자연어 추론 태스크에서의 프롬프트 엔지니어링 기법 및 언어 모델 성능 비교 연구

최효림○ 강원대학교빅데이터메디컬융합학과, 황현선 강원대학교컴퓨터공학과, 이창기 강원대학교컴퓨터공학과

rimi11e@kangwon.ac.kr, hhs4322@kangwon.ac.kr, leeck@kangwon.ac.kr

A Comparative Study of Prompt Engineering Techniques and Language Model Performance in Natural Language Inference Tasks

HyoRim Choi○, Hyunsun Hwang, Changki Lee

Department of Big Data Medical Convergence, Kangwon National University

요 약

자연어 추론(Natural Language Inference)은 전제와 가설 사이의 논리적 관계를 이해하고 분석하는 자연어처리 태스크로, 기계번역, 질의응답 등 다양한 자연어 처리 응용 분야에서 활용될 수 있다. 최근 거대 언어모델(Large Language Model)이 자연어처리의 다양한 태스크에 적용되면서, 사용자의 의도를 정확히 파악하고 최적의 답변을 얻기 위한 프롬프트(Prompt) 설계 기법인 프롬프트 엔지니어링에 대한 연구가 활발히 진행되고 있다. 프롬프트 엔지니어링 분야에서는 Chain-of-Thought(CoT), Tab-CoT, Plan and Solve 등의 다양한 기법들이 제안되고 있다. 본 연구에서는 이러한 프롬프트 엔지니어링 기법들을 자연어 추론 태스크에 적용하고, GPT3.5, GPT4, Llama3(8B, 70B), Phi3-mini 등 다양한 규모와 종류의 언어 모델을 대상으로 실험을 수행한다. 이를 통해 프롬프트 설계 방식과 언어 모델의 특성에 따른 자연어 추론 태스크의 성능 변화를 종합적으로 분석하고, 최적의 조합을 도출하고자 한다.

1. 서 론

자연어 추론(Natural Language Inference)은 전제(Premise)와 가설(Hypothesis) 사이의 논리적인 관계를 이해하고 분석하는 자연어처리 태스크이다. 전제와 가설에서 담고 있는 정보가 서로 연관성이 있는 경우 '함의(entailment)', 두 정보가 서로 반대되는 경우 '모순(contradiction), 두 정보가 서로 연관성이 없는 경우 '중립(neutral)'으로 분류되며 기계번역이나 질의응답 등 다양한 응용 분야에서 사용될 수 있다

최근 여러 거대 언어 모델(Large Language Model)과 경량 언어 모델(small Language Model)이 개발되면서 사용자의 의도를 보다 정확하게 파악하고 목적에 부합하는 최적의 답변을 제공하기 위한 프롬프트(Prompt) 연구의 중요성이 대두되고 있다. 이에 따라 프롬프트 엔지니어링 분야에서는 다양한 연구가 활발히 진행되고 있다. 대표적인 연구로는 Chain-of-Thought(CoT)[1], Plan and Solve[2], 그리고 Tab-CoT[3] 등이 있다. Chain-of-Thought(CoT)는 모델이 여러 단계의 추론 과정을 생성하도록 유도하는 프롬프트를 제공하여 추론의 정확성을 크게 향상시키는 기법이다[1]. Plan and Solve는 모델이 추론을 위한 계획을 먼저 수립하고, 그에 따라 문제를 단계별로 해결해 나가도록 하는 접근 방식이다[2]. Tab-CoT는 프롬프트를 표 형식으로 구성하여 모델에 입력함으로써, 구조화된 정보를 활용하여 더욱 효과적인 추론을 가능하게 한다[3].

본 논문에서는 기존에 진행된 다양한 프롬프트 엔지니어링 연구 방식을 자연어 추론 태스크에 적용하는 것과 더불어, GPT3.5, GPT4, Llama3(8B, 70B), Phi3-mini 등의 다양한 규모의 언어 모델에 대해서도 실험을 진행한다. 이를 통해 프롬프트 설계 방식과 언어 모델의 크기 및 종류에 따른 자연어 추론 태스크의 성능 변화를 종합적으로 분석하고, 최적의 조합을 도출하고자 한다.

2. 관련 연구

최근 거대 언어 모델의 성능을 향상시키기 위해 사용자의 의도를 더 정확하게 전달하고 모델이 목적을 명확히 이해할 수 있도록 하는 최적의 프롬프트를 찾는 연구가 많이 진행되고 있다. [1]에서는 "Let's think step by step"이라는

문장을 프롬프트에 포함시켜, 모델이 문제를 세분화하여 단계적으로 사고하고 해결할 수 있도록 유도한다. [2]에서는 "Let's first understand the problem and devise a plan to solve the problem. Then, let's carry out the plan and solve the problem step by step" 와 같은 문장으로 프롬프트를 구성하여 모델이 문제 해결을 위한 구체적인 계획을 수립하고, 그 계획에 따라 단계별로 문제를 해결할 수 있도록 유도한다. [3]은 [1]에서 제안한 CoT 프롬프트 기법처럼 모델이 문제를 세분화하여 추론할 수 있도록 하는 방식이나 세분화된 추론 과정을 표 형식으로 구성하여 복잡한 추론 과정을 명시적으로 나타낼 수 있도록 한다.

3. 제안기법

3.1 자연어 추론을 위한 프롬프트 구성

본 논문에서는 기존에 연구된 여러 프롬프트 방식을 자연어 추론 태스크에 맞게 내용을 재구성하여 적용하였다. 표 1은 Zero-shot 세팅 실험을 위해 프롬프트를 구성한 예시이다. Zero-shot 세팅에서는 태스크에 대한 설명과 추론 방법에 대한 설명으로 프롬프트가 구성된다. 3가지 기법 중 Plan and Solve 방식은 원래 모델이 스스로 문제 해결을 위한 계획을 수립하고, 해당 계획을 단계 별로 실행하며 문제를 해결하는 방식이나, 자연어 추론 태스크의 특성 상 수립할 수 있는 계획이 제한적이므로 1) 전제와 가설을 이해하고, 2) 전제와 가설 사이의 차이점을 파악하고, 3) 전제와 가설 사이의 논리적 관계를 추론하는 3단계로 Plan을 구성하였다. 또한 Plan의 단계 수를 다양화하기 위해 2단계 및 4단계 Plan도 추가로 구성하였다. 표 2는 3단계로 구성하였던 Plan and Solve를 각각 2단계, 4단계로 다양화하여 프롬프트를 구성한 예시이다. 2단계 Plan and Solve는 3단계 구성의 첫 번째 단계를 생략하여 구성하였고, 4단계 Plan and Solve는 각 라벨에 대한 간략한 설명으로 구성하여(step1:entailment, step2:contradiction, step3:neutral) 주어진 전제 가설이 어느 관계에 해당하는 지 판단할 수 있도록 구성하였다.

표 1. 자연어 추론을 위한 Zero-shot 프롬프트 구성

CoT	Infer whether the relationship between premise and hypothesis corresponds to 'contradiction', 'entailment', or 'neutral'. **Explain the reason for your inference** and at the end, answer with one of the following: 'contradiction', 'entailment', or 'neutral'. **Let's think step by step**
3 단계 Plan and Solve	Solve the problem step by step, following a plan to predict whether the hypothesis is "contradiction", "entailment" or "neutral" with respect to the premises, as shown in the example. **Step1**. Understand the premise and hypothesis **Step2.** Determine if there is a difference between the information that can be obtained from the premises and the information that can be obtained from the hypothesis. **Step3**.Based on what you understand, determine whether the premise falls under 'contradiciton', 'entailment', or 'neutral'.
Tab-CoT	Infer that the relationship between premises and hypothesis is 'contradiction', 'entailment', or 'neutral', as in the following examples. **| step | subquestion | results |** **| --- | --- | --- |**

표 2. 2단계, 3단계 4단계 Plan and Solve

2 단계 Plan and Solve	**Step1.** Determine if there is a difference between the information that can be obtained from the premises and the information that can be obtained from the hypothesis. **Step2.** Based on what you understand, determine whether the premise falls under 'contradiciton', 'entailment', or 'neutral'.
4 단계 Plan and Solve	**Step1.** Determine whether there is a superordinate or synonymous relationship between the words that make up the premise and the words that make up the hypothesis. **Step2.** Determined whether the words forming the premise and the words forming the hypothesis have an antonym relationship or a negative relationship, or whether they have no relationship with each other. **Step3.** Determine whether a hypothesis has been formed by adding additional information to the information obtained from the premises. **Step4.** Based on the judgments, infer whether the premise corresponds to 'contradiction', 'entailment', or 'neutral'.

3.2 In-Context Learning

In-context Few-shot Learning은 사전 학습된 거대 언어 모델에 해당 태스크의 설명과 몇 개의 예시를 입력으로 제공하여 모델의 성능을 향상시키는 방법이다[4]. 본 논문에서는 프롬프트 방식의 효과를 보다 정확하게 비교하기 위해 Zero-shot 세팅에서의 실험뿐만 아니라 In-context Few-shot Learning 세팅에서의 실험도 진행하였다. 구체적으로는 라벨별로 각각 4개의 예제를 제공하는 4-shot 세팅을 사용하였다. 4-shot 세팅의 프롬프트는 1) 자연어 추론 태스크에 대한 설명과 2) 추론 방식에 대한 설명 그리고 3) 추론 방식을 적용한 결과에 대한 예제(각 라벨별로 4개씩)로 구성된다

4. 실험 및 결과

본 논문에서는 자연어 추론 태스크에 맞게 구성한 다양한 프롬프트를 이용하여 다양한 규모의 언어 모델의 성능을 비교 분석하였다. 실험에는 OpenAI사에서 제공하는 GPT-3를 Fine-tuning한 GPT-3.5-turbo모델과 GPT-4-turbo모델, MetaAI사에서 제공하는 LLaMa3모델[5,6], 그리고 Microsoft사에서 제공하는 Phi-3-mini 모델을 사용하였으며,

표 4. Gpt-3.5-turbo/Gpt-4-turbo/LLaMa3/Phi-3 모델 실험결과

	ACC									
	gpt-3.5-turbo			LLaMa3		gpt-4-turbo			phi-3-mini(4K)	
	Zero-shot	4shot	4shot+self	4shot (8B)	4shot (70B)	Zero-shot	4shot	4shot+self	Zero-shot	4shot
Baseline	51.08	63.66	66.80	58.00	75.22	84.45	90.44	90.14	78.02	85.48
CoT	58.54	67.60	68.07	69.93	79.75	83.22	90.00	66.23	77.52	82.79
Tab-CoT	37.86	66.89	-	69.40	80.39	-	-	-	79.19	84.31
2단계 Plan and Solve	63.04	59.94	59.67	57.91	81.22	-	-	-	63.40	85.08
3단계 Plan and Solve	64.34	72.82	74.53	57.64	76.09	80.05	83.58	86.48	69.10	76.96
4단계 Plan and Solve	54.48	64.16	67.67	73.39	74.56	-	-	-	70.60	54.91

Zero-shot 세팅과 4-shot 세팅에서의 성능을 평가하였다. 실험에 사용된 데이터셋은 e-SNLI 데이터셋의 test셋으로, entailment 997개, neutral 988개, contradiction 1018개로 구성되어 있으며, 총 3,003개의 데이터로 이루어져 있다. LLaMa3-70B 모델의 경우, 메모리 사용량을 줄이기 위해 4bit 양자화 방식을 적용하여 실험을 진행하였다. 표 3는 GPT-3.5-turbo, GPT-4-turbo, LLaMa3(8B,70B), phi-3-mini 모델을 사용하여 각 프롬프트 방식에 따른 실험한 결과를 보여준다. 표 4에서 '4-shot+self'는 Self-consistency 기법을 적용한 실험 결과를 나타낸다. Self-consistency는 모델이 동일한 입력에 대해 여러 개의 답변을 생성하고, 그 중에 가장 일관성 있는 답변을 최종 추론 결과로 선택하는 방법이다[7]. 실험 결과, 자연어 추론 태스크에서 GPT-4-turbo 모델이 가장 우수한 성능을 보였으며, 특히 전제와 가설만 제공하는 Baseline 방식에서 다른 프롬프트 구성 방식보다 더 나은 결과를 보였다. 이는 GPT-4-turbo 모델이 추가적인 프롬프트 없이도 전제와 가설 간의 관계를 효과적으로 추론할 수 있는 능력을 가지고 있음을 시사한다. 반면, 다른 모델들의 경우에는 프롬프트 구성 방식에 따라 상이한 결과를 나타냈다. GPT-3.5-turbo 모델은 3단계 Plan and Solve 방식이 가장 좋은 성능을 보였으며, LLaMa3 모델의 경우 모델 크기에 따라 최적의 프롬프트 방식이 달랐다. 8B 모델은 2단계 Plan and Solve, 70B 모델은 4단계 Plan and Solve 방식에서 가장 우수한 성능을 보였다. 경량 언어 모델인 Phi-3-mini모델은 GPT-4-turbo와 유사하게 4shot 환경의 BaseLine 방식에서 가장 좋은 성능을 보였다. 또한, GPT-4-turbo를 제외한 나머지 언어 모델들에서는 CoT나 Tab-CoT보다 Plan and Solve 방식이 우수한 성능을 보였다. 이는 Plan and Solve 방식이 전제와 가설의 내용을 이해하고 두 문장의 차이점을 파악하여 논리적 관계를 단계적으로 추론하도록 유도하기 때문에, 자연어 추론 태스크에 더욱 적합한 프롬프트 구성 방식임을 시사한다.

이러한 결과는 언어 모델의 규모나 특성에 따라 최적의 프롬프트 구성 방식이 달라질 수 있음을 보여준다. 따라서 특정 태스크의 성능을 향상시키기 위해서는 언어 모델의 특성과 태스크의 유형을 고려하여 적절한 프롬프트 엔지니어링 전략을 수립하는 것이 중요하다.

5. 결론

본 연구에서는 자연어 추론 태스크에 대해 GPT-4-turbo, GPT-3.5-turbo, LLaMA-3, Phi-3-mini 등 다양한 언어 모델과 프롬프트 구성 방식을 비교 분석하였다. 실험 결과, GPT-4-turbo 모델의 Baseline 방식이 가장 좋은 결과를 나타냈다. 반면, 다른 모델들의 경우에는 프롬프트 구성 방식에 따라 상이한 결과를 보였으며, GPT-4-turbo를 제외한 나머지 언어 모델들에서 CoT나 Tab-CoT보다 Plan and Solve 방식이 우수한 성능을 보였다. 향후 연구로는 자연어추론의 성능을 향상시키면서도 추론 결과의 해석 가능성을 높일 수 있는 프롬프트 방법을 연구할 계획이다.

감사의 글

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

참고문헌

[1] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.

[2] Wang, Lei, et al. "Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models." *arXiv preprint arXiv:2305.04091* (2023).

[3] Jin, Ziqi, and Wei Lu. "Tab-cot: Zero-shot tabular chain of thought." *arXiv preprint arXiv:2305.17812* (2023).

[4] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004.

[5] Dettmers, Tim, et al. "Llm. int8 (): 8-bit matrix multiplication for transformers at scale, 2022." *CoRR abs/2208.07339*.

[6] Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." *Advances in Neural Information Processing Systems* 36 (2024).

[7] Wang, Xuezhi, et al. "Self-consistency improves chain of thought

reasoning in language models." arXiv preprint arXiv:2203.11171 (2022).