한국어 자연어 추론을 위한 다양한 프롬프트 방법

최요한*○ 이창기* 배경만**
강원대학교* 한국전자통신연구원**
choiyohan@kangwon.ac.kr* leeck@kangwon.ac.kr* kyoungman.bae@etri.re.kr**

Various Prompt Methods for Korean Natural Language Inference

Yohan Choi*○ Changki Lee* Kyungman Bae**
Kangwon National University* Electronics and Telecommunications Research Institute**

요 약

자연어 추론은 전제 문장과 가설 문장의 관계를 함의, 중립, 모순으로 분류하는 자연어 처리 태스크이다. 최근 여러 자연어 처리 태스크에서 딥러닝 모델을 이용하는 방법이 우수한 성능을 보이고 있지만, 이는 미세 조정과정에 드는 비용이 많다는 점과 모델 출력의 근거, 과정을 사람이 이해하기 어려운 한계가 있다. 이러한 이유로 최근에는 소량의 입력, 출력 예시를 포함한 프롬프트를 이용한 방법론과 모델 출력에 대한 근거를 생성, 활용하는 방법에 관한 많은 연구가 진행되고 있다. 본 논문에서는 퓨샷 학습 환경의 한국어 자연어 추론 태스크를 위한 세 가지 프롬프트 방법과 이들을 조합하여 적용하는 방법을 제안한다. 이를 통해 '해석 가능성'과 자연어 추론 성능을 모두 향상시킬 수 있음을 보인다.

주제어: 자연어 추론, 프롬프트, 프롬프트 엔지니어링

1. 서론

자연어 추론(Natural Language Inference)은 전제 문장과 가설 문장의 관계를 함의(Entailment), 중립(Neutral), 모순(Contradiction)으로 분류하는 자연어 처리 작업으로, 전제 문장의 내용을 기준으로 가설 문장의 내용이 전제 문장과 합치하면 함의, 충돌하면 모순, 알 수 없는 경우 중립으로 분류한다. 예를 들어 "시인의 편지가 시대를 녹였다."라는 전제 문장과 "시대를 시인의 편지가 녹였다."라는 두 문장의 관계는 내용이 같기 때문에 '함의'로 분류한다.

최근 자연어 처리에서는 사전 학습된 언어 모델을 미세 조정(Fine-Tuning)하는 방법을 주로 사용하며 우수한 성능을 보이지만, 언어 모델의 크기가 커질수록 학습에 드는 비용이 많이 들며, 또한 모델의 출력의 근거를 사람이 이해할 수 없는 한계가 있다. 이러한 이유로 최근에는 소량의 입력 및 출력 예시를 포함한 프롬프트(Prompt)를 거대 언어 모델의 입력으로 사용하는 퓨샷 학습(Few-shot Learning) 방법론[1]과 모델 출력에 대한 근거를 생성 및 활용하여 딥러닝 모델의 '해석 가능성'을 향상시키는 방법[2,3]이 연구되었다.

퓨샷 학습 방법론은 언어 모델의 크기가 커질수록 성능도 미세 조정 방법론을 상회하는 반면, 복잡한 추론 능력이 요구되는 태스크에 대해서는 모델의 크기가 커지는 것이 도움이 되지 않는 경우가 많다는 문제가 있다. 이러한 문제를 해결하기 위해 프롬프트 내부에 입력부터 최종 출력을 도출하기까지의 중간 과정을 기술해주는 등 프롬프트의 구조를 다양하게 변형하는 방법 또한 연구되고 있다[4].

본 논문에서는 학습 데이터가 존재하지 않는 신규 도메인에 한국어 자연어 추론 태스크를 적용하기 위해 퓨샷 학습 방법을 적용하며, 성능 향상을 위해 전제 문장과 가설 문장으로부터 지식 기반의 설명문을 생성하여 활용하는 방법[5], 한국어 자연어 추론 태스크를 QA 태스크로의 변환하는 방법, 한국어 자연어 추론을 위한 Chain-of-Thought 방법 등의 세 가지 형식의 프롬프트와 이들을 조합하여 적용하는 방법을 제안한다.

2. 관련 연구

프롬프트 엔지니어링(Prompt Engineering)은 모델에 자연어로 기술된 프롬프트를 주어 더욱 정확한 출력을 이끌어내어 성능이 향상될 수 있도록 프롬프트를 구성하는 방법이다. [3]에서는 지식을 생성하는 프롬프트를 이용해 지식을 생성하여 상식 추론 태스크의 성능을 향상시킬 수 있음을 보여주었다. 본 논문에서는 자연어 추론을 위해 전제 문장과 가설 문장으로부터 지식을 이용한 설명문을 생성하여 전제 문장과 가설 문장에 추가하여 자연어 추론의 성능을 향상 시킨다.

[4]에서는 언어 모델의 크기를 확장하는 것만으로는 높은 성능을 달성하기 어려운 까다로운 작업에 대해 최종 출력을 도출해내기 위한 중간 과정을 기술하는 예시를 포함한 Chain-of-Thought 프롬프트가 산수, 기호 추론 등과 같은 태스크의 성능을 향상할 수 있음을 보여주었다. 본 논문에서는 이러한 Chain-of-Thought 프롬프트를 한국어 자연어 추론에 적용하여 성능을 향상 시킨다.

'설명 가능한 인공지능'은 기존 딥러닝 모델의 출력의 근거를 사람이 이해하기 어려운 한계를 극복하기 위해 근거를 생성 및 활용하여 모델에 대한 '해석 가능성'을 높이기 위한 방법론이다. [2]에서는 자연어 추론 태스크를 중심으로 딥러닝 모델의 '해석 가능성'을 위한 자연어 추론 설명 생성기를 제안하였고, [3, 5]에서는 프롬프트를 이용해 상식 추론, 자연어 추론에 대한 지식을 생성하여 추론 성능을 향상시킬 수 있음을 보여

주었다.

[6]에서는 거대 언어 모델로도 성능 향상이 더딘 복잡한 태스크인 상호참조해결을 비교적 간단한 QA 태스크로 변환하는 프롬프트를 이용하여 비교적 나은 추론을 할 수 있음을 보여준다. 본 논문에서는 한국어 자연어 추론 태스크를 QA 태스크로의 변환하는 프롬프트를 이용하여 자연어 추론의 성능을 향상시킨다.

3. 한국어 자연어 추론을 위한 다양한 프롬프팅 방법

본 논문에서는 퓨샷 환경에서의 한국어 자연어 추론을 위한 세 가지 프롬프팅 방법과 이를 조합하여 적용하는 방법을 제안한다. 자연어 추론 태스크를 위한 세 가지 프롬프트의 예시는 표 1과 같다.

표 1. 한국어 자연어 추론을 위한 프롬프트 구성

지식 기반 설명 생성
Premise: 머무는 6일 동안 불편함 없이 잘 지냈습니다. Hypothesis: 6일 동안 편히 머물렀습니다. Knowledge: 편함과 불편함은 의미가 대립되는 반의어이므로 불편함이 없다는 것은 편하다고 이해할 수 있다. ... more example Premise: {premise} Hypothesis: {hypothesis} Knowledge:
QA 변환
다음의 예시와 같이 의문형으로 변환 가설: 인교진 씨는 무대 뒤편을 바라보았다. 변환: 인교진 씨는 무대 뒤편을 바라보았습니까? ... more example 가설: {hypothesis} 변환:
CoT
Premise: 이후에도 인교진 씨는 울컥할 때마다 무대 뒤편을 바라보며 눈물을 삼켰다. Hypothesis: 인교진 씨는 울컥하지 않았다. 순서대로 위 문장의 관계를 유추한다. 1. Premise에서 인교진 씨는 울컥한다는 정보가 포함된다. 2. 반면, Hypothesis는 인교진 씨가 울컥하지 않았다고 진술한다. 3. Premise와 Hypothesis의 진술이 엇갈린다. 따라서 정답은 contradiction이다. ... more example Premise: {premise} Hypothesis: {hypothesis} 순서대로 위 문장 쌍의 관계를 [entailment, neutral, contradiction]으로 추론한다.

지식 기반 설명 생성 지식 기반 설명 생성은 표 1과 같이 자연어 추론의 전제 문장과 가설 문장, 지식 기반의 설명문 예시들을 포함한 프롬프트를 이용해 새로운 전제 문장과 가설 문장에 대한 지식 기반 설명문을 생성한다. 자연어 추론 실험을 위해서 생성한 지식 기반 설명문의 개수는 5개이며, 추론 단계에서는 5개의 설명문을 전제 문장 및 가설 문장과 하나씩 결합하여 총 5번을 추론하고 이 중에서 가장 많이 답변한 레이블을 최종 레이블로 결정한다. 그림 1은 지식 기반 설명문을 생성하여 추론하는 과정을 표현한 그림이다.

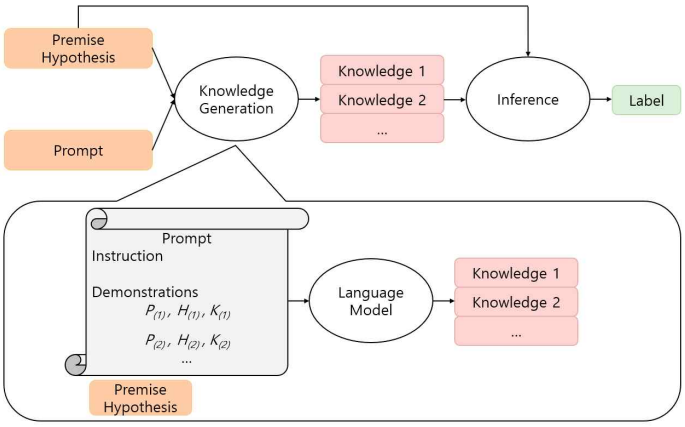

그림 1. 지식 기반 설명 생성 프롬프트를 이용한 추론

QA 변환 QA 변환은 전제 문장과 가설 문장의 관계를 추론하는 자연어 추론 태스크를 전제 문장을 Context로 하고 가설 문장을 Question으로 바꾸어 QA 태스크로 변환하는 방법이다. 표 1과 같이 가설 문장을 질문으로 변환하는 몇 가지 예시를 포함한 프롬프트를 이용해 새로운 가설 문장들을 질문으로 변환한 후 전제 문장을 Context로, 가설 문장을 Context에 대한 질문으로 바꾸어 모델에 {예, 아니오, 알 수 없음} 중 하나의 답변을 전달받고 이를 {예 -> Entailment, 아니오 -> Contradiction, 알 수 없음 -> Neutral}로 맵핑하여 최종 레이블로 결정한다.

Chain-of-Thought CoT는 전제 문장과 가설 문장의 레이블 사이에 어떠한 과정으로 해당 레이블이 도출되었는지를 기술하는 예시와 '순서대로 위 문장 쌍의 관계를 추론한다.' 라는 문장을 포함하는 프롬프트를 이용해 새로운 전제 문장과 가설 문장의 관계를 추론한다. 이 경우 모델은 프롬프트의 예시와 같이 추론 단계, 레이블 순으로 답변을 출력하고 답변에서 레이블만을 추출하여 최종 레이블로 결정한다.

4. 실험

본 논문에서는 자연어 추론과 지식 기반 설명문 생성을 위한 언어 모델로 GPT-3를 대화에 Fine-tuning한 GPT-3.5-turbo를 사용하였다. 한국어 자연어 추론 실험에 사용한 데이터는 KLUE-NLU 자연어 추론 데이터 셋의 검증셋으로 레이블 당 1,000개씩 총 3,000개로 이루어져 있다.

지식 기반 설명 생성과 CoT 프롬프트에 포함된 예시는 레이블 당 최대 8개씩 총 24개이고, QA 변환 프롬프트에 포함된 예시는 레이블 당 8개씩, 총 16개이다. 실험 결과는 레이블당 예시 0, 1, 2, 4, 8개 각각의 추론 성능

을 측정하였다.

실험에 사용된 GPT-3.5-turbo 모델은 채팅에 최적화되어 있기 때문에 기본 프롬프트(Baseline)와 지식 기반 설명 생성, QA 변환의 경우 레이블만을 반환받기 위해 출력 토큰의 최대 크기를 3으로 제한하고 소문자 변환, 공백 제거 등의 후처리 과정을 거친다. 지식 기반 설명문 생성에서 생성되는 하나의 설명문의 최대 토큰 수는 128토큰으로 제한하였다. CoT의 경우 실험에 쓰인 출력 토큰의 최대 크기는 256이다. 본 논문에서 사용한 기본 프롬프트(Baseline)는 세 가지 프롬프트 방법을 이용하지 않고 입력(전제 문장과 가설 문장)과 출력의 예시만 추가하여 추론한 결과이다.

4.1 실험 결과

표 2는 기본 프롬프트(Baseline)와 세 가지 프롬프트를 이용한 방법들의 성능을 비교한 표이다. “QA 변환 + 지식 기반 설명 생성”은 전제 문장은 Context로 사용하고 가설 문장으로 변환된 질문과, 생성된 지식 기반 설명을 결합하여 추론한 결과이다. “QA 변환 + 지식 기반 설명 생성 + CoT”는 표 1 CoT 프롬프트의 Premise와 Hypothesis를 Context와 Question으로 변환하고 중간 추론 과정과 생성된 지식 기반 설명을 결합하여 추론한 결과이다.

표 2를 통해 세 가지 프롬프트 방법 모두 기본 프롬프트(Baseline)보다 성능이 향상됨을 알 수 있다. 세 가지 방법을 각각 적용했을 때는 지식 기반 설명 생성이 가장 성능이 높았으며, 지식 기반 설명 생성과 QA 변환, CoT를 모두 적용한 추론의 경우 성능이 가장 높았다. CoT가 결합된 경우, 예시를 주지 않을 때 표 1의 CoT 예시와 같은 형식의 출력을 일관적으로 출력하지 못하므로 0-shot 실험 결과는 제외하였다.

표 2. 퓨샷 학습 환경의 프롬프트 방법 성능

	정확도(Acc)				
	0-shot	1-shot	2-shot	4-shot	8-shot
기본 프롬프트 (Baseline)	64.30	66.76	67.09	67.46	67.54
QA 변환	64.85	68.92	69.11	69.39	69.48
지식 기반 설명 생성[5]	65.41	73.83	75.50	76.03	76.17
CoT	-	70.14	70.79	71.20	71.47
QA 변환 + 지식 기반 설명 생성	**65.68**	74.08	75.59	76.14	76.35
QA 변환 + 지식 기반 설명 생성 + CoT	-	**74.19**	**75.87**	**76.39**	**76.58**

표 3과 4, 5는 실험에 사용된 데이터셋의 실제 입력과 출력 중 하나를 샘플링한 것이다. 표 4는 표 2 결과에 대한 모델의 실제 입력이다. QA + 지식 기반 설명 생성 + CoT 방법의 경우, 실험에 필요한 토큰 수와 시간 대비 성능 향상이 부진하여 한 번 만을 추론하여 최종 레이블을 결정하였다. 모든 입력의 앞에 표 2의 예시 개수에 따른 Example은 생략하였다.

표 3은 QA 변환과 지식 기반 설명 생성 프롬프팅으로 변환 및 생성된 질문과 지식 기반 설명이다.

표 3. 질문으로 변환된 가설 문장과 지식 기반 설명

{premise} 광주광역시 해외시장개척단이 러시아와 독일 수출 판로를 활짝 열었다.	{hyothesis} 광주광역시 해외시장개척단이 수출판로를 연 것은 독일뿐이다.
질문으로 변환된 가설 문장	
{question}: 광주광역시 해외시장개척단이 수출판로를 연 것은 독일뿐입니까?	
생성된 지식 기반 설명	
1. 전제 문장에서는 러시아와 독일 모두 수출판로를 열었다는 정보가 있다. 2. 러시아와 독일 모두 수출판로를 개척했다는 정보가 있으므로, 독일에만 수출판로를 개척했다는 추론은 부적절하다. 3. 전제 문장에서는 러시아와 독일이 모두 수출판로로 열렸다는 것을 언급하고 있다. 따라서 두 나라를 모두 포함한 활동으로 이해하는 것이 옳다. 4. 광주광역시 해외시장개척단이 러시아와 독일 수출판로를 모두 활짝 열었다는 것을 명시하고 있으므로 두 나라를 모두 포함한다는 것을 알 수 있다. 5. 러시아와 독일 두 나라의 수출판로를 개척했다는 정보가 있으므로, 광주시 해외시장개척단이 개척한 수출판로는 러시아와 독일 두 나라 모두 포함된다.	

표 4. 세 가지 프롬프트에 대한 입력

{premise} 광주광역시 해외시장개척단이 러시아와 독일 수출 판로를 활짝 열었다.	{hyothesis} 광주광역시 해외시장개척단이 수출판로를 연 것은 독일뿐이다.
기본 프롬프트(Baseline)	
Premise: {premise} Hypothesis: {hypothesis} Answer:	
QA 변환	
Context: {premise} Question: {question)} Answer:	
지식 기반 설명 생성	
Premise: {premise} Hypothesis: {hypothesis} Knowledge: {knowledge} Answer: (5개의 생성된 지식을 각각 결합하여 5번 추론 후 Voting)	
CoT	
Context: {premise} Question: {question} Answer: 순서대로 위 질문의 정답을 [예, 아니오, 알 수 없음]으로 추론한다.	
QA 변환 + 지식 기반 설명 생성	
Context: {premise} Question: {question} Knowledge: {knowledge} Answer: (5개의 생성된 지식을 각각 결합하여 5번 추론 후 Voting)	
QA + 지식 기반 설명 생성 + CoT	
Context: {premise} Question: {question} Knowledge: {knowledge} Answer: 순서대로 위 질문의 정답을 [예, 아니오, 알 수 없음]으로 추론한다.	

표 5는 표 4 입력에 대한 모델의 실제 출력이다.

표 5. 세 가지 프롬프트에 대한 출력

{premise} 광주광역시 해외시장개척단이 러시아와 독일 수출 판로를 활짝 열었다.	{hyothesis} 광주광역시 해외시장개척단이 수출판로를 연 것은 독일뿐이다.
Baseline contradiction	
QA 변환 아니오	
지식 기반 설명 생성 contradiction	
CoT 1. Premise에서 러시아와 독일 모두 수출판로를 열었다는 정보가 있다. 2. 하지만 Premise에서는 독일에만 수출판로를 개척했다고 추론했다. 3. Premise에 대한 Hypothesis의 추론은 부적절하다. 따라서 정답은 contradiction이다.	
QA 변환 + 지식 기반 설명 생성 아니오	
QA 변환 + 지식 기반 설명 생성 + CoT 1. 주어진 Context에서 러시아와 독일 양쪽으로 수출판로가 열렸다는 것을 명시하고 있다. 2. “러시아와 독일” 두 국가를 목적지로 나타내고 있다. 3. 하지만 Question은 독일에 대한 수출판로만을 언급하고 있다. 따라서 정답은 아니오이다.	

5. 결론

본 논문에서는 퓨샷 학습 환경의 한국어 자연어 추론에 지식 기반 설명 생성, QA 변환, CoT 프롬프트를 적용하고 이들을 조합하여 적용하는 방법을 제안하였다. 실험 결과 세 가지 프롬프트 방법 모두 기본 프롬프트보다 우수한 성능을 보였고, 세 가지 프롬프트 방법을 모두 조합하여 적용하였을 때 가장 우수한 성능을 보였다.

향후 연구로는 자연어 추론뿐만 아니라 상호참조해결 등과 같은 복잡한 태스크의 성능을 향상시킬 수 있는 다양한 프롬프트의 구조에 대해서 연구할 예정이다.

감사의 글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2022- 0-00369, (4세부) 전문지식 대상 판단결과의 이유/근거를 설명 가능한 전문가 의사결정 지원 인공지능 기술개발)

참고문헌

[1] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

[2] Kumar, Sawan, and Partha Talukdar. "NILE: Natural language inference with faithful natural language explanations." arXiv preprint arXiv:2005.12116 (2020).

[3] Liu, Jiacheng, et al. "Generated knowledge prompting for commonsense reasoning." arXiv preprint arXiv:2110.08387 (2021).

[4] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in Neural Information Processing Systems 35 (2022): 24824-24837.

[5] 최요한, 이창기, 배경만, "한국어 자연어 추론을 위한 거대 언어 모델 기반 지식 생성," 한국정보과학회 학술발표논문집, 2023, pp. 1066-1068.

[6] Yang, Xiaohan, et al. "What gpt knows about who is who." arXiv preprint arXiv:2205.07407 (2022).