entailment)
[추론] 중립(Neutral)

▸가설 도출 전략 8: 복합 명사구 속 단어를 수정하여 속이기

[선행 담화] 하이트진로의 지난해 일본 소주 수출액은 증가했다. 조사 결과, 전년대비 약 27% 증가한 것으로 밝혀졌다.
[가설] 하이트진로의 지난해 일본 맥주 수출액은 전년대비 약 27% 감소했다.
[함의 관계] 중립(Neutral)
[추론] 모순(Contradict)

▸가설 도출 전략 9: 약어의 구성 요소를 분리하여 속이기

[선행 담화] 전문가들은 MZ세대(밀레니얼+Z세대)가 자유롭고 개방된 문화를 추구한다고 분석했다. 덧붙여, 이들에게 경직된 공무원 사회의 문화가 거부감을 키울 수 있다고 전했다.
[가설] Z 세대가 추구하는 문화는 밀레니얼 세대가 추구하는 문화와 그 성격이 다르다.
[함의 관계] 모순(Contradict)
[추론] 함의(Entailment)

(5) 속임수 Tricky

▸가설 도출 전략 10: 구문 변환이나 재정렬을 통한 속이기

[선행 담화] 경남도는 지난해 12월 3일부터 19일까지 화재안전등급이 낮은 38개 시장을 대상으로 소방서 등과 합동으로 화재대응 실태 점검을 벌인 바 있다. 도는 자동화재속보기 등 소방시설 작동유무·소방시설 주변 적치물 방치·소방통로 미확보 등을 중점 점검했다.
[가설] 경상남도는 지난해 연말, 화재안전등급을 기준으로 하여, 일부 시장의 화재대응 실태를 점검했다.
[함의 관계] 함의(Entailment)
[추론] 중립(Neutral)

(6) 추론과 사실 Reasoning & Facts

▸가설 도출 전략 11: 세상의 지식 및 상식(예: 한국의 이름 문화)를 사용하여 속이기

[선행 담화] 2020년 초는 코로나 유행이 본격화하기 직전이었다. 당시 직장인 임모 cm, 몸무게 64kg으로 정상 체중에 속했다.
[가설] 직장인 임모 씨의 친가는 임씨 집안이다.
[함의 관계] 함의(Entailment)
[추론] 중립(Neutral)

2. 적대적 가설 문장 생성 과정

앞 절에서 제안한 적대적 가설 문장을 생성하기 위한 전략은 Nie et al.(2020) ANLI 연구의 Bottom-up 형태를 변용한 것으로서, Top-down 형태를 따른 것이다.

이러한 거시적인 전략에 따라 다음과 같은 세부적인 절차를 거쳐서 적대적 가설 문장의 생성 과정이 진행된다.

(1) 6개의 추론 방식에 근거한 지침 수립(지침의 세부 내용은 <부록>을 참조)
(2) 작업자 실습 교육
(3) 지침에 따라 적대적 가설 문장 생성
(4) 작업자 간 검수
(5) 적대적 사례를 모델에 입력하여 속이기(fooling) 테스트(작업자가 부착한 라벨과 모델의 예측값이 불일치하는 경우 속이기에 성공한 것임)

본 작업에는 모두 6개조(3인씩 구성), 18명의 연구원이 투입되었다. 다수의 작업자가 참여할 뿐 아니라 실제 가설 생성과 관련한 실습이 필요하기 때문에 작업자 실습 교육을 온라인/오프라인으로 철저하게 수행하였다. 특히 애써 생성한 적대적 가설 문장이 모델 속이기에 실패할 경우가 지나치게 많이 나오지 않도록 다음과 같은 사항에 유의하도록 교육 과정에 제시하였다.

▸가설 문장 생성 시 언어 모델을 속이기 위해 유의할 점

(1) 인공지능 모델은 문장을 선형적으로 인식하지 않는다.

→ 즉 기존 정보를 단순히 순서만 바꾸어 뒤섞는 것만으로는 속이기(fooling)에 성공하지 못할 가능성이 크다는 점에 유의해야 한다.

(2) 인공지능 모델은 특정한 일반상식이 결여되어 있다.
→ 나이 계산법, 호칭법 등과 같이 세상과 관련한 일반상식을 포함하도록 가설을 생성할 경우 속이기에 성공할 가능성이 높아진다.

(3) 인공지능 모델은 한 문장 안에 두 군데 이상 변형을 가할 경우 명제 인식에 어려움을 겪는다.
→ 복합 명사구에 포함된 내부 단어를 바꾼다면 맥락 속 서술어 등 다른 단어도 함께 바꾸어야 속이기에 성공할 가능성이 커진다.

(4) 인공지능 모델은 생략된 명사구의 복원에 어려움을 겪을 수 있다.
→ 생략되어도 복원 가능한 명사구의 경우, 가설에서 이를 생략해서 제시하면 속이기에 성공할 가능성이 높아진다.

(5) 인공지능 모델은 은유나 환유를 이해하는 데에 어려움을 겪는다.
→ 맥락에서 명시적으로 제시된 내용을 가설에서 은유나 환유를 통해 제시하면 속이기에 성공할 가능성이 커진다.

지침 교육과 한편으로 다수의 작업자가 효율적으로 가설 문장을 생성하기 위해서는 공동 작업을 위한 워크벤치(작업 도구)가 필요하다. 본 사업에서는 사업 수행 초기부터 워크벤치 개발을 상당수 완료해 두었으며, 작업자의 검증과 피드백을 거쳐

실제 작업에 활용하기 위한 작업 도구를 개발, 활용하였다. 본 작업을 위해 활용한 작업용 워크벤치를 보이면 <그림 8>과 같다.

<그림 8> 작업자를 위한 함의분석 말뭉치 작업용 워크벤치

<그림 8>의 워크벤치는 작업의 효율을 위해 가설 단계에서'부적절'을 선택할 수 있게 하는 등 작업자의 편의를 최대한 배려하여 설계되었다.

결과적으로 53,214개의 적대적 가설 문장을 생성하였다.

3. 작업자 간 검수

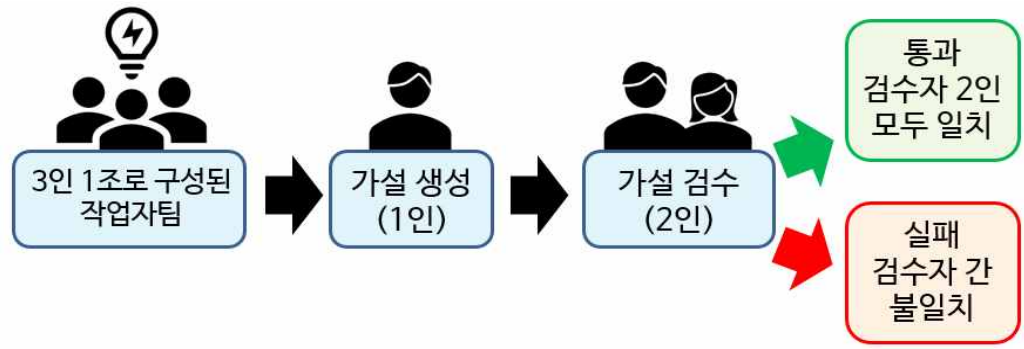
생성한 적대적 사례는 <표 7>과 같이 선행 담화, 가설, 생성된 가설에 대해 원 작업자가 부착한 라벨, 라벨 부착에 대한 원 작업자의 설명, 가설 생성 시 사용된 추론 방식에 대한 주석(annotation), 원 작업자 외에 타 작업자 2인의 가설에 대한 판단(1, 2), 그리고 해당 가설에 대한 모델의 예측값으로 구성된다.

선행 담화	가설	라벨	설명	주석	1	2	모델 예측
국립경주박물관은 1981년 실시한 금척리 고분 18기 발굴조사 결과를 최근 40년 만에 보고서('경주 금척리 신라묘')로 펴냈다.	국립경주박물관이 1981년 실시한 금척리 고분 18기 발굴조사 결과를 2020년대에 와서 보고서로 펴냈다.	E	주어진 문맥에서 근 40년이 지났다고 했으므로 2020년대에 펴냈음을 추론할 수 있다.	Numerical & Quant, Reasoning & Facts	E	E	C

<표 7> 적대적 가설 문장 생성 예

<표 7>에서 가설에 대한 판단은 '주석'에 표시되어 있는데(1, 2) 이것이 최초 작업자의 추론에 대한 다른 작업자 2인의 판정 결과를 표시한 것이다. 즉 <표 7>에 따르면 '라벨'의 E는 최초 작업자가 '함의'로 결정하였음을 표시한 것이고, '주석 1, 2'는 다른 작업자 2인도 최초 작업자와 같이 '함의'(E)로 판정하였음을 나타낸다. 결국 최초 작업자와 다른 작업자(즉 검수자) 2인의 결정이 모두 일치하므로 이들은 그다음 단계인 모델 속이기 과정으로 넘어갈 수 있다.

작업자 간 검수 진행 과정을 도식화해서 보이면 <그림 9>와 같다.



<그림 9> 작업자 간 검수 진행 과정

이러한 과정을 통해 최초 작업자가 생성한 적대적 가설 문장 총 53,214개 가운데 검수자 간 검수를 거친 최종 가설 문장의 수는 38,018개로 집계되었다. 검수자들의 검수를 거치는 과정에서 15,196개의 문장이 배제되었는데 이는 전체 가설 문장의 28%에 해당하는 수치로, 이는 인간의 추론도 기대보다 서로 상이하다는 점을 잘 보여주는 결과로 보인다.

제 4 장

언어 모델 속이기 실험

1. 언어 모델 속이기 실험 설계

본 사업에서는 자연어 이해를 위한 강건한 언어데이터 구축을 위해 두 가지 언어 모델을 대상으로 하여 속이기(fooling) 실험을 수행하였다. 즉 언어 모델이 정확하게 예측하지 못하는 '적대적' 데이터를 구축하기 위해 가설 문장을 생성하고, 검수자 간의 검수를 통과한 문장들에 대해서 '속이기' 실험을 실시하였다. 과업의 초반에는 하나의 모델만을 대상으로 실험적으로 모델 속이기 실험을 진행하다가 최종적으로는 다음과 같이 두 종류의 언어 모델을 활용하여 속이기 실험을 진행하였다 (Park et al., 2021; Lee et al., 2020).

모델명	파라미터 크기
KLUE-RoBERTa base	110M
KRBERT	99M

<표 8> 속이기 실험에 활용된 언어모델과 파라미터 크기

즉, 두 가지 모델을 활용하여 각 모델이 가설과 관련하여 추론한 라벨과 작업자들이 추론한 모델이 일치하는지를 비교하였으며, 이때 두 라벨이 일치하지 않을 경우 속이기(fooling)에 성공한 것으로, 두 라벨이 일치할 경우 속이기에 실패한 것으로 판정된다.

최종적으로는 다음 <표 9>와 같이 두 모델에 대해 모두 속이기에 성공한 경우에 한해 모델 속이기 실험에 통과한 것으로 간주함으로써 매우 강건한 적대적 가설을 확보할 수 있도록 하였다.

작업자 라벨	KLUE-RoBERTa 예측 라벨	KRBERT 라벨	속이기 성공 여부
함의	모순	함의	실패
함의	함의	모순	실패
함의	**모순**	**모순**	**성공**

<표 9> 작업자 라벨과 언어 모델 예측 라벨의 속이기 실험 판정

2. 언어 모델 속이기 실험 결과

언어 모델에 대한 속이기 실험은 모두 7차례의 라운드별로 수행되었다. 각 라운드별 수행 결과를 제시하면 다음과 같다(<표 10>~<표 17>). 각 라운드별 수행 결과의 경우, KLUE-RoBERTa base만을 활용하여 fooling을 진행하였다.

▸ 1라운드 총 9,396 문장 생성 완료, 검수 통과 5,999 문장

의미 관계	대상 문장 수	풀링 성공	풀링 실패	풀링 성공률
함의	3,050	2,346	704	76.92%
모순	1,869	1,583	286	84.70%
중립	1,080	210	870	19.44%
총합계	**5,999**	**4,139**	**1,860**	**68.99%**

<표 10> 1라운드 중 검수를 통과한 적대적 가설 문장의 풀링 실험 결과

▸ 2라운드 총 8,449 문장 생성 완료, 검수 통과 5,641 문장

의미 관계	대상 문장 수	풀링 성공	풀링 실패	풀링 성공률
함의	2,419	1,907	512	78.83%
모순	1,903	1,548	355	81.35%
중립	1,319	288	1,031	21.83%
총합계	**5,641**	**3,743**	**1,898**	**66.35%**

<표 11> 2라운드 중 검수를 통과한 적대적 가설 문장의 풀링 실험 결과

▸ 3라운드 총 8,535 문장 생성 완료, 검수 통과 6,045 문장

의미 관계	대상 문장 수	풀링 성공	풀링 실패	풀링 성공률
함의	2,537	2,022	515	79.70%
모순	2,018	1,618	400	80.18%
중립	1,490	251	1,239	16.85%
총합계	**6,045**	**3,891**	**2,154**	**64.37%**

<표 12> 3라운드 중 검수를 통과한 적대적 가설 문장의 풀링 실험 결과

▸ 4라운드 총 8,570 문장 생성 완료, 검수 통과 6,259 문장

의미 관계	대상 문장 수	풀링 성공	풀링 실패	풀링 성공률
함의	2,670	2,215	455	82.96%
모순	2,067	1,658	409	80.21%
중립	1,522	279	1,243	18.33%
총합계	**6,259**	**4,152**	**2,107**	**66.34%**

<표 13> 4라운드 중 검수를 통과한 적대적 가설 문장의 풀링 실험 결과

▸5라운드 5,972 문장 생성 완료, 검수 통과 4,543 문장

의미 관계	대상 문장 수	풀링 성공	풀링 실패	풀링 성공률
함의	1,905	1,554	351	82%
모순	1,672	1,396	276	83%
중립	966	154	812	16%
총합계	4,543	3,104	1,439	68%

<표 14> 5라운드 중 검수를 통과한 적대적 가설 문장의 풀링 실험 결과

▸6라운드 5,594 문장 생성 완료, 검수 통과 4,283 문장

의미 관계	대상 문장 수	풀링 성공	풀링 실패	풀링 성공률
함의	1,865	1,564	301	84%
모순	1,553	1,261	292	81%
중립	865	114	751	13%
총합계	4,283	2,939	1,344	69%

<표 15> 6라운드 중 검수를 통과한 적대적 가설 문장의 풀링 실험 결과

▸7라운드 6,698 문장 생성 완료, 검수 통과 문장 5,248 문장

의미 관계	대상 문장 수	풀링 성공	풀링 실패	풀링 성공률
함의	2,259	1,866	393	83%
모순	2,026	1,672	354	83%
중립	963	188	775	20%
총합계	5,248	3,726	1,522	71%

<표 16> 7라운드 중 검수를 통과한 적대적 가설 문장의 풀링 실험 결과

위의 1~7라운드를 거쳐 수행한 결과를 종합하면 <표 17>과 같다. 앞에서도 언급한 바와 같이 이 결과는 하나의 언어 모델만을 사용하여 풀링을 수행한 것이다.

의미 관계	대상 문장 수	풀링 성공	풀링 실패	풀링 성공률
함의	16,705	13,474	3,231	81%
모순	13,108	10,736	2,372	82%
중립	8,205	1,484	6,721	18%
총합계	**38,018**	**25,694**	**12,324**	**68%**

<표 17> 1~7라운드 중 검수를 통과한 적대적 가설 문장의 풀링 실험 결과

한편 다음은 의미 관계별 풀링에 성공한 예시와 실패한 예시이다.

(1) 함의

▸[담화] 이는 사회적 거리두기 2단계 상향 기준지표 중 하나인 '일일 확진자 수 50~100명'에 해당한다. 이달 들어 국내 신규 확진자 수는 20~40명대를 오르내렸지만 지난 10일부터는 28명→34명→54명→56명→103명 등 가파른 증가세를 나타

내고 있다.
[가설] 최근의 확진자 수가 가파르게 증가하면서 사회적 거리두기가 2단계로 상향될 가능성이 높아졌다.
[라벨] E
[설명] 사회적 거리두기 2단계 상향 기준지표를 충족하였으므로 2단계로 상향될 가능성이 높아진 것이다.
[추론] standard, tricky
[KLUE-RoBERTa base 예측 라벨] N
[KRBERT 라벨] N
[속이기 성공 여부] 성공

▸[담화] 결국 재판부는 7일 전 목사에 대한 보석을 취소하고 서울구치소에 재수감하는 결정을 내렸다. 재판부는 “형사소송법 제102조 제2항 제5호 ‘법원이 정한 조건을 위반한 때’에 해당하는 사유가 있다”며 “피고인에 대한 보석을 취소하고 보석보증금 중 3000만원을 몰취한다”고 밝혔다.
[가설] 전 목사는 3000만원 이상의 보석 보증금을 냈다.
[라벨] E
[설명] 보석 보증금 중 3000만 원을 몰취한다고 하였으므로, 전 목사는 그 이상의 금액을 보석 보증금으로 낸 적이 있음을 추론할 수 있다.
[추론] standard, tricky
[KLUE-RoBERTa base 예측 라벨] N
[KRBERT 라벨] E

[속이기 성공 여부] 실패

(2) 중립

▸[담화] 수시모집 원서 접수 일정은 전국의 모든 전문대가 동일하게 운영된다. 전형기간 내 면접 실시 등의 고사 일정은 각 대학이 자율적으로 정하고 대학 간 복수 지원 및 입학 지원 횟수도 제한이 없다.
[가설] 수시모집 원서 접수 일정은 전국의 모든 대학교에서 동일하게 운영된다.
[라벨] N
[설명] 수시모집 원서 접수 일정은 전국의 모든 전문대에서 동일하게 운영되지만 모든 대학교에서도 그런지는 알 수 없으므로 중립에 해당한다.
[추론] Reasoning, tricky
[KLUE-RoBERTa base 예측 라벨] E
[KRBERT 라벨] E
[속이기 성공 여부] 성공

▸[담화] 우선 포항시는 지난해 12월 포항과 러시아 블라디보스토크를 오가는 국제 크루즈선 시범 운항을 성공적으로 마쳤다. 올해는 영일만항을 출발하거나 거쳐 가는 국제 크루즈를 5차례 운항할 예정이다.
[가설] 영일만항을 출발하는 국제 크루즈에는 다섯 노선이 있다.
[라벨] N
[설명] 영일만항을 출발하는 국제 크루즈 노선이 몇 개인지는 알 수 없으므로 중립에 해당한다.

[추론] Reasoning, standard
[KLUE-RoBERTa base 예측 라벨] E
[KRBERT 라벨] N
[속이기 성공 여부] 실패

(3) 모순

▸[담화] 이에 수능 난이도 조정에 대한 요구도 높아지는 모양새다. 전국시도교육감협의회는 오는 9일 충남에서 총회를 열고 올해 수능 난이도 조정 등 고3 재학생을 실질적으로 지원할 수 있는 방안에 대해 논의할 예정이다.
[가설] 전국시도교육감협의회는 올해 수능 난이도 조정 방안을 확정하였다.
[라벨] C
[설명] 전국시도교육감협의회는 올해 수능 난이도 조정 방안 등을 논의할 예정이므로, 이를 확정하였다는 것은 모순에 해당한다.
[추론] standard, tricky
[KLUE-RoBERTa base 예측 라벨] N
[KRBERT 라벨] N
[속이기 성공 여부] 성공

▸[담화] 서울의 코로나19 신규 확진자는 18일 하루 14명이 늘어나 19일 0시 기준 누적 5,702명으로 집계됐다. 새로운 집단감염은 없었고, 진행 중인 집단감염 사례인 도봉구 다나병원 관련 2명(환자 2명), 송파구 잠언의료기 및 강남구 콜센터 관련 2명이 나왔다.

[가설] 도봉구 다나병원에서는 지금까지 총 2명의 감염자가 나왔다.

[라벨] C

[설명] 도봉구 다나병원은 집단감염이 진행중이라고 하였으므로, 신규 감염자인 2명보다 많은 감염자가 나왔을 것임을 알 수 있다.

[추론] Numerical & Quant, tricky, reasoning

[KLUE-RoBERTa base 예측 라벨] N

[KRBERT 라벨] C

[속이기 성공 여부] 실패

또한 KLUE-RoBERTa base 모델을 대상으로 했을 때 어떤 가설 생성 방식으로 만들어진 가설이 풀링 성공 비율이 높았는지도 정리해 보았다. <표 18>은 그 결과를 보여주고 있는데 이에 따르면 '속이기(tricky)' 방식으로 생성된 가설이 풀링 성공률이 가장 높은 것으로 나타났다.

	num.	refer.	std	lexical	tricky	reason	계
풀링 성공 수	7,385	3,420	9,882	8,234	14,952	7,245	24,330
풀링 성공 비율	**30.3%**	**14.1%**	**40.6%**	**33.8%**	**61.5%**	**29.8%**	**100.0%**

<표 18> KLUE-RoBERTa base 모델에서의 추론 방식별 풀링 성공 비율

한편 데이터 세트의 강건성을 확보하기 위해 기존의 언어 모델(KLUE-RoBERTa base)에 하나의 언어 모델(KRBERT)을 추가하여 풀링을 수행하였다. 이를 통해 적대성의 정도를 더욱 높이게 됨으로써 결과적으로 언어 모델의 추론 능력 향상에 기

여할 수 있을 것이다(<표 19>).

의미 관계	대상 문장 수	풀링 성공	풀링 실패	풀링 성공률
함의	16,705	11,952	4,753	71.55%
모순	13,108	8,275	4,833	63.13%
중립	8,205	1,073	7,132	13.08%
총합계	**38,018**	**21,300**	**16,718**	**56.03%**

<표 19> 1~7라운드 중 검수를 통과한 적대적 가설 문장의 풀링 실험 결과

최종 데이터의 경우, 7라운드에 걸쳐 생성된 적대적 가설 중 두 가지 모델 모두(KRBERT, KLUE-RoBERTa base)에 대해 fooling에 성공한 가설만을 취했다. 그 결과 총 21,300개의 사례가 집계되었다.

집계된 21,300개의 사례 중 가설 문장 자체에 정치, 사회적으로 윤리적 문제의 소지가 있거나 혹은 오타 및 오류가 있는 경우 혹은 가설 문장의 전제 문장에 상기에 언급된 문제가 있는 경우를 제외하는 사후 재검수 작업을 통해 최종적으로 총 20,052개의 사례가 수합되었다.

제 5 장

검증

1. 검증 절차와 결과

국립국어원의 2021년 신문 말뭉치로부터 선행 담화를 추출하고, 이를 토대로 적대적 가설을 생성한 다음 작업자 간 검수와 언어 모델 속이기 실험까지 통과한 최종적인 적대적 가설 말뭉치가 구축되었다. 이렇게 구축된 적대적 가설 말뭉치는 언어 모델의 학습 데이터로서의 강건성을 유지할 것으로 기대되었으나 실제로 어느 정도 기대에 충족될 수 있는지에 대해 점검할 필요가 있다. 이 장에서는 본 사업에서 구축된 적대적 가설 말뭉치에 대하여 총 5개의 언어 모델을 활용하여 데이터 세트의 적대성을 검증하도록 하였다.

▸모델링 개요

다음의 <표 20>은 적대성 검증에 활용된 언어 모델과 파라미터를 정리한 것이다.

모델명	파라미터 크기
KLUE-RoBERTa large	337M
KLUE-RoBERTa base	110M
KoELECTRA-Base-v3 (이하 KoELECTRA)	112M
KRBERT	99M
KLUE-BERT	110M

<표 20> 검증에 활용된 언어 모델과 파라미터

가설의 적대성 검증은 다음과 같은 과정으로 진행되었다.

먼저, 기존의 한국어 NLI 데이터 세트(KLUE-NLI)를 가지고 미세조정된 모델들이 재검수 작업을 거치기 이전의 총 21,300건의 ANLI 데이터 세트에 대해 어느 정도의 성능을 보이는지 확인하였다. 두 번째로, 기존 한국어 NLI 데이터 세트를 가지고 미세조정된 모델들과 ANLI를 가지고 미세조정을 진행한 모델들이 ANLI 데이터 세트에 보이는 성능을 비교하였다. 마지막으로, ANLI를 가지고 미세조정을 진행한 모델들이 기존 한국어 NLI 데이터 세트에 어떠한 성능을 보이는지 평가하도록 하였다.

▸실험 배경과 기대

이번 적대적 함의 분석 데이터 세트의 기반이 된 연구(Nie et al., 2019)에서와 마찬가지로, 구축된 데이터 세트의 품질을 검증하기 위하여 평가 실험을 진행하였다. 이번 사업의 중요 동기 중 하나는 현재 벤치마크의 개발 속도가 언어 모델의 급속한 발전 속도를 따라잡지 못하고 있다는 것이었다(Nie et al., 2020). 즉, 이번 사업을 통해 구축된 ANLI가 적절한 방식과 절차를 통해 구축되고 검수가 이루어졌다면, 현재 언어 모델과 앞으로 등장할 언어 모델의 평가가 가능하도록 기존 자연어추론 데이터 세트에 비해 그 난도가 높아야 할 것이다.

또한, ANLI는 적대적 방법론을 활용하였다. 적대적 사례란, 인공지능 모델의 약점을 공략할 수 있는 사례를 일컫는 것으로 인공지능 언어 모델들이 벤치마크 데이터 세트에 존재하는 통계적 편향을 부적절하게 학습하거나 활용하는 것을 막을 것이다. 이번 사업에서는 구축 과정에서 데이터의 적대성을 확인하기 위해 모델 속이기 방식을 적용한 바 있다. 만약, ANLI 데이터 세트의 구축에 있어 적대적 방법론이 적절하게 적용되었다면, 데이터 세트의 적대적인 난도와는 별개로 모델 속이기에 활용된 모델들은 특히 더 낮은 성능을 보여주어야 할 것이다.

이 두 가지 관점에서, 만약 ANLI가 적절하게 구축되었다면 첫 번째 평가에서 기존 데이터 세트를 통해 미세조정이 진행된 모델들은 ANLI에 매우 낮은 성능을 보여야 하며 또한, 모델 풀링에 활용된 KLUE-RoBERTa base와 KRBERT는 ANLI에 대해 다른 모델보다도 훨씬 떨어지는 성능을 보여야 한다.

▸한국어 인공지능 모델 개발에의 활용 가능성

선행 연구에서 모델의 학습이 적대적 사례를 기반으로 한 벤치마크 데이터 세트를 통해 이루어진 경우 모델의 견고성이 향상되고 과업 수행의 측면에서도 개선된 성능을 보여주었다. 즉, 기존 데이터 세트에 비해 '적대성'을 활용한 ANLI 데이터 세트가 한국어 인공지능 모델의 성능 개발에 기여할 수 있을 것으로 기대된다. 이러한 가능성을 일부 확인하기 위해서는, ANLI 데이터 세트를 통해 미세조정을 진행한 모델들이 데이터 세트를 통해 미세조정이 진행된 모델들보다 개선된 성능을 보여야 한다. 여기에 더해, ANLI 데이터 세트를 통해 미세조정을 진행한 모델들이 기존 데이터 세트에 대해서도 개선된 성능을 보여야 할 것이다.

▸미제조정(fine-tuning) 및 평가 데이터

5개의 모델에 대해 각각 두 종류의 데이터 세트를 활용하여 미세조정을 진행하였다. 이 중 한 데이터 세트는 기존에 적대적 방법론을 활용하지 않은 방식으로 구축된 KLUE-NLI 데이터 세트, 나머지 한 데이터는 이번 사업을 통해 구축된 ANLI 데이터 세트를 활용하였다.

KLUE-NLI 데이터 세트(Park et al., 2021) 중, 훈련 데이터 24,998건과 개발 데이터 세트 3,000건을 활용하여 5개 모델에 대한 미세조정을 진행하였다.

ANLI데이터의 경우, 생성 후 검수를 통과한 38,018건의 데이터 중, KLUE-NLI

데이터를 통해 미세조정한 KLUE-RoBERTa base와 KRBERT를 속이는데 모두 성공한 데이터 21,300의 데이터를 가지고 실험을 진행하였다.

이때, 전체 21,300건의 데이터는 학습, 검증, 시험을 각각 8:1:1의 비율로 나누어 활용하였다. 즉, 최종 적대성 검증에서 5개의 모델을 학습시키는데 사용된 데이터는 21,300건의 80%에 해당하는 17,040건이다. 이들 각각의 세부적인 내용은 <표 21>과 같다.

	{|Train|, |Dev|, |Test|}	원문 데이터	문체
KLUE-NLI	24998, 3000, 3000	뉴스, 위키피디아, 리뷰	구어, 문어
ANLI	17040, 2130, 2130	뉴스	문어

<표 21> 5개 모델 학습에 활용된 데이터 세부 내역

▸평가

평가에는 미세조정 시에 활용된 두 개의 데이터 세트에서 총 세 종류의 하위 데이터 세트를 추출하여 활용하였다. 먼저, 첫 번째 평가에서는 KLUE-NLI를 통해 미세조정을 진행한 총 5개의 모델들에 대해 ANLI 데이터 세트 전체에 대한 성능을 확인하였고, 두 번째 평가에서는 KLUE-NLI를 통해 미세조정을 수행한 모델과 ANLI를 통해 미세조정을 진행한 총 10개의 모델(5개의 모델 * 2개의 미세조정 데이터 세트)을 ANLI 시험 데이터 세트를 가지고 평가하였다.

이때, 성능 비교에 사용된 ANLI 시험 데이터 세트는 미세조정 과정에서 전혀 사용되지 않은 별개의 데이터로, ANLI 훈련 데이터를 가지고 미세조정이 진행된 모델 또한 KLUE-NLI를 가지고 미세조정된 모델들과 마찬가지로 시험 데이터를 처음

접하도록 설계하였다. 마지막 평가에서는 ANLI를 통해 미세조정을 진행한 총 5개의 모델들에 대해 KLUE-NLI 데이터 세트 중 훈련, 검증 데이터 세트들에 대한 성능을 확인하는 과정을 거쳤다.

▸실험 방법

세 종류의 평가 모두 같은 방식으로 진행되었는데, 각각의 모델들이 평가 데이터 세트의 라벨을 어떤 식으로 예측하는지 확인하였다. 이때 평가 기준으로는 두 가지 척도, 정확도(accuracy)와 F1 점수를 활용하였다.

즉 기존 데이터 세트를 통한 모델 성능 평가에서는 정확도를 활용하였으며(Park et al., 2021), 여기에 더해 두 데이터 세트에 존재하는 라벨 간 불균형에 대응하기 위하여 F1 점수 또한 평가 기준으로 활용하였다.

▸실험 결과

먼저, ANLI 데이터 세트에 대해 기존 한국어 NLI 데이터 세트(KLUE-NLI)를 가지고 미세조정된 모델들이 보이는 성능을 확인하였으며 그 결과는 <표 22>와 같다.

모델	미세조정 데이터	ANLI 전체	
		F1	정확도
KLUE-RoBERTa large	KLUE-NLI	14.11	12.71
KLUE-RoBERTa base	KLUE-NLI	0.00	0.00
KoELECTRA-Base-v3	KLUE-NLI	11.30	11.24
KRBERT	KLUE-NLI	0.00	0.00
KLUE-BERT	KLUE-NLI	11.40	10.89

<표 22> 기존 한국어 NLI 데이터(KLUE-NLI)로 미세 조정된 모델의 성능

검증을 진행한 결과, 모델의 파라미터 수나 모델의 종류와 관계없이 기존 한국어 자연어추론 데이터를 통해 미세조정이 진행된 모델들의 경우 ANLI 데이터에 대해서는 성능이 크게 떨어지고 있음을 확인하였다.

파라미터 수가 337,000개로 이번 검증에 사용된 모델 중 그 규모가 가장 큰 KLUE-RoBERTa large 모델의 점수조차 각각 정확도 12.71, F1 점수 14.11로 매우 낮은 수준의 성능을 보여주었다. 그러나 가장 큰 모델인 만큼 다른 모델보다는 상대적으로 높은 점수를 보여주었다.

그러나 ANLI 구축의 속이기 과정에 활용된 KLUE-RoBERTa base와 KRBERT의 경우, 정확도와 F1점수 모두에서 0점을 보여주었다.

그다음 과정으로 KLUE-NLI와 ANLI의 훈련 데이터를 통해 미세조정을 진행한 모델들이 같은 ANLI 시험 데이터 세트에 대해 보이는 성능을 비교하였으며 그 결과는 <표 23>과 같다.

모델	미세조정 데이터	ANLI 시험	
		F1	정확도
KLUE-RoBERTa large	KLUE-NLI	15.70	14.46
KLUE-RoBERTa base	KLUE-NLI	0.00	0.00
KoELECTRA-Base-v3	KLUE-NLI	10.96	10.85
KRBERT	KLUE-NLI	0.00	0.00
KLUE-BERT	KLUE-NLI	11.53	11.13
KLUE-RoBERTa large	ANLI	61.83	73.71
KLUE-RoBERTa base	ANLI	52.84	67.89
KoELECTRA-Base-v3	ANLI	57.83	68.87
KRBERT	ANLI	55.89	68.31
KLUE-BERT	ANLI	58.58	70.28

<표 23> KLUE-NLI와 ANLI의 훈련 데이터로 미세조정을 진행한 모델들의 성능

성능 평가를 진행한 결과, KLUE-NLI를 통해 미세조정을 진행한 모델들은 ANLI 훈련 데이터를 통해 미세조정이 진행된 모델들에 비해 매우 떨어지는 성능을 보였다.

먼저 KLUE-NLI를 통해 미세조정을 진행한 모델의 경우, 구축 과정의 적대성 검증에 사용된 KLUE-RoBERTa base와 KRBERT를 제외하면 일반적으로 11점에서 16점 사이의 점수 분포를 보이고 있다.

반면, ANLI 훈련 데이터를 통해 미세조정이 진행된 모델들은, F1 점수 중 가장 낮은 점수가 52.84, 정확도 중 가장 낮은 점수가 67.89일 만큼 기존 데이터를 통해 미세조정이 진행된 데이터들에 비해 월등히 높은 성능을 보여주고 있다.

그러나 ANLI를 통해 미세조정이 진행된 데이터들 또한 뛰어난 성능을 보여주고 있지는 못했다. 기존에 KLUE-RoBERTa large, KLUE-RoBERTa base, 그리고 KLUE-BERT는 KLUE-NLI 데이터 세트에 대해서 정확도가 각각 89.17, 84.83, 81.63으로, 모두 80점이 넘는 성능을 보였다(Park et al., 2021).

이에 비해 ANLI 데이터 세트의 경우, 정확도의 최고점이 KLUE-RoBERTa large의 63.71점으로 80점을 채 넘기지 못하고 70점대 초반에 머물러 있었으며, 다른 모델들은 모두 70점을 넘기지 못하는 상당히 낮은 점수 분포를 나타냈다.

추가적으로, ANLI를 통해 미세조정을 진행한 모델을 대상으로 KLUE-NLI 데이터에 대한 성능 평가를 진행하였으며 그 결과는 <표 24>와 같다.

모델	미세조정 데이터	KLUE-NLI	
		F1	정확도
KLUE-RoBERTa large	ANLI	14.87	15.33
KLUE-RoBERTa base	ANLI	14.35	15.91
KoELECTRA-Base-v3	ANLI	14.57	14.35
KRBERT	ANLI	17.13	17.55
KLUE-BERT	ANLI	16.60	16.43

<표 24> ANLI로 미세조정을 진행한 모델의 성능 평가

이에 따르면 ANLI로 미세조정을 진행한 모델들 또한 전반적으로 14점에서 17점 사이의 10점대 중반의 점수 상대적으로 떨어지는 성능을 보였다. 그중 가장 좋은

성능을 보여준 모델은 KRBERT로 F1 점수와 정확도 모두 17점을 넘는 가장 높은 점수를 보여주었다. 그러나 이는 다른 모델들의 성능이 충분하지 못하기 때문으로, 상대적으로 높게 측정된 이러한 성능이 유의미하다고 보기는 어렵다. 물론 KLUE-NLI로 미세조정을 진행하여 ANLI 데이터에 대한 성능을 확인했던 결과보다는 향상된 결과를 보여주기는 하였다.

2. 실험 결과의 해석

첫 번째 평가와 두 번째 평가에서 모두, 모델의 파라미터 수나 모델의 종류와 관계없이, 기존 한국어 자연어추론 데이터를 통해 미세조정이 진행된 모델들의 경우 ANLI 데이터에 대해 굉장히 떨어지는 성능을 보여주었다. 이러한 결과는, ANLI 데이터 세트가 가지는 높은 난도를 보여주는 것이며, ANLI 구축의 목적 중 하나인 '오래가는' 데이터 세트가 되기 위한 중요한 요건 중 하나이다.

즉, 실험 평가 결과를 통해 ANLI 사업의 가장 중요한 목적 중 하나인 '고난도의 자연어추론 데이터 세트 구축'이 성공적으로 이루어졌음을 확인하였다.

▸ANLI 구축 시 모델 활용의 적절성

ANLI 구축 시 속이기 과정에 활용된 KLUE-RoBERTa base와 KRBERT의 경우, 첫 번째와 두 번째 평가 모두에서 0점의 정확도와 F1점수를 보였다. 이는 ANLI 데이터 세트 구축 과정에서 데이터의 적대성 검증을 할 때 두 모델을 속일 수 있는지 여부로 진행했기 때문이다. 이러한 결과는 이번 과제가 참고한 선행 연구(Nie et al., 2019)에서도 동일하게 발견된 것이었다. 또한 두 모델의 정확도와 F1점수가 모두 0점이 나온 것은, 구축 과정에서 두 모델에 대한 적대성 검증이 충실하게 진행되었으며, 두 모델의 약점을 충분히 활용하는 방향으로 ANLI가 구축되었음을 잘 보여주는 것이다.

▸ANLI에 대한 모델 성능의 경향

첫 번째와 두 번째 평가 미세조정 데이터와 평가 데이터에 상관없이, 모두 파라

미터 크기가 가장 큰 KLUE-RoBERTa large 모델이 가장 높은 성능을 보여주었다. 이는 언어 모델 평가의 영역에서 일반적으로 기대되는 결과이다. 동시에, 모델의 종류에 상관없이 ANLI에 대해서는 전반적으로 낮은 점수를 보였다.

한편 검증에 사용된 5개의 모델은 크게 BERT 기반 모델, ELECTRA 기반 모델, ROBERTA 기반 모델로 나누어 볼 수 있다. 다양한 종류의 모델이 검증에 사용되었지만, 기존 데이터 세트를 통해 미세조정이 이루어진 경우 그 종류에 상관없이 전반적으로 15점을 넘지 못하는 낮은 성능을 보여주었다.

즉, 구축 과정에서의 모델 속이기 단계에는 KLUE-RoBERTa base와 KRBERT의 두 모델을 활용하였지만, 이 두 모델뿐만 아니라 다양한 아키텍처의 모델에 대해서도 매우 난도가 높은 데이터라는 것을 확인하였다. 즉, 이번 사업을 통해 구축된 적대적 사례 데이터는 다양한 모델들의 성능 평가에 활용될 수 있는 강건성을 갖춘 데이터 세트라는 것을 실험을 통해 최종적으로 확인하였다.

▸데이터의 일반화 가능성

ANLI로 미세조정을 한 데이터와 KLUE-NLI로 미세조정을 진행한 모든 모델에서, 각각 자신의 미세조정 데이터 세트과 같은 방식으로 구축된 시험 데이터에서는 더 나은 성능을 보인 반면, 다른 방식으로 구축된 데이터 세트의 시험 데이터는 매우 부족한 성능을 보여주었다. 이러한 경향은, 자연어 추론 데이터의 일반화 가능성과 연관이 있는 것으로 해석할 수 있는데, 이에 대해서는 향후 연구가 지속될 필요가 있을 것이다.

3. 자문회의를 통한 최종 검토

본 사업의 결과인 함의 분석 말뭉치는 인공지능을 비롯하여 다양한 영역에서 기초 자원으로 활용될 수 있기 때문에 구축된 말뭉치의 방향성, 확장 및 활용 가능성 등에 대한 외부 전문가들의 자문을 구하는 과정이 필요하다. 본 사업에서는 학계와 산업계에서 왕성하게 활동하고 있는 5명의 자문위원단을 국립국어원과 함께 구성하고 10월과 12월, 두 차례에 걸쳐 자문회의를 실시함으로써 말뭉치의 신뢰성과 정확성, 범용성 등을 개선하고자 하였다. 10월에 열린 자문회의는 온라인 회의로, 12월의 자문회의는 서면 조사 형식으로 개최되었다.

▸1차 자문회의(2023년 10월 5일, 온라인 회의로 진행)

• 주요 자문 내용

- 작업의 추론 방식과 관련한 균형성 문제(6개의 추론 방식에 따른 구축량의 비율은 어떻게 구성하고 있느냐는 자문단의 의견)
- '설명' 부분에 대한 작성을 좀 더 정교화할 필요가 있다는 자문단의 조언
- 모델의 속이기(fooling)에 있어 대규모 언어 모델을 대상으로 수행하였는지에 대한 지적(대규모 언어 모델이 포함되어 있어야 한다고 조언해 주었으며, 사업단에서는 이를 확인하였음)
- 속이기의 성공 비율과 이에 따른 작업량 계산을 적절히 염두에 두고 있는지에 대한 확인 조언
- 선행 담화가 모두 신문 기사로 한정되어 있다는 것에 대한 문제를 고려하고 있어야 한다는 조언, 그리고 섹션별로 어떻게 선행 담화를 구성하였는지도 고려해야 한다고 지적함.

- 전반적인 사업의 방향성과 결과에 대해서는 모두 긍정적인 반응

▸2차 자문회의(2023년 12월, 서면조사 형식으로 진행)

- 1차 자문회의에서 매우 상세하게 의견을 공유하였고, 사업단에서도 자문위원단의 조언을 적극 수용하여 사업을 진행한 결과 2차 자문회의는 서면 형식으로 간략히 진행하였음.
- 1차 자문회의에서 혹시 누락된 부분, 추가적인 의견, 최종 결과물인 JSON 형식 등에 대해 자문을 요청하였으며 이에 대해 자문위원단은 사업 결과물에 대해 높은 기대감을 가지고 있으며, 최종 결과물에 형식 오류가 없도록 유의해 줄 것을 당부하였음. 아울러 선행 담화의 길이가 다소 길다는 의견도 있었는데, 이는 선행 담화가 너무 길 경우 추론에 부정적인 영향을 줄 수 있기 때문이라고 함. 한편 최종 공개 파일 형식을 JSON뿐 아니라 csv 등의 형식도 제공해 주는 것이 다양한 분야의 연구자들이 활용하기 좋을 것이라는 의견도 나왔음.

제 6 장

결론

지금까지 서술된 내용을 간략히 요약하고, 향후 과제를 제언하는 것으로 결론을 삼고자 한다.

1. 요약

본 사업은 한국어 인공지능의 과업 수행 능력 일반과 높은 상관관계를 갖고 있는 언어 추론인 '함의 분석'을 위해 신문 텍스트로부터 선행 담화를 추출하고, 이를 기반으로 적대적 사례를 생성하여 인공지능 평가용 벤치마크로 가공함으로써 자연어 이해 벤치마크의 취약성 문제를 보완하기 위한 방안을 마련하는 데 목적이 있다.

이를 위해 2021년 국립국어원 신문 말뭉치로부터 7차례의 라운드를 통해 모두 630,000건의 선행 담화를 추출하고 이를 대상으로 적대적 가설을 생성하였다. 적대적 가설 문장을 생성하기 위해서는 선행 연구들을 참조하여 6가지의 가설 생성 전략을 기반으로 한 구축 지침을 수립하여 적용하였고, 작업의 효율성을 최대한으로 높이기 위해서 작업용 벤치마크를 개발, 활용하였다. 이를 통해 모두 53,214개의 적대적 가설을 생성하였다. 한편 생성된 적대적 가설은 추론 판단의 정확성을 위해 2명의 작업자의 검수를 거치도록 하였다. 결과적으로 3인의 작업자가 추론 판정에 동의한 문장만이 다음 단계에 진입할 수 있게 되었다. 이렇게 작업자 간 검수까지 통과한 적대적 가설은 모두 38,018개에 해당하였다.

적대적 가설 데이터 세트가 인공지능 언어 모델의 강건성을 확보하는 데 유효한 것인지를 판단하기 위해 언어 모델의 속이기(fooling) 실험을 수행하였다. 최종적으로 KLUE-RoBERTa base와 KRBERT 두 개의 언어 모델에 대한 속이기 실험을 통과한 적대적 가설만으로 데이터 세트를 구성하였다. 이 데이터 세트는 모두 21,300

개의 문장으로 구성되었다. 마지막으로 5개의 언어 모델을 활용하여 본 사업을 통해 구축된 데이터 세트가 충분히 '적절한' 데이터인지를 검증하였고, 실험 결과 본 사업을 통해 구축된 적대적 가설 데이터가 언어 모델의 성능 개선에 유의미한 결과를 보인다는 점을 확인하였다.

2. 제언

이번 사업의 선행 담화는 모두 2021년 국립국어원 신문 말뭉치(2020년 신문 기사로 구성)로부터 추출하였다. 이는 선행 담화를 신문이라는 텍스트로 한정하였다는 점에서 신문에 특성화된 데이터라는 특징을 가진다. 이는 일반적, 전반적인 언어 모델을 위해서는 한계로 작용할 수 있다. 신문이라는 텍스트가 가지는 장르적 성격에서 자유로울 수 없기 때문이다. 특히 2020년 신문 기사에는 코로나 관련 뉴스의 비중이 엄청나게 높기 때문에 아무리 균형 있게 자료를 추출한다고 하더라도 텍스트 내용의 편향성을 모두 극복했다고 보기에는 한계가 있다. 향후 연구에서는 선행 담화의 추출 대상 텍스트를 더욱 다양하게 확장함으로써 언어 모델의 범용성과 강건성을 강화하는 데 도움이 될 것이다.

한편 2만여 건의 적대적 가설을 생성, 검수하고 복수의 언어 모델을 대상으로 속이기 실험까지 수행하는 일련의 과정은 매우 복잡하고 많은 단계가 필요할 뿐 아니라 지침의 개발과 교육, 그리고 이를 토대로 한 문장 생성에 이르기까지 많은 연구 참여 인력과 노력, 경험이 요구되는 과정이었다. 본 사업의 결과물은 이러한 복잡하고 지난한 과정을 통해 산출되었으며, 여러 차례의 검증과 수정, 보완을 거침으로써 언어 모델의 추론 능력을 강건하게 만드는 데 기여할 수 있을 것으로 기대한다.

참고문헌

Bentivogli, L., Clark, P., Dagan, I., & Giampiccolo, D. (2009). The Fifth PASCAL Recognizing Textual Entailment Challenge. In TAC.

Bos, J., & Markert, K. (2005). Recognising textual entailment with logical inference. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (pp. 628-635).

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In NAACL.

Condoravdi, C., Crouch, D., De Paiva, V., Stolle, R., & Bobrow, D. (2003). Entailment, intensionality and text understanding. In Proceedings of the HLT-NAACL 2003 workshop on Text meaning (pp. 38-45).

Dagan, I., Glickman, O., & Magnini, B. (2005, April). The pascal recognising textual entailment challenge. In Machine Learning Challenges Workshop (pp. 177-190). Springer, Berlin, Heidelberg.

Fyodorov, Y., Winter, Y., & Francez, N. (2000). A natural logic inference system. In Proceedings of the 2nd Workshop on Inference in Computational Semantics (ICoS-2).

Ham, J., Choe, Y. J., Park, K., Choi, I., & Soh, H. (2020). Kornli and korsts: New benchmark datasets for korean natural language understanding. arXiv preprint arXiv:2004.03289.

Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020, April). Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In Proceedings of the AAAI conference

on artificial intelligence (Vol. 34, No. 05, pp. 8018-8025).

Khot, T., Sabharwal, A., & Clark, P. (2018). Scitail: A textual entailment dataset from science question answering. In Thirty-Second AAAI Conference on Artificial Intelligence.

Lee, S., Jang, H., Baik, Y., Park, S., & Shin, H. (2020). Kr-bert: A small-scale korean-specific language model. arXiv preprint arXiv:2008.03979.

MacCartney, B., & Manning, C. D. (2009, January). An extended model of natural logic. In Proceedings of the eight international conference on computational semantics (pp. 140-156).

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2019). Adversarial NLI: A new benchmark for natural language understanding. arXiv preprint arXiv:1910.14599.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020). Adversarial NLI: A New Benchmark for Natural Language Understanding, ACL 2020.

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K. (2021). Klue: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. Advances in neural information processing systems, 32

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R.

(2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Williams, A., Nangia, N., & Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426.

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. ACM Transactions on Intelligent Systems and Technology (TIST), 11(3), 1-41.

<부록> 2022년 말뭉치 함의 분석 연구 및 구축 지침

2022년 말뭉치 함의 분석 및 연구 구축 지침

1. 개요

1.1. 본 과제의 목적

본 과제 '2022년 말뭉치 함의 분석 및 연구'에서는 관련한 선행 연구(Nie et al. 2020)를 기반으로 적대적 사례에 기반한 자연어추론 데이터 세트를 구축하는 것을 목표로 한다. '화이트해커'인 주석자를 활용하여 적대적 사례를 구축하고 작업자로 하여금 문맥에 맞는 가설을 생성하도록 하여 라벨과 다른 대답을 내놓도록 모델을 속임(fooling)으로써 새로운 데이터 세트을 만드는 것이 주요한 목적이다.

1.2. 적대적 사례 개괄

적대적 사례 말뭉치는 6개의 의미화용적 추론방식(numerical&quant, reference, standard, lexical, tricky, reasoning)을 이용하여 함의 분석된 적대적 사례에 기반한 자연어 문장 데이터 세트이다(Nie et al. 2020). 함의 분석이란 주어진 대상 문장(담화)와 가설 문장에 대해 둘 간의 관계를 참과 거짓의 정도에 근거하여 추론하는 작업이다. 함의(entailment), 중립(neutral), 모순(contradiction)의 세 가지 중 어떤 경우에 속하는지를 추론하여 라벨을 부착한다. 대상 문장과 가설 문장 간의 관계 추론 예시는 아래와 같다:

(1) 대상 문장: 올빼미 버스는 오후 11시부터 오전 6시까지 운영된다.

가설 문장: 올빼미 버스의 운영 시간은 7시간이다.
라벨: 함의(E)
설명: 오후 11시부터 오전 6시는 총 7시간이므로 운영 시간은 7시간이다.

1.3. 주요 용어

가. 대상 담화: 선행 문장과 대상 문장으로 최대 두 문장으로 구성된다.

(2) 커튼을 걷고 밖을 내다보자 어제는 어둡고 경황이 없어서 볼 수 없었던 새로운 풍경이 눈앞에 펼쳐졌다. 숙소 바로 뒤쪽으로 작은 호수가 있었고, 잔잔한 물결에 반짝반짝 비치는 빛이 그동안의 모든 걱정을 한 번에 씻어내는 듯하다.

나. 대상 문장: 가설 문장과 함의 관계를 이루는 문장으로 추론 방식이 적용될 수 있도록 명사/대명사 등의 인칭 정보와 수/양/날짜 등의 수치적 정보, 반의어/유의어/접속/부정어 등의 논리적 정보, 구문변환/재정렬 등의 통사적 정보, 인과관계, 작가의도 추론과 같은 문맥적 정보를 포함한다.

(3) 숙소 바로 뒤쪽으로 작은 호수가 있었고, 잔잔한 물결에 반짝반짝 비치는 빛이 그동안의 모든 걱정을 한 번에 씻어내는 듯하다.

다. 가설 문장: 대상 문장과 함의 관계를 이루는 문장으로 적대적 사례의 역할을 하는 문장이다.

(4) 숙소 바로 뒤쪽에는 작은 호수가 있다.

라. 라벨: 대상 문장과 가설 문장 간의 함의 관계이다. 함의, 중립, 모순 중 한 가지로 선택된다. 아래는 대상 문장 (3)과 가설 문장 (4) 가설 문장 간의 함의 관계이다.

(5) 함의(E)

마. 설명: 작업자가 라벨을 선택한 이유에 대한 설명이다.

(6) '숙소 바로 뒤쪽으로 작은 호수가 있었다'는 명제가 대상 문장에 접속되어 연결되므로 대상 문장이 참이면 가설 문장도 반드시 참이다.

바. 함의 관계: 대상 문장과 가설 문장 간의 의미화용적 관계이다. 본 과제에서는 논리학적인(logical) 관계뿐만 아니라 자연어 추론에서 다루는 함의 관계까지 확장시켜 다음과 같이 함의 관계를 정의한다(Dagan and Glickman 2004; Glickman 2006).

i. 함의(entailment): 대상 문장의 진리치가 참일 때, 가설의 진리치가 반드시 참인 경우이다(wirte one sentence that is **definitely correct** about the situation or event in the line).

(7) 대상 문장: 숙소 바로 뒤쪽으로 작은 호수가 있었고, 잔잔한 물결에 반짝반짝 비치는 빛이 그동안의 모든 걱정을 한 번에 씻어내는 듯하다.
가설 문장: 숙소 바로 뒤쪽에는 작은 호수가 있다.
레이블: 함의(E)

ii. 중립(neutral): 대상 문장의 진리치가 참일 때, 가설의 진리치가 참인지 거짓인지 판단할 수 없을 경우이다(wirte one sentence that is **might be correct** about the situation or event in the line).

(8) 대상 문장: 숙소 바로 뒤쪽으로 작은 호수가 있었고, 잔잔한 물결에 반짝반짝 비치는 빛이 그동안의 모든 걱정을 한 번에 씻어내는 듯하다.
가설 문장: 숙소 바로 뒤쪽으로 큰 나무가 있다.
레이블: 중립(N)

iii. 모순(contradiction): 대상 문장의 진리치가 참일 때, 가설의 진리치가 반드시 거짓인 경우이다(write one sentence that is **definitely incorrect** about the situation or event in the line).

(9) 대상 문장: 숙소 바로 뒤쪽으로 작은 호수가 있었고, 잔잔한 물결에 반짝반짝 비치는 빛이 그동안의 모든 걱정을 한 번에 씻어내는 듯하다.
가설 문장: 숙소 바로 뒤쪽에는 작은 호수가 없다.

레이블: 모순(C)

☞ 주어진 언어적 정보와 일반 지식을 기반으로 하여 일반 언중의 직관에 따라 가설 문장의 '함의/중립/모순'을 판단한다(한지윤 2020).

자. numerical & quant(수와 양): 기수 및 서수에 대한 추론, 숫자에서 날짜 및 연령 추론.

차. reference & name(지시와 이름): 대명사와 고유 명사 간의 상호 참조, 이름, 성별에 대한 추론

카. standard(표준): 접속, 부정, 인과관계, 비교급과 최상급 추론

타. lexical(어휘): 동의어, 반의어에 대한 어휘 정보

파. tricky(속임수): 말장난(wordplay), 구문 변환(syntactic transformation), 재정렬(reordering)/문맥에서 작가 의도 추론과 같은 언어 전략 추론

하. reasoning & facts(추론 및 사실): 세상에 대한 외부 지식 또는 추가 사실로부터 추론

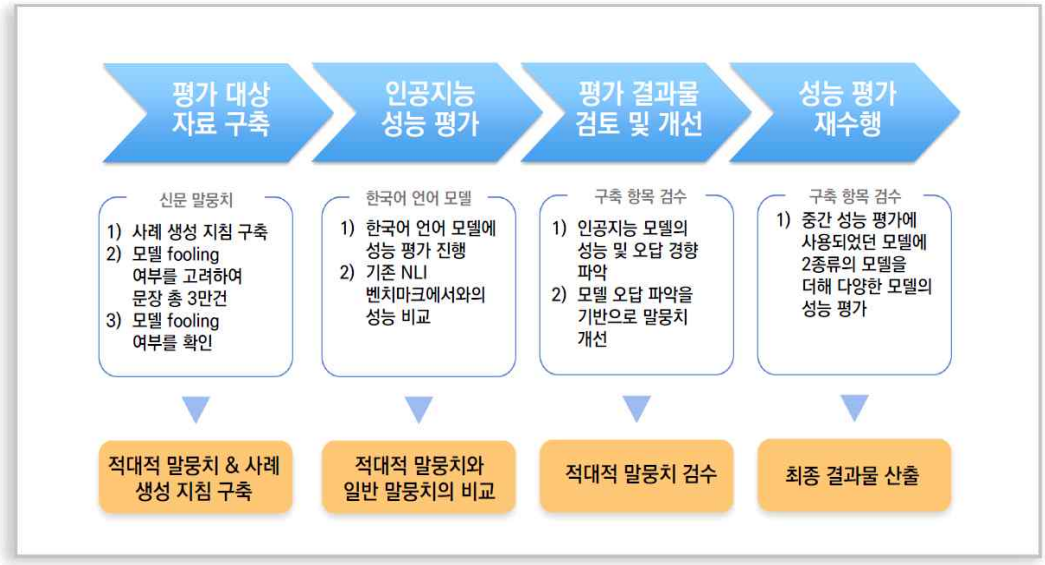
2. 작업 절차

2.1 구축 절차

구체적인 사업 내용은 다음 세 가지로 구성된다. 첫째, 적대적 함의 관계 분석 지침을 수립한다. 지침 수립의 내용은 분석 대상을 선정하여 분석 방법론을 수립하고 모델 학습 및 평가 결과를 검증하고 개선을 위한 방법론을 수립하는 것이다. 둘째, 적대적 함의 관계 말뭉치를 구축한다. 구축의 단계는 다음과 같다. 먼저 공개된 국립국어원 신문 말뭉치 2021(2020년 생산된 신문 기사

729,280건)에서 분석 대상을 선정한다. 그다음 적대적 함의 관계 생성을 위한 가설 도출 및 설명 정보를 부착한다. 마지막으로 말뭉치로 가공 및 정비(2만 건 이상)한다. 셋째, 검증 평가 체계 및 품질 관리 계획을 수립한다. 모델을 활용한 인공지능 성능 평가를 2회 이상 진행하고 이후 검수 및 개선 체계를 수립한다.

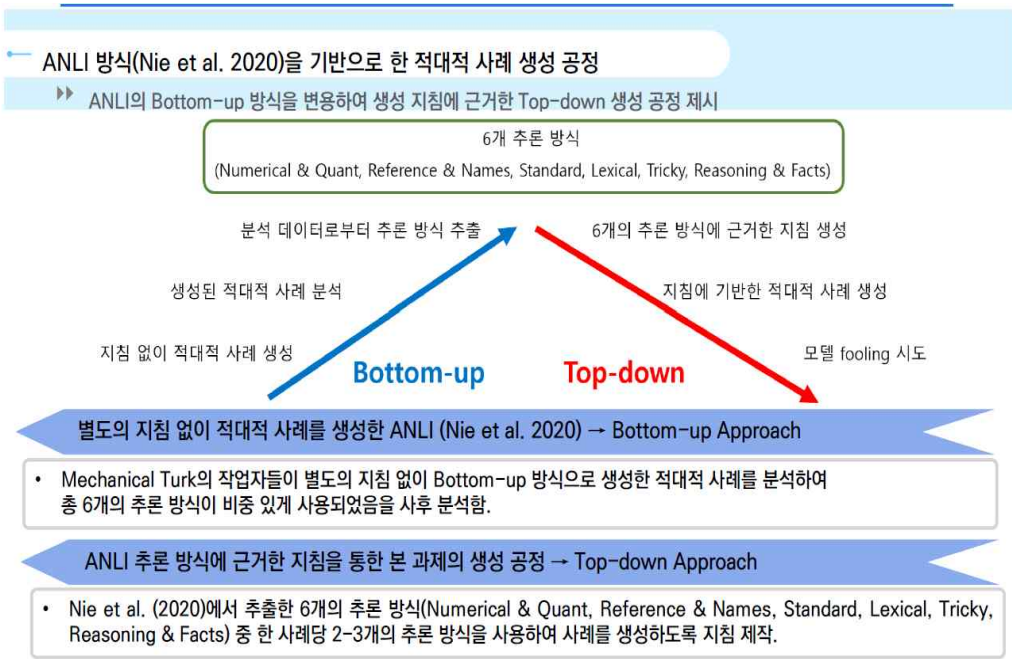
[그림 1] 과업수행 절차



주석자를 '화이트해커'로 간주하고 모델이 취약한 적대적 자연어추론 사례를 만드는 역할을 부여한다. 라운드(round) 1-3에서 주석자는 가설을 제작하고 BERT, RoBERTa 모형이 정답 라벨을 반복적으로 예측하게 한다. 신경망 모형의 예측이 틀리면 주석자의 임무는 완수되고, round를 거듭할수록 더 다양한 장르의 말뭉치에서 보다 적대적인 사례를 제작한다.

[그림 2] 적대적 사례 생성 공정

[그림 3] 적대적 함의 분석 생성 공정(top-down)



[그림 4] 적대적 사례 생성 예시

문맥	가설	label	설명	annotation	1	2	모델 예측
국립경주박물관은 1981년 실시한 금척리 고분 18기 발굴조사 결과를 최근 40년 만에 보고서('경주 금척리 신라묘')로 펴냈다.	국립경주박물관이 1981년 실시한 금척리 고분 18기 발굴조사 결과를 2020년대에 와서 보고서로 펴냈다.	E	근 40년이 지났다고 했으므로 2020년대에 펴냈음을 추론할 수 있다.	Numerical & Quant, Reasoning & Facts	E	E	C

원 작업자(가설 생성자) 1인과 타 작업자(1,2) 2인의 의견(함의 관계)이 모두 일치하는 경우에만 적절한 사례인 것으로 인정하고 수합한다.

3. 주석 지침

3.1 가설 문장 생성 원칙

★ 가설 문장 생성은 다음의 요건을 따른다.

① 가설 문장은 대상 담화에 작업자의 추론을 적용하여 각색/수정한 문장이어야 한다.
② 문맥 정보를 최대한 활용한다.

(10) [담화] 아빠의 가방에는 엄마의 시계가 들어 있었다. 이탈리아에서 가져온 시계. 엄마가 이탈리아에서 일할 때 돈 많은 백작이 마지막 선물로 준 것이라고 했다.

[가설 1 단계] **엄마와 아빠는** 이탈리아에서 일할 때 돈 많은 백작이 마지막 선물로 준 것이라고 했다.

--> 표준(standard) 적용: 접속

[가설 2 단계] 엄마와 아빠는 이탈리아에서 일할 때 돈 많은 백작에게서 마지막 선물로 **시계를 받았다.**

--> 속임수(tricky) 적용: 구문 변환

(11) [담화] 서울시는 거리두기 해제를 앞두고 **오후 11시부터 오전 6시까지 운행하는 '올빼미버스'를 대폭 확대 운영한다**고 밝혔지만 지하철과 택시는 운영 확대가 쉽지 않다.

[가설 1단계] **올빼미버스는** 오후 11시부터 오전 6시까지 운행한다.

--> 속임수(tricky) 적용: 구문 변환

[가설 2단계] 올빼미버스는 **7시간 동안** 운영된다.

--> 추론 및 사실(reasoning&fact) 적용: 세상에 대한 지식-사칙연산

★ 가설 문장은 다음을 참고하여 생성한다(한지윤 2019, 2020, 2021).

① 관계 주석별 언어 현상

가) 함의 관계

어휘적

- 동의관계, 상하의 관계, 부분관계, 일반 함의 관계

통사적

- 어순 뒤바꾸기, 능동/피동, 대립어 교체 구문, 처소 논항 교체 구문, 장/단형 사동 구문, 격 교체, 수식, 목록, 관계절

논증

- 시간, 공간, 수량

나) 모순 관계

- 대립 이의 관계, 상보 반의 관계, 관계반의어 중 역의관계, 양립 불가능 관계

다) 중립

- 함의 관계 및 모순 관계가 성립하지 않는 경우

② 세계 지식

시간 추론, 공간 추론, 양적 추론, 일반 상식

아래는 사례별 예시이다.

1. 유의어

(12) [담화] 지난 5월 이기훈 하나금융투자 연구원은 빅히트엔터테인먼트의 예상 기업가치로 3조 9000억~5조2000억원을 제시했다. 무엇보다 BTS의 인기가 최고조인데다 실제 **가파른 실적 성장**을 구가하고 있다는 점이 강점이다.

[가설] BTS의 인기가 최고조인데다 실제 실적도 **가파른 성장 가도**를 달리고 있다는 점이 강점이다.

[함의] E

2. 반의어

(13) [담화] 신종 코로나바이러스 감염증(코로나19) 장기화로 성수기임에도 공실을 채워야 하는 호텔업계가 비대면 마케팅에 초점을 맞춰 라이브 커머스에서 할인 판매에

나선 모습이다. 6일 호텔업계에 따르면 롯데호텔 월드는 오는 13일 라이브 커머스 '잼라이브'를 통해 '호캉스(호텔+바캉스)' 상품 **판매에 나선다**.

[가설] 롯데호텔 월드는 '잼라이브'를 통해 '호캉스(호텔+바캉스)' 상품 **구입에 나선다**.

[함의] C

3. 부분관계(양화사)

(14) [담화] 아울러 WCIF는 음악·엔터테인먼트 산업에서 동·서양의 교류와 협력에 기여한 인물에게 수여하는 'WCIF Award'를 제정해 1회 수상자로 K-POP 해외진출의 선구자인 보아(BOA)를 선정했다. 이번 행사는 WCIF 유튜브 채널을 통해 28일 **오전 10시부터** 생중계되며, 한-영 동시통역 서비스가 제공된다.

[가설] 이번 행사는 WCIF 유튜브 채널을 통해 28일 **하루종일** 생중계되며, 한-영 동시통역 서비스가 제공된다.

[함의] C

4. 핵심논항(처소 논항 교체 구문)

압구정 본점에서는 19일까지 지하 식품관에서'와 가설 문장의 '19일부터 압구정 본점은 지하 식품관을 통해'는 보조사 변경과 처소 논항 삭제의 모순에 해당한다. 추가 전략으로 어순 변경도 사용되었다.

(15) [담화] 이탈리안 레스토랑 살바토레쿠오모, 베트남 식당 타마린드, 한정식집 화니, 떡갈비 전문점 덕인관 등에서 식사 뒤 신세계 제휴 카드로 결제하면 30% 할인이 적용된다. 현대백화점 **압구정 본점에서는 19일까지 지하 식품관에서** 서울에서 가장 오래된 빵집인 '태극당' 팝업 매장을 운영한다.

[가설] **19일부터 현대백화점 압구정 본점은 지하 식품관을 통해** 서울에서 가장 오래된 빵집인 '태극당'의 팝업 매장을 운영한다.

[함의] C

'진주지원에서'과 가설 문장 '진주지원만을'은 핵심 논항 교체와 부분/전체를 복합적으로 나타내는 보조사를 사용한 구문으로 중립에 해당한다.

(16) [담화] 또 경찰이 사천시청 집무실 등을 압수수색할 때 집에 있던 돈을 아내와

측근인사 등을 통해 은닉하도록 하고, 2016년 11월 건설업자 A씨등 2명으로부터 의류 1000만원 어치와 상품권 300만원을 받은 혐의(부정청탁금지법 위반)로 불구속 기소됐다. 송 시장에 대한 선고공판은 오는 5월 28일 창원지법 진주지원에서 진행될 예정이다.

[가설] 송 시장에 대한 선고공판은 오는 5월 28일로 창원지법 진주지원만을 예정지로 보고 있다.

[함의] N

5. 능동/피동

(17) [담화] 미국 CNN방송은 박 시장이 시민운동가로 활동하던 2011년 여당 후보를 누르고 처음 서울시장에 당선된 이후 대선 후보로까지 부상한 정치 경력을 다뤘다. 뉴욕타임스(NYT)는 박 시장이 한국 최초의 성희롱 사건에서 승소한 인권변호사 출신이라는 점에 주목하면서 최근 몇년 동안 '미투 운동'이 한국 사회를 **강타한 사실**도 함께 전했다

[가설] 뉴욕타임스(NYT)는 한국사회가 미투 운동에 강한 **타격을 받은 사실**도 함께 전했다.

[함의] E

(18) [담화] 특히 농어촌 민박 통합 홈페이지 구축을 지원해 자체 통합 예약·결제 시스템을 구축해 수수료 등 경영부담을 줄이고, 무신고 숙박시설의 참여를 제한해 안전한 관광 환경을 조성할 방침이다. 정부는 향후 실증특례 운영 실적과 신사업이 농촌 경제·사회에 미치는 영향 등을 **종합적으로 고려해** 법·제도 정비를 검토하기로 했다.

[가설] 법·제도 정비는 향후 실증특례 운영 실적과 신사업이 농촌 경제·사회에 미치는 영향 등이 **종합적으로 고려될** 것이다.

[함의] E

6. 주제-논평구조

(19) 초점의 위치가 다른 경우

[담화] 과천시는 지난 24일 과천래미안센트럴스위트 단지 카페에서 마을문제 해결을 위한 주민 간 소통행사 '소통 대화마당'을 개최했다. 이번 행사는 과천시가 마을 내에 있는 문제와 갈등에 대해 주민들이 스스로 대화와 토론을 통해 해결할 수 있도록 하기 위한 취지로 기획된 것이다.

[가설] 과천시는 이번 행사를 통해 같은 마을에서 사는 주민들이 기관의 도움 없이 문제를 해결하도록 하고자 했다.

[함의] E

(20) 분열문

[담화] 평소와 다른 보호자들의 행동에 당황하는 봉식이와 단호한 행동으로 봉식이를 대하는 보호자들의 모습이 몰입을 배가시켰다. 봉식이를 자식처럼 여겨온 만큼 봉식이의 잘못된 행동을 고치고자 하는 보호자 부부의 의지가 더욱 불타올랐다.

[가설] 보호자 부부의 의지가 불타오른 것은 봉식이를 자신이 직접 낳은 자식처럼 여기기 때문이다.

[함의] E

7. 어순 뒤섞기(scrambling)

(21) [담화] 이밖에 이단 신천지, 기독교복음선교회(구 JMS)등과 관련된 추가 확진자는 현재까지 없는 것으로 알려졌다. 7명의 확진자가 나온 아산시 역시 최근 4일간 추가 확진자가 나오지 않으면서 한숨 돌리는 분위기다.

[가설] 새 확진자 수가 사일 동안 늘지 않으면서 일곱 명의 확진자가 발생했던 아산시도 한숨 돌리는 분위기이다.

[함의] E

8. 관계절

(22) [담화] 그 결과 간암을 일으키는 발암물질인 '아플라톡신'에 노출하면 시토신(C) 염기가 티민(T) 염기로 치환되지만, 감마선에 노출되면 티민(T)이 아데닌(A)이나 시토신(C)으로 치환되는 것으로 나타났다. 또 같은 발암물질에 노출되더라도 DNA 복구 기능에 결함이 있는 꼬마선충은 정상인 대조군에 비해 돌연변이 시그니처의 발생이 급격히 증가했다.

[가설] 아플라톡신은 간에 암을 유발할 수도 있다고 알려져 있는 물질 중 하나이다.

[함의] E

9. 세계 지식

(23) [담화] 이 동영상이 지난달 23일 공개되자 경찰은 하루 만에 용의자의 신원을 파악했으나, 아직 구체적인 혐의가 확인되지 않아 추가 수사를 하고 있다. 지난주에는 **밴쿠버 차이나타운**의 중국 문화센터에서 복면을 한 백인 남자가 '혐오스러운 낙서'로 창문을 훼손해 경찰이 수사에 나섰다.

[가설] **캐나다에 있는** 중국 문화센터에서 백인 남자에 의한 혐오 범죄가 있었다.

[함의] E

10. 일반 상식

(24) [담화] 조 의원은 이런 이유로, 현재 구성된 특위로 이번 문제를 다루는 것이 부적합하다고 지적했다. 9명의 특위 의원 가운데 조 의원을 제외한 8명이 모두 남성 의원이고 그 중 다수가 성인지 감수성이 부족하다는 주장이다.

[가설] 조 의원은 여성으로 특위 의원에 포함되어 있다.

[함의] E

3.2 가설 문장 기술 원칙

★ 가설 문장 기술 시 다음의 원칙들을 따른다.

1) 오류

▮ 맞춤법 오류, 기호 사용 오류 등은 임의로 수정하지 않고 부적절 처리한다.

(25) 선행 문장: 동 주민센터 직원, 대학생 아르바이트생을 비롯해,

대상 문장: 희망일자리 사업 참여 공공근로 800여명, 노인일자리 사업 참여 어르신 449명 등 추가 인력을 활용하여 주요 지역 및 시설 등을 지속적으로 소독하고 방역수칙을 홍보하는 등 주민 불안감 해소와 안전 보호에 힘쓴다.

(26) 선행 문장: 시는 사회적 거리두기 실천으로 가정에서 갓김치의 주문이 평년에 비해 1.

대상 문장: 5배 이상 증가했으며, 이로 인한 생 갓 품귀현상으로 높은 가격이 형성되고 '봄 갓' 재배 면적 확대와 생산량 증가로 이어진 것으로 분석했다.

2) 복원

▮ 대상 담화에 주어나 목적어가 없을 경우 가설 문장에서는 문맥상 예측할 수 있는 문장 요소를 복원하여 제시한다.

(27) [담화] 이 마스크는 에탄올 살균 세척 실험을 진행한 결과 20회 반복 세척 후에도 초기 여과 효율 94% 이상을 유지하고 나노섬유 멤브레인의 구조 변화가 전혀 일어나지 않는 것이 관찰을 통해 확인했다. 특히 에탄올에 3시간 이상 담가도 나노섬유가 녹거나 멤브레인의 뒤틀림 현상이 없어 에탄올을 이용해 살균·세척할 경우 한 달 이상 사용이 가능한 장점이 있다.

[가설] 에탄올은 **이 마스크를** 한 달 이상 사용 가능하게 만든다.

[함의] E

3) 문맥적 정보 우선

▮ 사전적 정보와 문맥적 정보가 상충할 때, 문맥적 정보를 우선한다.

(28) [담화] 우한 체류 교민들은 이날 오전 5시(현지시간)께 출발해 오전 8시께 김포공항에 도착했다. 귀국을 희망한 **교민과 유학생** 720명 중 369명이 우한 공항에 모였지만 **1명**은 중국 당국의 사전 검역에서 우한 폐렴 의심 증세가 확인돼 비행기에 오르지 못하고 귀가 조치됐다.

[가설] **교민** 중 1명은 중국 당국의 사전 검역에서 우한 폐렴 의심 증세가 확인돼 비행기에 오르지 못하고 귀가 조치됐다.

[함의] N

☞ 교민의 사전적 정의에 따르면 교민은 '다른 나라에 살고 있는 동포. 아예 정착하여 살고 있는 교포나 일시적으로 머무르는 유학생, 주재원 등'을 지칭하나 주어진 문맥에서는 '교민과 유학생'으로 '교민'의 범위를 교포로 한정하여 교민과 유학생의 의미가 구분되어 해석된다.

▮ 작업자의 상식적 정보와 문맥적 정보가 상충할 때, 문맥적 정보를 우선한다.

(29) [담화] 특히 **탄산음료**는 편의점 배달 수요가 10배 이상 증가했던 지난 5월 배달로 가장 많이 판매된 음료 10개 중 3개를 차지했다. 작년 5월에는 커피를 제외하면 미네랄워터, 제주삼다수, 옥수수 수염차 등 생수와 차음료가 매출 상위권을 차지했지만 올해는 **밀키스, 마운틴듀, 코라콜라제로 등**이 생수를 밀어내고 매출 톱10에 이름을 올렸다.

[가설] 올해 5월 생수를 밀어내고 매출 톱10에 이름을 올린 음료 중 밀키스, 마운틴듀, 코카콜라제로는 탄산음료이다.

[함의] E

☞ 작업자가 밀키스, 마운틴듀, 코카콜라제로가 탄산음료가 아니라고 알고 있었다 하더라도 문맥 정보상 탄산음료라 해석되고 있으므로 이를 따른다.

4) 정보구조

▮ 선후행 문맥을 기반으로 한 정보구조를 통해 동일하게 적용될 수 있는 화제인지를 파악한다.

(30) [담화] 향후 정부 방침에 따라 '생활 속 거리두기'로의 전환 시 면밀한 검토를 거쳐 다음달 6일 이후 **도덕산캠핑장, 광명국민체육센터 등 사업장**도 추가로 재개장 할 계획이다. 휴장 기간 동안 공사는 종합적인 시설 점검을 실시해 ▷**광명골프연습장** 그물망 보수 공사와 타석 인조잔디 교체공사 ▷광명국민체육센터 창문 가림막 공사, 실내 바닥 샌딩 공사, 안내데스크 개선 공사 ▷**도덕산캠핑장** 생태연못 울타리 보수 공사, 데크 수선 공사 ▷**시립족구장** 가압 펌프 역류 방지기와 배관 누수 수리 공사 등 전반적인 시설 개보수 공사를 진행하며 재개장 시 고객 이용에 불편함이 없도록 만전을 기했다.

[가설] 광명골프연습장, 광명국민체육센터, 도덕산캠핑장, 시립족구장은 다음달 6일까지 휴장한다.

[함의] E

☞ 문맥상 도덕산캠핑장, 광명국민체육센터 등에 광명골프연습장, 시립족구장이 포함된다.

5) 중립 판단 기준

▮ 가설은 모두 문맥적 정보를 통해 추출되어야 한다. 따라서 실제 세계에서는 사실일지 모르나 문맥 정보 상에서 사실 여부를 판단할 수 없는 정보는 중립으로 처리한다.

(31) [담화] 동국제강은 탄력적 조업이 가능한 전기로 사업의 장점을 극대화하는 데 주력했다. 여기에 더해 차별화된 봉형강과 컬러 강판 신제품 등 고부가가치 제품과

설루션 마케팅 도입 등 시장을 선도하는 초격차 전략을 지속하며 수익성 확대에 **매진했다**.

[가설] 동국제강의 봉형강과 컬러강판 신제품은 **매진되었다**.

[함의] N

☞ 실제 세계에서 동국제강의 봉형강과 컬러강판 신제품이 매진되었더라도 담화 문맥에서 사실 여부를 확인할 수 없으므로 중립으로 처리한다.

6) 민감사항

▮ 성(gender), 정치, 종교와 관련한 민감 사례를 담았거나, 실명이 언급되어 논란의 여지가 있을 수 있는 담화들은 부적절 처리한다.

(32) [담화] 이 남성은 “그냥 개XX XXX 욕하는 거면 모른다. 애들도 있는 동네에서 남성의 성기, 여성의 성기 들먹이며 떠드는데 교육적으로 좋겠느냐”며 “저런 XX 하나 있는 것만으로도 충분한데 왜 동네를 쓰레기로 만드느냐”고 토로했다.

3.3 추론 적용 원칙

★ 추론 적용 시 다음의 원칙들을 따른다.

1) numerical & quant(수와 양)

▮ 대상 문장에서 숫자와 관련한 단어를 변환하여 숫자/날짜/연령의 해석을 속임(fooling)

(33) [담화] 이 지사는 이날 선포를 통해 건축계와 사회전반에 스며들어 있는 표절에 대해 경종을 울리고 앞으로 상생을 위한 비전을 제시하면서 지역사회의 좋은 평가를 이끌었다. 한편 경주타워와 관련한 저작권 소송은 **지난 2004년 디자인 공모를 통해 2007년 완공된** 경주타워의 모습이 공모전에 출품한 유 선생의 디자인과 흡사하다는 점이 지적되면서 같은 해 연말부터 시작됐다.

[가설] 경주타워는 디자인 공모를 한 지 **3년 만**에 완공되었다.

[추론] numerical&quant, reasoning&fact
[함의] E

(34) [담화] 지난 **4월 한달 동안 500여건**이 접수된 것을 시작으로 **지난 5월에는 1000여건**이 접수될 정도로 많은 소비자들이 잎새레터에 참여하며 관심이 높아지고 있다. 부모님과 연관된 추억에서부터 잎새주 모델인 송가인 팬들의 뜨거운 관심까지 다양한 사연이 접수됐다.
[가설] 지난 4월과 5월에 **총 1500여건**이 접수되었다.
[추론] numerical&quant, reasoning&fact
[함의] E

(35) [담화] 이에 대해 재판부는 이날 "검찰 측이 항소를 했다하더라도 동일한 형이 선고됐을 것"이라고 말했다. A씨 부부는 지난해 5월26일부터 **닷새 동안**, 인천 부평구에 있는 한 아파트에 **생후 7개월 된 딸**을 반려견 두 마리와 함께 방치해 숨지게 한 혐의로 재판에 넘겨졌다.
[가설] A씨 부부는 **7개월간** 딸을 방치하였다.
[추론] numerical&quant, tricky
[함의] C

▮ 문맥에서 제시되는 집합-부분집합-여집합 간의 관계를 속임(fooling)

(36) [담화] 이날 거소투표에는 시설 거주 이용인 **52명 중 40명이 참여**했고 경기도 선관위 관계자를 비롯한 투표 참관인 4명, 시설 사회복지사 7명이 함께 했다. 투표 과정을 안내하던 김인선 사회복지사(38)는 "시설 거주 이용인들이 무사히 투표권 행사를 마쳐 기쁘다"고 소감을 밝혔다.
[가설] 거소투표에는 시설 거주 이용인 **과반수 이상**이 참여했다.
[추론] numerical&quant, standard, lexical
[함의] E

2) reference & name(지시와 이름)

▮ 상호참조하는 일반명사, 고유명사 혹은 (대)명사 간의 관계를 속임(fooling)

(37) [담화] 공포에 질린 그는 "나으리, 경비들 때문에 나갈 수나 있겠습니까?"라며 걱정스러운 마음을 여실히 드러내지만 성이겸은 의미심장한 눈길로 어딘가를 지켜보고 있어 무언가

속셈이 있는 것인지 이목이 쏠린다. 성이겸과 박춘삼은 시간이 되어 노역꾼들 틈에 섞여 집결하고, **통솔자**가 나타나 살벌한 말투로 명령을 내리고 있어 눈길을 끈다.

[가설] **성이겸과 박춘삼**은 살벌한 말투로 명령을 했다.
[추론] reference&name, standard
[함의] C

(38) [담화] 5개 대학은 경희대학교(용인), 성균관대학교(수원), 연세대학교(인천), 한양대학교(안산), 한국외국어대학교(용인)다. 2012년부터 시작한 '삼성드림클래스'는 **교육 여건이 부족한 지역의 중학생**에게 대학생이 멘토가 돼 학습을 지원하는 삼성전자의 대표적인 교육 사회공헌 프로그램이다.

[가설] 삼성드림클래스는 **중학생이라면 누구나** 참여할 수 있다.
[추론] reference&names, tricky
[함의] C

▮ 상호참조하는 일반명사, 고유명사 혹은 (대)명사를 복원하여 속임(fooling)

(39) [담화] 미국 경제전문 매체 CNBC는 21일(현지시간) 다수의 행동 심리학자들을 인용, 긍정적 결과를 더욱 강조하는 성향이 있다면 다른 사람들보다 감염 위험에 노출될 가능성이 더 높다고 보도했다. 이른바 '낙관주의 편향성'이 강하다면 실제 위험에 대해 잘 알고 있음에도 불구하고 **본인**이 처한 위험을 축소하는 경향이 강하다는 설명이다.

[가설] **낙관주의 편향성이 강한 사람**은 실제로 **자신**이 어떤 위기에 처해 있는지 알지 못한다.
[추론] reference&names, standard
[함의] C

3) standard(표준)

▮ 인과적 논리 관계를 속임(fooling)

(40) [담화] 청와대 국민청원 게시판에는 10일 '을왕리 음주운전 역주행으로 참변을 당한 50대 가장의 딸입니다'라는 제목의 글이 올라왔다. 전날 새벽 술을 마시고 역주행 하는 차량에 치여 숨진 A(54)씨의 딸이라고 자신을 소개한 청원인은 당시 가해자들이 사고 현장에서 **119보다 변호사를 먼저 찾았다**는 목격담을 확인했다고 주장했다.

[가설] 10일 을왕리 음주운전 역주행 사건 가해자들은 **119를 부르고 그다음 변호사를 찾았다.**

[추론] standard, tricky

[함의] C

(41) [담화] 오는 2022년까지 추진되는 이 사업은 국비 235억원, 시비 215억원 등 총 450억원이 투입된다. 건조사는 **외부 평가위원회의 심의를 거쳐 ㈜현대미포조선이 선정**됐다.

[가설] **현대미포조선이 외부 평가위원회를 심의했다.**

[추론] standard

[함의] C

(42) [담화] 하정우는 **본인의 휴대전화를 해킹한 협박범**과 대화로 시간을 벌며 경찰의 수사를 도왔지만 끝내 직접 대화를 한 닉네임 '고호'는 중국으로 도피한 상태다. 승재현 형사정책 연구원 연구위원은 휴대전화 해킹 방지책으로 세가지 지침을 꼭 지켜야 한다고 강조했다.

[가설] **하정우는 휴대전화를 해킹했다.**

[추론] reference&name, tricky

[함의] C

▮ 부정 관계를 속임(fooling)

(43) [담화] 이 조치는 고국으로 돌아오는 유럽 시민에게는 적용되지 않으며, 장기 EU 거주자, EU 회원국 국민의 가족, 외교관, 의사, 코로나19 확산 방지를 위해 일하는 연구자, 상품 운송 인력 등도 면제 대상이라고 폰데어라이엔 위원장은 덧붙였다. 이번 금지 조치는 **EU 27개 회원국 가운데 아일랜드를 제외한 26개국**과 노르웨이, 스위스, 아이슬란드, 리히텐슈타인 등 셍겐 협정에 가입된 4개 EU 비회원국 등 30개 국가를 아우르게 될 것으로 예상된다.

[가설] **EU 회원국에 아일랜드는 포함되어 있지 않다.**

[추론] numerical&quant, reasoning&facts

[함의] C

(44) [담화] 앞서 25일에는 정책 추진을 일단 보류하고 코로나19 안정화 뒤 협의

체에서 논의를 진행하자는 정부의 제안을 대한의사협회가 수용하지 않자 정부는 전공의에 대한 업무개시명령을 **수도권에서 비수도권으로까지 확대함**과 동시에 현장에 복귀하지 않은 10명을경찰에 고발했다. 이에 의협은 복지부 간부를 맞고발하고 오는 7일부터 무기한 총파업을 예고하는 등 정부와 의료계 갈등은 이어지고 있다.

[가설] 정부의 전공의에 대한 업무개시명령은 **수도권이 아닌 지역은 해당되지 않는다.**

[추론] standard, lexical

[함의] C

▮ 비교 관계를 속임(fooling)

(45) [담화] 완도군 공동관에서는 활전복과 활광어, 해초 샐러드, 해초 돈가스, 해초 국수, 전복장, 전복 절편 등 다양한 제품을 국내외 바이어들과 소비자들에게 선보였다. 공동관에 참가한 업체는 **누리영어조합법인(전복), 완도사랑S&F(전복), 해성인터내셔널(광어), 해청정(해조류), 세계로수산(해조류), 하나물산(해조류), 바다향기(해조류), 완도친환경협동조합(해조류)**이다.

[가설] 공동관에 참가한 업체 중 **해조류를 판매하는 업체보다 전복을 판매하는 업체가 더 많다.**

[추론] numerical&quant, standard, reasoning&facts

[함의] C

(46) [담화] 정부는 애초 긴급재난지원금 예산을 9조 7000억원(**2차 추경 7조 6000억원+지방정부 분담금 2조 1000억원**)으로 잡았지만, 민주당 입장대로 지급 대상을 전 국민으로 확대하면 예산 규모는 13조원으로 늘어난다. 민주당은 추가로 소요되는 재원 3조~4조원을 지출조정과 국채발행 등을 통해 확보할 수 있다는 구상이다.

[가설] 긴급재난지원금 예산 중 **2차 추경에 해당하는 금액이 지방정부 분담금보다 많지 않다**.

[추론] numerical&quant, standard

[함의] C

3) lexical(어휘)

▮ 동의어 혹은 반의어 관계 간의 어휘를 수정하여 속임(fooling)

(47) [담화] 2020 스포츠서울 라이프특집 혁신한국인&파워코리아에 선정된 수성키즈스트레칭전문학원에서는 유아동을 위한 스트레칭/리듬체조 수업과 함께 성인, 가족 단위 수강생을 대상으로 스트레칭 수업을 진행하는데 호응이 뜨겁다. 고 원장은 매년 두 차례 열리는 비선수 리듬체조 대회가 **무기한 연기되자** 수강생들의 맨손 체조 작품을 체조와 무용 특성에 맞게 창작하여 **3주간 훈련시켰다.**

[가설] 비선수 리듬체조 대회가 **3주간 연기되었다.**

[추론] lexical

[함의] C

(48) [담화] 올해 우편투표가 4년 전보다 폭증하고 총투표도 2000만여 표 늘 전망이지만, 특별한 문제가 없다면 예년처럼 3일 밤 11시(한국 시각 4일 오후 1시)쯤 윤곽이 나오는 게 가능하다고 미 언론들은 전망한다. **최대 경합주인 플로리다**와 노스캐롤라이나, 애리조나 등이 우편투표를 선거 전부터 개표해 당일 밤 최종 결과를 알리겠다고 공언해 왔기 때문에 투표일 밤에 승자 윤곽을 점칠 수도 있다.

[가설] **플로리다는 최소 경합주**에 해당하는 주이다.

[추론] reference&names. lexical

[함의] C

(49) [담화] 다만 테슬라가 최근 계속되는 품질 이슈와 푸조·르노 등 대중 수입 전기차들의 등장, 지자체별 보조금 상황 등은 발목 잡히는 주요 요인으로 작용할 수 있다. 특히 도장 품질이나 차체 패널 단차(어긋남) 등 관련 문제가 지속 제기되는 가운데 **국내 공식 서비스센터는 단 2곳(서울 강서, 성남 분당)**에 불과한 만큼 서비스 지연 불만도 계속 불거지는 상태다.

[가설] 테슬러의 국내 공식 서비스텐터는 **서울시에 두 곳**이 있다.

[추론] lexical, reasoning&facts

[함의] C

▮ 연접(conjunction) 관계에 있는 어휘를 수정하여 속임(fooling)

(50) [담화] **삼성전자와 LG전자가 각각 주력으로 하는 QLED TV와 유기발광다이오드(OLED) TV**의 명암도 엇갈렸다. QLED 판매량은 지난해 4분기 252만대에서 154만대로, OLED 판매량은 111만대에서 62만대로 일제히 감소했고 판매량 격차는 2.3배 수준이었다.

[가설] **삼성전자는 유기발광다이오드 TV에 초점을 맞추고 LG전자는 QLED TV에 초점을 맞춘다.**

[추론] lexical, tricky
[함의] C

5) tricky(속임수)

▮ 구문변환이나 재정렬을 통한 속임(fooling)

(51) [담화] 자신의 시즌 세 번째 대회에서 공동 2위에 올라 **CME 그룹 투어 챔피언십 출전권을 손에 쥔 고진영**은 이번 대회 우승으로 단숨에 상금왕 고지까지 올랐다. 마지막 2개 대회의 우승 상금 규모가 US여자오픈 100만 달러, CME 그룹 투어 챔피언십 110만 달러로 올해 LPGA 투어 대회 가운데 가장 컸고, 고진영은 그 2개 대회에서 우승, 준우승을 연달아서 하며 상금왕에 오르는 원동력으로 삼았다.

[가설] 고진영은 **CME 그룹 투어 챔피언십에서 우승하였다.**
[추론] tricky, reasoning&facts
[함의] C

(52) [담화] **광주시**는 19일 시청 1층 시민홀에서 '광주형 AI-그린뉴딜 2차 시민보고회'를 열고 **'2045 탄소중립 에너지 자립도시 실현'**을 위한 3대 전략과 8대 핵심과제를 발표했다. 광주시는 지난달 21일 **국내 최초로 2045년까지 탄소중립 에너지 자립도시를 만들겠다**는 광주형 AI-그린뉴딜의 목표를 제시하고 전력부문을 100% 신재생 에너지로 전환하는 1차 로드맵을 발표한 바 있다.

[가설] **국내 최초의 탄소중립 에너지 자립도시는 광주이다.**
[추론] lexical, tricky
[함의] C

(53) [담화] 앞서 박지윤, 최동석 아나운서 가족은 음주 상태로 고속도로에서 역주행하던 화물차에 들이 받히는 사고를 당했다. 부산경찰청에 따르면 전날 오후 8시30분쯤 부산 금정구 선두구동 경부고속도로에서 **만취 운전자 A씨(49-남)가 몰던 2.5t 화물차가 반대 차선으로 역주행**해 **마주 오던 볼보 승용차**와 충돌했다.

[가설] **볼보 승용차가 역주행**을 하였다.
[추론] reference&names, tricky
[함의] C

6) reasoning & fact(추론 및 사실)

▮ 세상의 지식 및 상식을 사용하여 속임(fooling)

(54) [담화] 시는 지난 1일 나무가 고사한 현장에 현수막을 내걸고 **나무를 훼손한 현장 목격자**를 찾고 있다. 시는 공공 시설물 훼손과 산림자원법 위반 혐의로 경찰에 수사 의뢰할 예정이다.

[가설] 나무를 훼손한 사람이 있다.
[추론] reasoning&facts, tricky
[함의] E

(55) [담화] 지난 3일부터 4일 6시까지 도내 강우량은 예산이 218㎜로 가장 많고, 천안 212㎜, **아산 187㎜**, 홍성 132㎜ 등을 기록했다. 1일 최대 강우량은 아산 송악 273㎜, 천안 북면 267㎜, 예산읍이 217㎜ 등이며, 시간당 강우량은 아산읍 63㎜, 천안 성거읍 51㎜, 예산읍 34.5㎜ 등으로 나타났다.

[가설] 지난 3일부터 4일 6시까지 도내 강우량은 **아산이 3위**를 기록했다.
[추론] numerical&quant, reasoning&facts
[함의] E

(56) [담화] 이런 가운데 교내 점심 방송에서 **백호랑의 신경세포를 강타한 사연**이 등장, "사랑하는 친구야, 우리 중3때 생일 파티 진짜 재밌었는데"라는 사연 한 줄에 아연실색한 백호랑의 모습이 호기심을 돋운다. 백호랑이 이 익명의 사연과 어떤 관련이 있는 것인지, 사연자의 정체는 누구일지, 그리고 중3 생일 파티 때 어떤 일이 있었던 것일지 물음표가 더해지는 상황.

[가설] 백호랑은 익명의 사연에게 **신경세포를 강타당한 후에 고통을 호소**했다.
[추론] standard, reasoning&facts
[함의] C

3.4 설명 작성 원칙

★ 설명 작성 시 다음의 원칙들을 따른다.

1) 완전한 문장 형태로 기술

▮ 완전한 문장 형태가 아닌 종결 어미는 배제한다. '~하기 때문에 함의/중립/모순에 해당한다'와 같은 형식을 취한다. 추론이 중립일 경우 '~하기 때문에 추론할 수 없다/추론이 불가능하다'와 같은 형식도 가능하다.

2) 대상 담화에서 사용된 단어나 문장 구조를 근거로 가져올 수 있음

▮ 설명 부착 시에 모든 기술을 새롭게 할 필요는 없으며 대상 담화에서 근거가 되는 단어나 문장 구조를 그대로 따 와서 설명에 사용할 수 있다.

(57) [담화] 하지만 신청 첫날 최고 24만 명이 동시 접속하는 등 접속자 폭주로 휴대폰 인증에 필요한 인증서버가 다운되면서 인증이 중단되는 불편이 발생했다. 경기도는 동시접속자가 최고 20만 명 이상 운영이 가능하도록 누리집을 설계했지만 접속자가 틀리면서 인증 중돈 사태를 빚은 것으로 파악했다.

[가설] 경기도가 설계한 누리집을 통해서 신청 첫 번째 날부터 순조롭게 인증을 할 수 있었다.

[추론] standard, lexical

[함의] C

[설명] 신청 첫날 최고 24만명이 동시 접속하는 바람에 접속자 폭주로 인증 서버가 다운되었다고 하므로 모순에 해당하는 진술이다.

3) 사용한 추론 방식을 근거로 함의/모순/중립에 대한 설명 기술

▮ 아래와 같은 경우, 반의어를 사용한 추론 방식(lexical)을 채택하였다면 본래 대상 담화에서는 반의어와 상충되는 내용이 나온다는 사실을 근거로 기술할 수 있다.

(58) [담화] 물론 코로나19에 따른 상흔이 워낙 커 매출 규모를 작년과 직접 비교하기엔 무리가 있지만, 소비심리가 최근 1개월여간 점진적으로 개선되고 있는 점만 해도 의미가 있다는 것이 업계 중론이다. 이와 관련해 백화점 업계는 모처럼 발걸음에 나서는 수요층을 사로잡기 위해 다양한 봄철 정기 세일 행사에 나선다.

[가설] 소비심리가 최근 한 달간 급진적으로 개선되고 있는 점만 해도 의미가 없는 것이 아니다.

[추론] standard, lexical

[함의] C

[설명] 급진적으로 개선되는 것이 아니라 점진적으로 개선되는 것이므로 모순에 해당한다.

참고 문헌

한지윤. 2019. 언어 추론 모델 개발을 위한 말뭉치 구축 방법론 연구. 언어사실과 관점 48, 351-384.

한지윤. 2020. 한국어 추론 말뭉치 구축을 위한 기초 연구. 한글 81, 949-975.

한지윤. 2021. 한국어 추론 벤치마크 데이터 구축 방법론 연구. 연세대학교 박사논문.

Dagan, I. and Glickman, O. 2004. Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. In *Proceedings of the PASCAL Workshop on Learning Methods for Text Understanding and Mining*. Grenoble, France.

Glickman, O. 2006. Applied Textual Entailment. Ph.D. Thesis. BarIlan University.

Yixin Nie, Adina Williams, Emily, Dinan, Mohit Bansal, Jason Weston, Douwe Kiela(2020), Adversarial NLI: A New Benchmark for Natural Language Understanding, ACL 2020.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020). Adversarial NLI: A new benchmark for natural language understanding.In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4885-4901.

Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In Proceedings of the AAAI conference on artificial intelligence, 34(05), pages 8018-8025.

<기획·연구>

국립국어원 강미영 언어정보과장
국립국어원 유희정 학예연구사
국립국어원 이민주 연구원
국립국어원 박미은 연구원
국립국어원 정영은 연구원

<사업 참여자>

책임 연구원 김일환(성신여자대학교)
공동 연구원 강아름(충남대학교)
김태우(부산대학교)
박진호(서울대학교)
박현아(고려대학교)
송상헌(고려대학교)
송영숙(주)나라지식정보
이도길(고려대학교)
이지은(고려대학교)
장하연(부산외국어대학교)
정슬아(성신여자대학교)
정연주(홍익대학교)
조경찬(주)나라지식정보
최윤지(인하대학교)
연구 보조원 김도현(고려대학교)
홍승혜(고려대학교)
신운섭(고려대학교)
이규민(고려대학교)
노강산(고려대학교)

황동진(고려대학교)
양진아(충남대학교)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2023년 1월 19일
발행일: 2023년 1월 19일
인 쇄: 성신POD

※ 이 책은 국립국어원의 용역비로 수행한 '2022년 말뭉치 함의 분석 및 연구' 사업의 결과물을 발간한 것입니다.