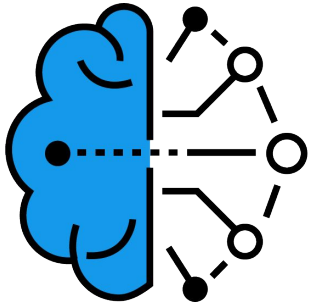
프롬프트 기반 학습을 통한 자연어 처리

Prompt-based Learning for Natural Language Processing



2021220699 이은찬

Kyungpook National University Brain AI Lab

2022-12-01 (목)

00

목차

Contents

0 Overview

NLP 모델의 엄청난 발전.

자연어 이해는 BERT-Family 모델로. 자연어 생성은 GPT-Family 모델로 구현 트렌드 양분화

자연어 처리 모델 용량의 엄청난 발전 (10억(B) 단위의 파라미터 수)

이러한 초대형 언어 모델을 다루는 기법들이 연구됨

최대 175B의 GPT-3 모델 연구를 통해 대중화된 방식인 '프롬프트'와 이를 이용한 '언어 모델의 퓨샷 학습' 제시

'생성 모델인 GPT-Family를 통해 자연어 이해를 퓨샷 학습으로 해결 한다'는 트렌드 자리 잡음

0

Overview

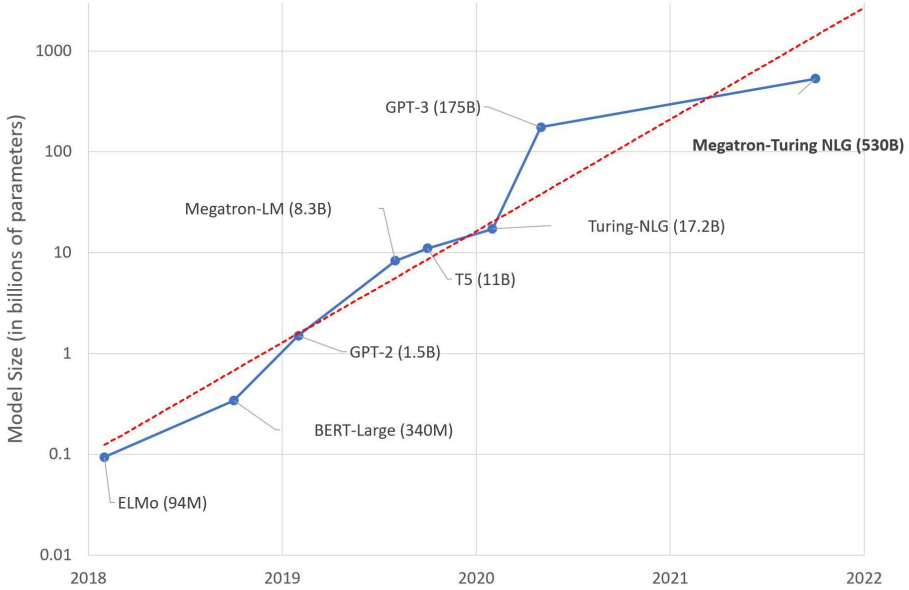
'프롬프트' 관련 연구가 2021년 NLP 연구의 하이라이트 중 하나로 선정. GPT-Family 생성 모델 기반의 '프롬프트 기반 퓨샷 러닝' 연구 트렌드 자리잡음 (2021 ~ now)

이러한 연구들은 굉장히 급격하게 발전 중이며 국제 학회 기준인 영어 기반의 NLP가 아닌, 한국어 등 소수의 언어들은 연구가 거의 이루어지지 못하고 있음

본 연구에서는 한국어 기반의 GPT-Family이자 1B 이상의 모델 용량을 가지는 KoGPT 모델을 통한 몇 가지 자연어 이해 문제를 '프롬프트 기반 퓨샷 학습'을 주요 Method로 사용하여 해결함

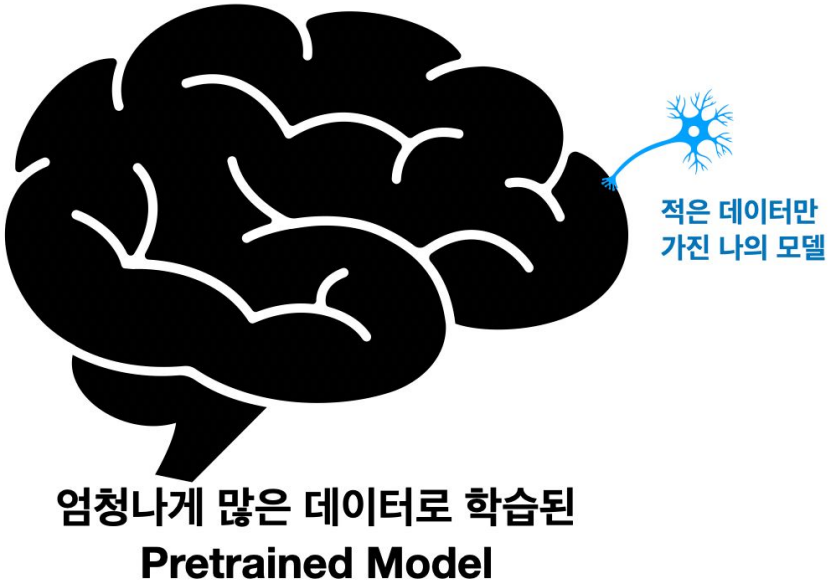
또한 여러 국제 학회의 프롬프트 관련 연구 실험을 공부 및 참고하여, 실험에 다양한 조작 변인을 주어서 어떠한 요소들이 한국어 기반 언어 모델의 프롬프트 기반 학습에 민감한 성능 변화 폭을 주는지 실험 결과를 제시하여 보고자 함

The tremendous growth of NLP Models



New ‘Moore’s Law’? 😆

Pretrained Model: NLP + Transfer Learning



Better performance using Pretrained Language Model (e.g. BERT, GPT)

MS Thesis Defense Presentation

Trend: BERT → Text Understanding, GPT → Text Generation!

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$BERT_{BASE}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

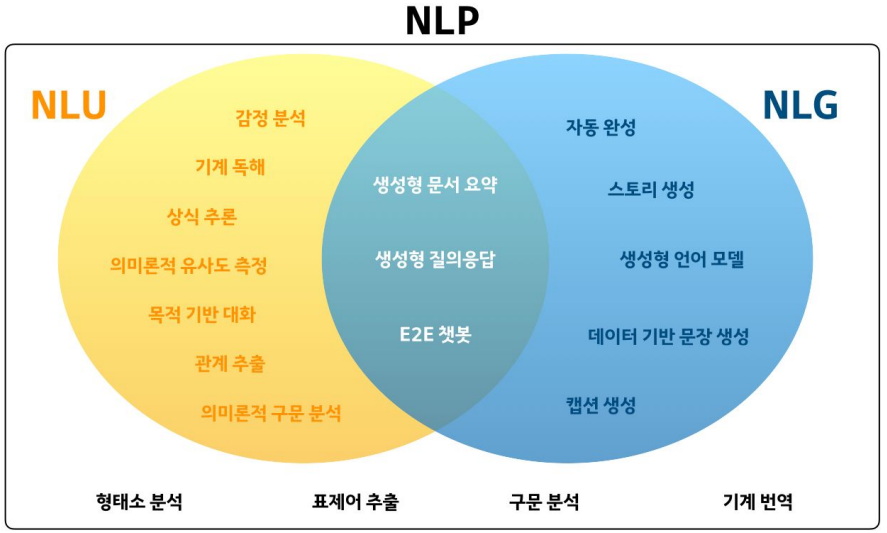
BERT with pre-train and fine-tune paradigm,
자연어 이해(NLU) SOTA

Language Models are Unsupervised Multitask Learners

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

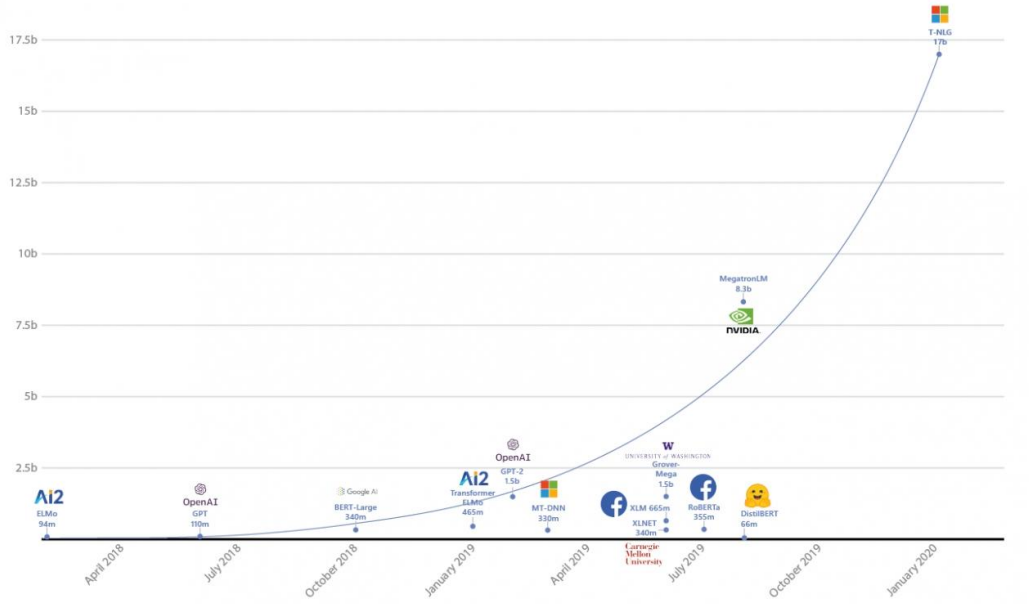
GPT with pre-train and fine-tune paradigm,
자연어 생성(NLG) SOTA



BERT and GPT-2,3 performance increases dramatically
with increasing model scale!

MS Thesis Defense Presentation

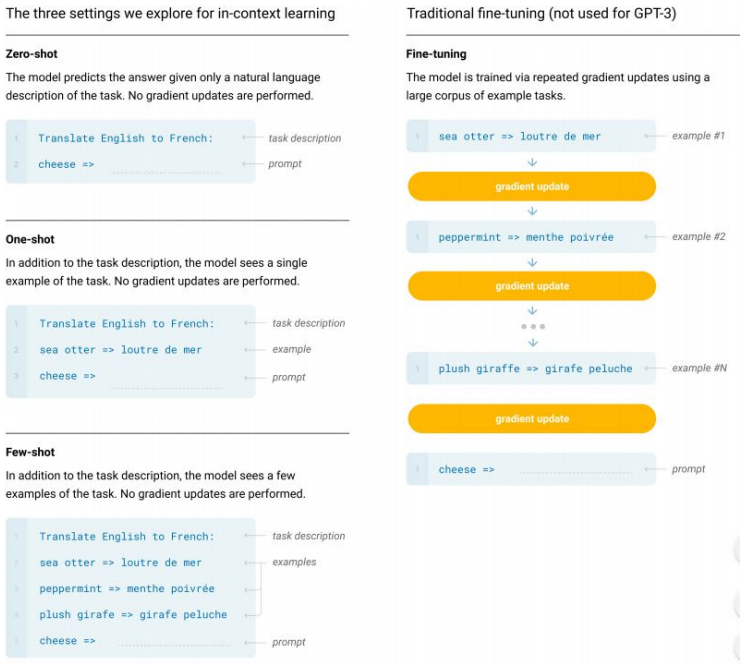
Huge advances in NLP model capacity (parameters in Billion+)



Larger than 1B+ parameters: 초대형 언어모델 (Large-Scale Language Models)

MS Thesis Defense Presentation

‘Prompt’: Studies on How to Handle Large-Scale LM



Language Models are Few-Shot Learners (2020, Brown et al)

1

Introduction

연구의 필요성

초대형 언어모델의 성능 고도화

- Large LM이 연구 트렌드가 됨
- GPT를 통한 Text Understanding 수행 가능성

한국어 기반 언어모델 인프라 이제 막 갖춰짐

- 한국어 초대형 언어모델의 등장
- 기존 한국어 Large LM Few-shot Learning 연구 매우 부족

프롬프트 연구 수요

- 프롬프트 기반의 학습을 한국어 모델, 태스크에 적용 수요 有
- 프롬프트, 불안정한 성능 → 요소들 비교 실험 필요

2 Background

NLP 모델의 발전.

자연어 이해는 BERT-Family 모델로, 자연어 생성은 GPT-Family 모델로 구현 트렌드 양분화

트랜스포머 구조+대용량 컴퓨팅 소스를 통한 자연어 처리 모델 용량의 발전. Google 등 주요 연구 기업체에서 전이학습에서 나온 개념인 **사전학습 기법**을 이용하여 대용량의 텍스트 데이터를 NLP 모델에 사전학습시켜 자연어 지식 이해 능력을 상당히 갖춘 상태의 모델을 오픈소스로 배포.

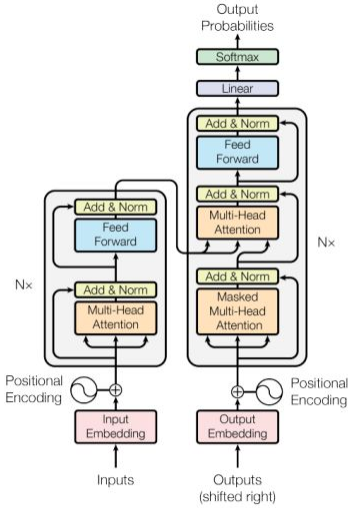
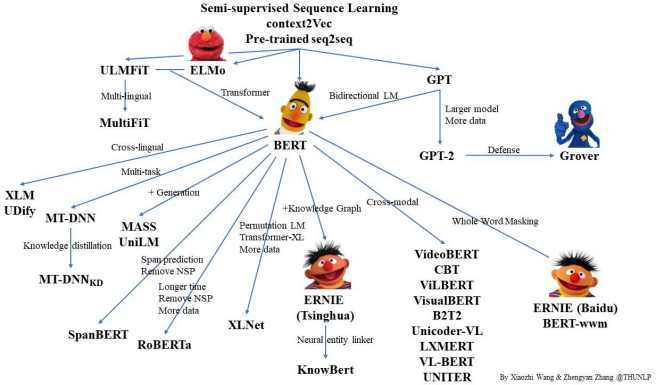


Fig. 2: The Transformer Architecture



2

Background

초대형 언어 모델 (Sec 2.3.3)

모델 파라미터 BERT 기준 100M, GPT 기준 1B 이상의 사전학습 된 트랜스포머 기반 초대형 모델들이 우후죽순 등장하고 있음

어떻게 모델 파라미터를 계산하는가?

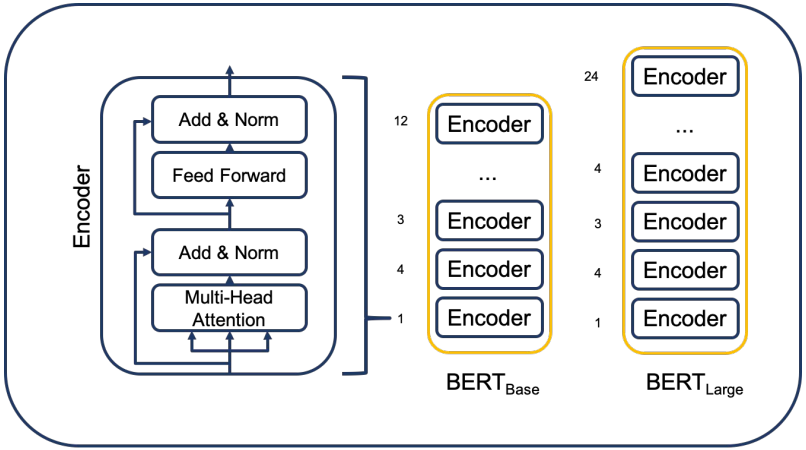


Table 1: Parameter information for major PLM

Models	nparams	nlayers	dmodel
GPT-3 6.7B	6.7B	32	4096
GPT-3 (Main, 175B)	175B	96	12288
GPT-2 XL	1.5B	48	1600
GPT-2	117M	12	768
BERT-base	110M	12	768
BERT-large	340M	24	1024
RoBERTa-large	340M	24	1024

BERT Layer	Key	Shape	Count
Embedding	embedding.word_embeddings.weight + position_embeddings.weight + token_type_embeddings.weight +LayerNorm.weight	[30522, 768] +[512,768] +[2,768] +[768] +[768]	23,837,184
Transformer Encoder x 12	encoder.layer.attention.self.query[weight,bias] +self.key[w,b] +self.value[w,b] +self.dense[w,b] +encoder.layer.intermediate.dense[w,b] +output.dense[w,b] +output.LayerNorm[w,b]	[768,768]+[768] x 4 [3072,768] + [3072] [768,3072] + [768] [768] x 2	7,087,872 x 12 = 85,054,464
Pooler	pooler.dense[w,b]	[768,768] +[768]	590,592

Fig. 3: Table showing how model parameter count by layer of BERT model is calculated

110M

2 Background

한국어 NLP 모델의 발전.

자연어 이해는 BERT-Family 모델로, 자연어 생성은 GPT-Family 모델로 구현 트렌드 양분화

이러한 흐름을 그대로 유지하고 학습 데이터를 한국어로 적용한 KoBERT, KoGPT, KoBART, KoELECTRA 등이 등장. 모델 파라미터, 학습 방식 등 거의 유사

한국어 태스크와 벤치마크:

주로 영어 기반의 데이터셋을 한국어로 변환한 느낌의 데이터가 많은 편. 최근 Korean Language Understanding Evaluation Dataset (KLUE)가 공개되어 자연어 추론(NLI), 대화 상태 추적(DST) 등 다양한 태스크 접근 용이해짐

AWS-SKT, 인공지능의 핵심... 한국어 자연어 처리(NLP) 모델 'KoGPT-2' 오픈 소스로 공개

구글의 버트, T-Brain의 KoBERT에 이은 'KoGPT-2'는 한국어로 학습된 오픈소스 기반 모델으로 챗봇 구축, 텍스트 감성 예측, 텍스트 분석 기반 응답 생성 등에

Models	nparams	nlayers	dmodel
GPT-3 XL	1.3B	24	2048
GPT-J 6B	6B	28	4096
GPT-3 6.7B	6.7B	32	4096
GPT-3 175B	175B	96	12288
KoGPT	6.16B	28	4096
HyperCLOVA	6.9B	28	4096

SST-2 [English]

contains very few laughs and even less surprises	**Negative (0)**
generous and subversive artworks!	**Positive (1)**
shot on ugly digital video	**Negative (0)**

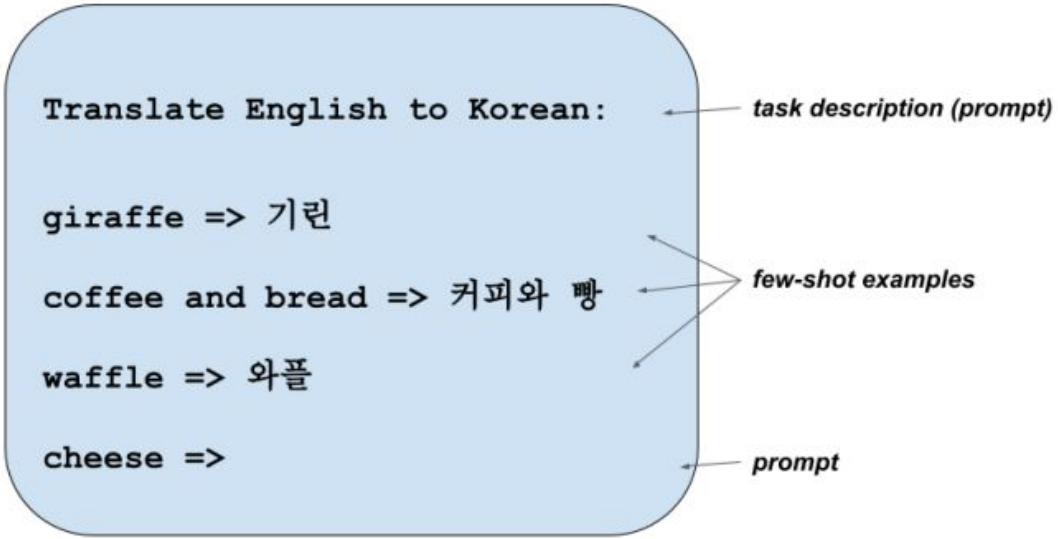
NSMC [Korean]

원작의 긴장감을 제대로 살려내지 못함	**Negative (0)**
사이먼 페그의 명연기가 돋보였던 영화!	**Positive (1)**
아 더빙 진짜 짜증나네요	**Negative (0)**

3

Method

Prompt-based Few-shot Learning



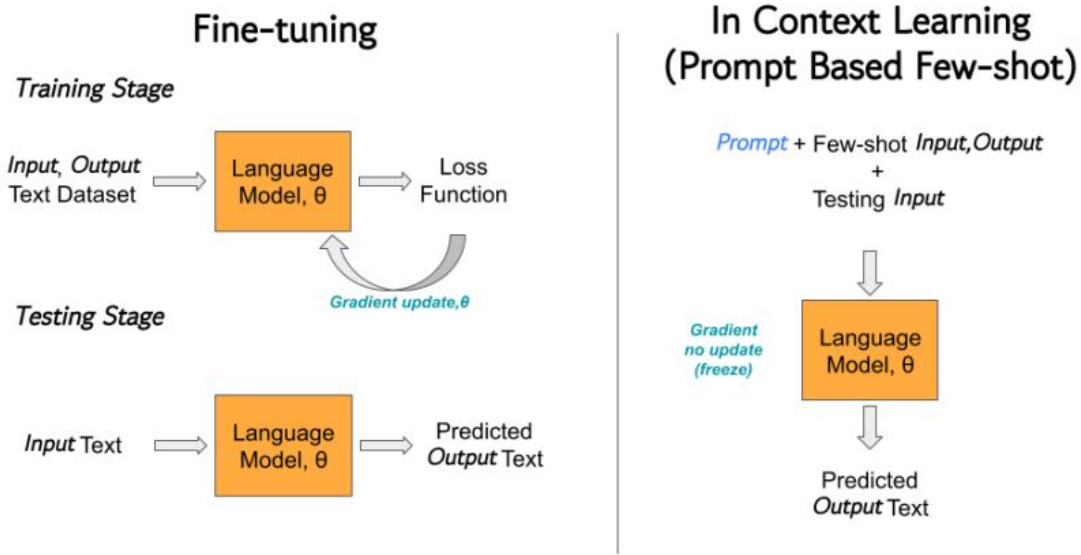
GPT-3에서 사용된 Prompt-based Few-shot Learning을 사용함

예시는 위와 같으며 모델은 적은 수의(퓨샷) 데이터와 '프롬프트'라는 텍스트 포맷의 조합으로 만들어진 텍스트를 입력으로 받아 맥락을 통해 태스크를 학습함. (Gradient-free)

3

Method

Fine-tuning vs Prompt-based Few-shot Learning



Prompt-based Few-shot Learning을 통해 Large LM은 텍스트의 맥락을 통해서 gradient 학습 없이 태스크를 학습함. 이 때문에 많은 데이터의 Training stage가 필요 없게됨

3

Method

Target Tasks: Korean NLU Dataset

SST-2 [English]

contains very few laughs and even less surprises	**Negative (0)**
generous and subversive artworks!	**Positive (1)**
shot on ugly digital video	**Negative (0)**

NSMC [Korean]

원작의 긴장감을 제대로 살려내지 못함	**Negative (0)**
사이먼 페그의 명연기가 돋보였던 영화!	**Positive (1)**
아 더빙 진짜 짜증나네요	**Negative (0)**

SQuAD [English]

The Lobund Institute grew out of pioneering research in germ-free-life which began in 1928 (...)	***Context***
When did study of a germ-free-life begin at Notre Dame?	***Question***
1928	***Answer***

KorQuAD [Korean]

1955년 프랑스의 탐험가인 Fernand Navarra가 발견한 목재 파편의 경우, 스페인의 임업 연구소에서 목재의 특성을 토대로 5000년 전의 것이라고 밝히긴 했으나 (...)	***Context***
1955년 프랑스 탐험가가 발견한 목재파편은 몇 년 전 것이라고 밝혀졌는가?	***Question***
5000년 전	***Answer***

Prompt-based Few-shot Learning을 통한 학습 성능을 검증하기 위한 타겟 태스크: 한국어 자연어 이해(NLU) 데이터셋 사용

1) **NSMC (Sentiment Analysis)**
2) **KorQuAD (Question Answering)**

Method

Large-scale Language Models

Table 2: Parameter information for major PLM

Models	nparams	nlayers	dmodel
GPT-3 XL	1.3B	24	2048
GPT-J 6B	6B	28	4096
GPT-3 6.7B	6.7B	32	4096
GPT-3 175B	175B	96	12288
KoGPT	6.16B	28	4096
HyperCLOVA	6.9B	28	4096

한국어 기반 Pretrained LM (PLM)인 KoGPT (made by Kakaobrain)를 한국어 초대형 언어 모델로 사용.

국제 학회에서 연구자들이 주로 사용하는 GPT-J (6B), GPT-3 (6.7B ver)과 nparam, nlayers, dmodel 매우 유사하여 연구 적용에 적합

Method

GPT Understand too

In-Context Learning (=prompt-based few-shot learning)

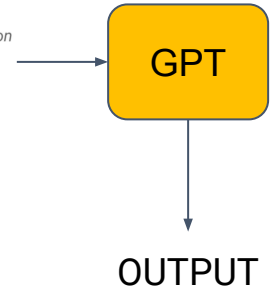
- Few-shot
- Prompt engineering
- Text-to-Text
- 생성 모델 (GPT)의 Text-to-text를 Few-shot NLU로 치환하여 적용 가능

few-shot + prompt

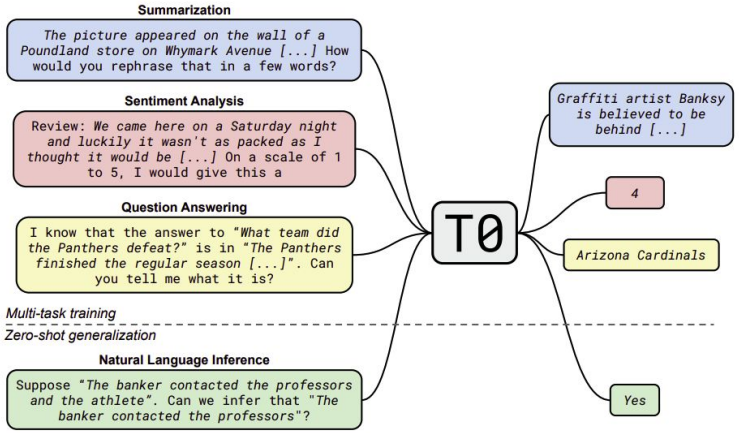
Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ← examples
4  plush girafe => girafe peluche      ← examples
5  cheese =>                           ← prompt
```



text-to-text



T0 (Sanh et al. 2021): Nice example of Text-to-Text

Method

Prompts for In-Context Learning

Prompt #1	{review} ({label})
Prompt #2	리뷰: {review} \n 감정: {label} \n
Prompt #3	리뷰: {review} \n 감정: {label} \n\n
Prompt #4	{{{review}}}\n 이 리뷰는 {{긍정,부정}}중 무엇?\n {label} \n\n

Fig. 8: Examples of 4 Prompts used for Prompt-based learning of NSMC datasets

Prompt #1	아래 지문을 읽고 물음에 답을 유추하시오. \n {context} \n 질문: {question} 답변: {answer}
Prompt #2	질의응답 태스크 수행: \n {context} \n 질문: {question} \n 답변: {answer}
Prompt #3	{context} \n {question}에 대해 답이 뭔지 아니? \n {answer} \n

Fig. 9: Prompt texts for Prompt-based learning for KorQuAD

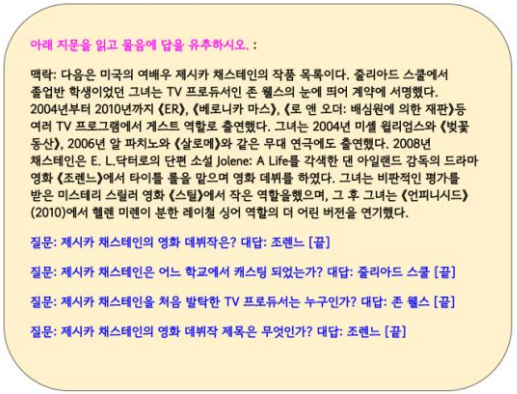


Fig. 10: Example of applying prompt engineering to the few-shot setting of the KorQuAD Dataset

적절한 프롬프트를 통해 모델이 퓨샷 러닝으로 데이터를 학습하도록 도움. 직접 관련 연구들에서 사용했던 프롬프트 형식들을 적절히 한국어로 구성하여 제작하고 실험에 비교군으로 사용 (sec 4)

3

Method

Prompt-based Few-shot Learning for New Custom Task

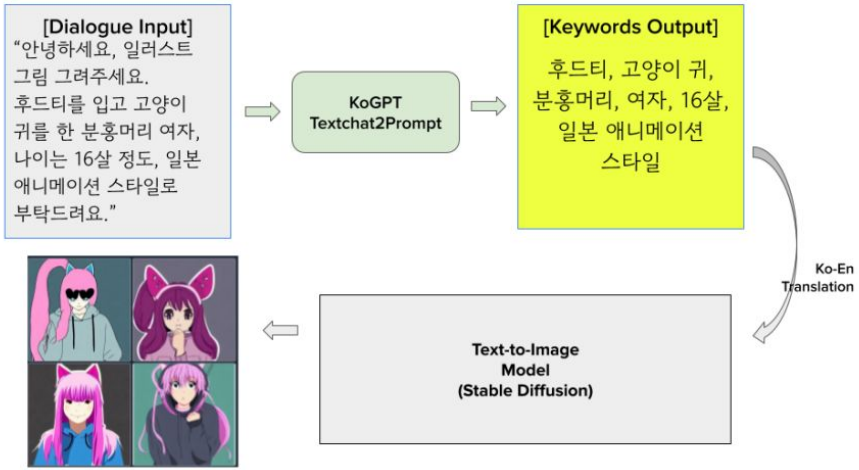


Fig. 11: Overview of the entire process of configuring Arbitrary Text Style Transfer as an interactive text-based image generation modeling

Used Prompt	입력:{input}\n 출력:{output}\n\n
Paired Data Example	**input:**웹툰 표지로 일러스트를 쓰고싶은데 그림 그려줘. 웹툰은 좀비 던전과 보물에 얽힌 비밀들에 관한 내용이고, 내가 필요 한 일러스트는 좀비 던전을 지나는 칼을 든 금발의 5살 여자 아이야. **output:**일러스트레이션, 웹툰, 좀비 던전, 칼을 든, 금발, 5살, 여자 아이

Fig. 12: Prompt text and few-shot data example for prompt-based arbitrary text style transfer (Textchat2prompt)

Prompt-based Few-shot Learning을 응용하여 새로운 Text Style Transfer 태스크를 제안

- 대화형 구어체 텍스트를 Text-to-Image Generation 모델에 적합한 키워드 텍스트로 변환
- 이러한 태스크에 대한 데이터셋이 없기 때문에. 퓨샷 러닝이 적합

Experimental Results

Experimental Setups

- Used Large-scale Language Model: KoGPT (6.16B) [Pretrained weight: 12.3 GB]
- Develop Environment: Colab Pro (GPU)
- Train Strategy: Prompt-based Few-shot Learning
- Model/Tokenizer Library: transformers (v4.22)

KoGPT 언어 모델을 불러오고 GPU를 통해 inference를 적용하기 위해 사용한 작업 환경 및 실험 설정

```
import torch
from transformers import AutoTokenizer, AutoModelForCausalLM

tokenizer = AutoTokenizer.from_pretrained(
  'kakaobrain/kogpt', revision='KoGPT6B-ryan1.5b-float16',  # or float32 ver
  bos_token='[BOS]', eos_token='[EOS]', unk_token='[UNK]', pad_token='[PAD]'
)
model = AutoModelForCausalLM.from_pretrained(
  'kakaobrain/kogpt', revision='KoGPT6B-ryan1.5b-float16',  # or float32 ver
  pad_token_id=tokenizer.eos_token_id,
  torch_dtype='auto', low_cpu_mem_usage=True
).to(device='cuda', non_blocking=True)
```

Fig. 13: Example of code to load Pre-trained KoGPT into huggingface's transformers

Experimental Results

Prompt-based Few-shot learning for Korean Sentiment Analysis

Setting

- Train Strategy: Prompt-based Few-shot Learning
- Used Pre-trained Language Model: Kakaobrain/KoGPT (6.16B ver)
- Used Benchmark Dataset: Naver Sentiment Movie Corpus (NSMC)
- Used Prompt formats: 4 prompt formats are used.

한국어 GPT + 프롬프트 기반 퓨샷 학습를 한국어 감정 인식 태스크인 NSMC에 적용하면 어떠한 성능을 얻을수 있는가? 또 어떠한 요인들이 성능에 작용하는가?를 보고자 함

Experiments

- Selection of prompt format : prompt #1 to prompt #4 are same as Sec 3.6 (Sec 4.2.1)
- Selection of Index of Example Data: Random select #1, #2, #3 (Sec 4.2.2)
- Selection of the number of Shots : 32,16,8,4,1 (Sec 4.2.3)

Prompt #1	{review} ({label})
Prompt #2	리뷰: {review} \n 감정: {label} \n
Prompt #3	리뷰: {review} \n 감정: {label} \n\n
Prompt #4	{{{review}}}\n 이 리뷰는 {{긍정,부정}}중 무엇?\n {label} \n\n

Fig. 8: Examples of 4 Prompts used for Prompt-based learning of NSMC datasets

NSMC [Korean]

원작의 긴장감을 제대로 살려내지 못함	Negative (0)
사이먼 페그의 명연기가 돋보였던 영화!	Positive (1)
아 더빙 진짜 짜증나네요	Negative (0)

Experimental Results

Prompt-based Few-shot learning for Korean Sentiment Analysis

4.2.1 Selection of prompt format

- The number of shots: k=32
- Few-shot data picking strategy: randomly picked.
- The number of test data that we used: 1000 (Test data[0:1000])
- The result was calculated as accuracy.

Matching Accuracy: 생성 기반 예측을 하는 모델이 올바른 레이블을 얼마나 예측하는지. 아예 상관없는 레이블 예측하는지를 나타내는 정확성 척도

Table 3: Results: Selection of Prompt Format

Prompt	Matching Accuracy	Acc (only Matching)	Acc (Total)
Prompt #1	0.660	**0.799**	0.460
Prompt #2	0.378	0.610	**0.242**
Prompt #3	**0.990**	0.640	**0.630**
Prompt #4	**0.998**	0.645	**0.643**

Prompt #1	{review} ({label})
Prompt #2	리뷰: {review} \n 감정: {label} \n
Prompt #3	리뷰: {review} \n 감정: {label} \n\n
Prompt #4	{{{review}}}\n 이 리뷰는 {{긍정,부정}}중 무엇?\n {label} \n\n

Fig. 8: Examples of 4 Prompts used for Prompt-based learning of NSMC datasets

Experimental Results

Prompt-based Few-shot learning for Korean Sentiment Analysis

4.2.1 Selection of prompt format (실험 결과 고찰)

Prompt #2와 #3의 개행문자 차이만으로 모델의 성능이 40% 가량 급변하였다. #2와 같이 개행문자가 퓨샷 데이터를 적절히 구분해주지 못한다면 모델은 불안정한 퓨샷 학습 성능을 보였다.

Prompt #1과 #4의 경우 더욱 직관적인 Prompt #4가 안정적인 성능을 보였다. #1의 경우는 NAVERCLOVA의 선행연구의 프롬프트 포맷을 사용했는데 성능이 높게 나오진 못했으며. 다만 Only Matching ACC가 가장 높게 나왔다.

결론적으로 prompt format의 선택은 같은 모델과 같은 퓨샷 데이터 조건에서도 매우 큰 성능에 영향을 미침을 확인하였다.

Table 3: Results: Selection of Prompt Format

Prompt	Matching Accuracy	Acc (only Matching)	Acc (Total)
Prompt #1	0.660	**0.799**	0.460
Prompt #2	0.378	0.610	**0.242**
Prompt #3	**0.990**	0.640	**0.630**
Prompt #4	**0.998**	0.645	**0.643**

Prompt #1	{review} ({label})
Prompt #2	리뷰: {review} \n 감정: {label} \n
Prompt #3	리뷰: {review} \n 감정: {label} \n\n
Prompt #4	{{{review}}}\n 이 리뷰는 {{긍정,부정}}중 무엇?\n {label} \n\n

Fig. 8: Examples of 4 Prompts used for Prompt-based learning of NSMC datasets

Experimental Results

Prompt-based Few-shot learning for Korean Sentiment Analysis

4.2.2 Selection of Index of Example Data

- The number of shots: k=32
- Few-shot data picking strategy: randomly picked.
- The number of test data that we used: 1000 (Test data[0:1000])
- Used Prompt format: Prompt 1 (Matching Acc: 0.660)

무작위의 index를 python code로 3번 추출하여 랜덤한 함수를 통해 32개의 index를 rule-based로 추출하여 few-shot을 위한 Example data가 바뀔 때 성능 변화를 보았다.

Table 4: Results: Selection of Index of Example data

Select	Matching Accuracy	Acc (only Matching)	Acc (Total)
Random Select #1	0.660	0.799	0.460
Random Select #2	0.702	0.806	**0.508**
Random Select #3	0.617	**0.859**	0.476

Prompt #1	{review} ({label})
Prompt #2	리뷰: {review} \n 감정: {label} \n
Prompt #3	리뷰: {review} \n 감정: {label} \n\n
Prompt #4	{{{review}}}\n 이 리뷰는 {{긍정,부정}}중 무엇?\n {label} \n\n

Fig. 8: Examples of 4 Prompts used for Prompt-based learning of NSMC datasets

Experimental Results

Prompt-based Few-shot learning for Korean Sentiment Analysis

4.2.2 Selection of Index of Example Data (실험결과고찰)

무작위의 index를 python code로 3번 추출하여 랜덤한 함수를 통해 32개의 index를 rule-based로 추출하여 few-shot을 위한 Example data가 바뀔 때 성능 변화를 보았다.

무작위로 32개의 퓨샷 학습을 위해 추출한 데이터의 index가 바뀜에 따라. 최대 4.8%의 성능 상승하락 폭이 있었다.

prompt select 실험에 비해서는 변화 폭이 적었으나. index 선택은 성능에 소폭 영향을 주었다. 이는 prompt-based learning의 모델 성능을 높이기 위해서 여러번의 indexing 작업을 통해 가장 좋은 결과를 사용하는 것이 좋음을 시사한다.

Table 4: Results: Selection of Index of Example data

Select	Matching Accuracy	Acc (only Matching)	Acc (Total)
Random Select #1	0.660	0.799	0.460
Random Select #2	0.702	0.806	**0.508**
Random Select #3	0.617	**0.859**	0.476

Prompt #1	{review} ({label})
Prompt #2	리뷰: {review} \n 감정: {label} \n
Prompt #3	리뷰: {review} \n 감정: {label} \n\n
Prompt #4	{{{review}}}\n 이 리뷰는 {{긍정,부정}}중 무엇?\n {label} \n\n

Fig. 8: Examples of 4 Prompts used for Prompt-based learning of NSMC datasets

Experimental Results

Prompt-based Few-shot learning for Korean Sentiment Analysis

4.2.3 Selection of the number of Shots

- The number of shots: k=32, 16, 8, 4, 1
- Few-shot data picking strategy: first randomly 32 → Then, randomly select k data from within 32
- The number of test data that we used: 1000 (Test data[0:1000])
- Used Prompt format: Prompt 4 (Matching Acc: 0.998)

퓨샷 학습을 위한 데이터 크기를 32개 부터 계속 줄이면 어떤 결과가 나올까?를 보고자 하였다. 이번에는 첫 실험에서 결과가 가장 좋았던 Prompt #4를 사용하였다.

Table 5: Results: Selection of the number of Shots

Shots	Matching Accuracy	Acc (only Matching)	Acc (Total)
k=32	0.998	0.645	0.643
k=16	0.997	0.620	0.616
k=8	0.983	0.644	0.627
k=4	0.940	0.578	0.518
k=1	**0.509**	**0.759**	**0.268**

Prompt #1	{review} ({label})
Prompt #2	리뷰: {review} \n 감정: {label} \n
Prompt #3	리뷰: {review} \n 감정: {label} \n\n
Prompt #4	{{{review}}}\n 이 리뷰는 {{긍정,부정}}중 무엇?\n {label} \n\n

Fig. 8: Examples of 4 Prompts used for Prompt-based learning of NSMC datasets

Experimental Results

Prompt-based Few-shot learning for Korean Sentiment Analysis

4.2.3 Selection of the number of Shots (결과 고찰)

퓨샷 학습을 위한 데이터 크기를 32개 부터 계속 줄이면 어떤 결과가 나올까?를 보고자 하였다. 이번에는 첫 실험에서 결과가 가장 좋았던 Prompt #4를 사용하였다.

k=32.16.8 까지는 변화가 미미하였으나. k가 4부터 1까지 떨어짐에 따라 처음에 비해 **최대 36%까지 급격하게 성능이 하락함을** 볼 수 있었다.

또한 99% 정도의 matching acc가 k=1에서 0.5로 급락하였는데. 이는 두개의 레이블인 NSMC 태스크를 학습하는데에. 한개의 데이터로는 binary classification 태스크에서 거의 레이블을 이해하지 못한다고 해석할 수 있다.

Table 5: Results: Selection of the number of Shots

Shots	Matching Accuracy	Acc (only Matching)	Acc (Total)
k=32	0.998	0.645	0.643
k=16	0.997	0.620	0.616
k=8	0.983	0.644	0.627
k=4	0.940	0.578	0.518
k=1	**0.509**	**0.759**	**0.268**

Prompt #1	{review} ({label})
Prompt #2	리뷰: {review} \n 감정: {label} \n
Prompt #3	리뷰: {review} \n 감정: {label} \n\n
Prompt #4	{{{review}}}\n 이 리뷰는 {{긍정,부정}}중 무엇?\n {label} \n\n

Fig. 8: Examples of 4 Prompts used for Prompt-based learning of NSMC datasets

Experimental Results

Prompt-based Few-shot learning for Korean QA

Setting

- Train Strategy: Prompt-based Few-shot Learning
- Used Pre-trained Language Model: Kakaobrain/KoGPT (6.16B ver)
- Used Benchmark Dataset: LG-CNS/ Korean Question Answering Dataset ver 1.0 (KorQuAD)
- Used Prompt formats: 3 prompt formats are used.

KorQuAD [Korean]

1955년 프랑스의 탐험가인 Fernand Navarra가 발견한 목재 파편의 경우, 스페인의 임업 연구소에서 목재의 특성을 토대로 5000년 전의 것이라고 밝히긴 했으나 (...)	*Context*
1955년 프랑스 탐험가가 발견한 목재파편은 몇 년 전 것이라고 밝혀졌는가?	*Question*
5000년 전	*Answer*

Experiments

- Selection of prompt format (prompt #1 to prompt #3, same as Sec 3.6) (Sec 4.3.1)
- Selection of label: Gold label vs random label (Sec 4.3.2)

Prompt #1	아래 지문을 읽고 물음에 답을 유추하시오. \n {context} \n 질문: {question} 답변: {answer}
Prompt #2	질의응답 태스크 수행: \n {context} \n 질문: {question} \n 답변: {answer}
Prompt #3	{context} \n {question}에 대해 답이 뭔지 아니? \n {answer} \n

Experimental Results

Prompt-based Few-shot learning for Korean QA

4.3.1 Selection of prompt format

- The number of shots: k=4
- Few-shot data picking strategy: (i) context: randomly → (ii) 4 QA pairs on the context are randomly selected.
- The number of test data that we used: 5774 (All Test data are used)
- Evaluation Metrics: F1 Score / Exact Match (EM)

선행연구를 참조 K=4를 적합한 데이터 수로 판단.

평가 방식으로는 F1 / EM를 사용했는데, EM은 완전히 정답을 맞추는 경우, F1은 맞춘 정답 텍스트 시퀀스에 정답이 포함되는 것도 정답으로 보는 조금 더 관대한 스코어 방식.

프롬프트 포맷으로 3가지 매뉴얼 준비. 어떤 변화/얼마정도의 성능차가 발생할까?를 보고자 함

Prompt	F1	Exact Match (EM)
Prompt #1	**62.889**	**41.756**
Prompt #2	59.387	35.036
Prompt #3	53.836	30.880

Prompt #1	아래 지문을 읽고 물음에 답을 유추하시오. \n {context} \n 질문: {question} 답변: {answer}
Prompt #2	질의응답 태스크 수행: \n {context} \n 질문: {question} \n 답변: {answer}
Prompt #3	{context} \n {question}에 대해 답이 뭔지 아니? \n {answer} \n

Experimental Results

Prompt-based Few-shot learning for Korean QA

4.3.1 Selection of prompt format (실험결과고찰)

선행연구를 참조 K=4를 적합한 데이터 수로 판단.

평가 방식으로는 F1 / EM를 사용했는데. EM은 완전히 정답을 맞추는 경우. F1은 맞춘 정답 텍스트 시퀀스에 정답이 포함되는 것도 정답으로 보는 조금 더 관대한 스코어 방식.

프롬프트 포맷으로 3가지 매뉴얼 준비. 어떤 변화/얼마정도의 성능차가 발생할까?를 보고자 함

QA 태스크는 앞선 섹션의 실험 NSMC (감성분류)보다 정답 예측이 어렵다. Prompt #1처럼 적절한 포맷의 프롬프트가 갖춰졌을때가 다소 구어체의 Prompt #3보다 성능이 최대 11% 상승하는 것을 볼 수 있었으며. 프롬프트에 태스크 관련 정보량이 많은 편인 #1이 #2보다 최대 6.7%의 EM 상승폭이 있었다.

이는 프롬프트의 올바른 선택이 Prompt-based Learning을 통한 QA태스크 성능에 영향을 주는 것으로 볼 수 있다.

Prompt	F1	Exact Match (EM)
Prompt #1	**62.889**	**41.756**
Prompt #2	59.387	35.036
Prompt #3	53.836	30.880

Prompt #1	아래 지문을 읽고 물음에 답을 유추하시오. \n {context} \n 질문: {question} 답변: {answer}
Prompt #2	질의응답 태스크 수행: \n {context} \n 질문: {question} \n 답변: {answer}
Prompt #3	{context} \n {question}에 대해 답이 뭔지 아니? \n {answer} \n

Experimental Results

Prompt-based Few-shot learning for Korean QA

4.3.2 Selection of label: Gold label vs random label

- The number of shots: k=4
- Few-shot data picking strategy: (i) context: randomly → (ii) 4 QA pairs on the context are randomly selected.
- The number of test data that we used: 5774 (All test data are used)
- Label → Gold label, Random label, zero-shot (k=0, only prompt)
- Used Prompt format: Prompt 1 (F1/EM:62.889, 41.756)

K=4의 퓨샷 학습 데이터의 label이 만약 정답이 아니라 무작위의 랜덤값, 혹은 아예 레이블이 없다면 결과는 어떻게 변화할 수 있을지를 보고자 함.
Prompt 포맷은 가장 성능이 좋았던 #1을 적용함

Table 7: Results: Selection of label: Gold label vs random label for KorQuAD

prompt	Label	F1	Exact Match (EM)
Prompt #1	gold label	62.889	41.756
Prompt #1 (same)	random label	56.064	32.577
Prompt #1 (same)	zero-shot	41.300	11.552

Prompt #1	아래 지문을 읽고 물음에 답을 유추하시오. \n {context} \n 질문: {question} 답변: {answer}
Prompt #2	질의응답 태스크 수행: \n {context} \n 질문: {question} \n 답변: {answer}
Prompt #3	{context} \n {question}에 대해 답이 뭔지 아니? \n {answer} \n

Experimental Results

Prompt-based Few-shot learning for Korean QA

4.3.2 Selection of label: Gold label vs random label

(실험결과 고찰)

K=4의 퓨샷 학습 데이터의 label이 만약 정답이 아니라 무작위의 랜덤값. 혹은 아예 레이블이 없다면 결과는 어떻게 변화할 수 있을지를 보고자 함.
Prompt 포맷은 가장 성능이 좋았던 #1을 적용함

결론적으로. 올바른 label이 random으로 대체되었을때 F1/EM은 6.8%. 9%로 크게 하락하였으며 이를 no label로 대체했을때는 F1/EM이 21.5%. 30%로 상징적으로 감소함을 확인할 수 있었다.

이는 선행연구에서 보여주었던 다른 자연어 이해 태스크에서 random label이 크게 중요하지 않을 수 있음을 보인것과 달리. Korean QA 태스크에서는 label의 정확성 여부가 성능을 크게 좌우함을 의미한다.

Table 7: Results: Selection of label: Gold label vs random label for KorQuAD

prompt	Label	F1	Exact Match (EM)
Prompt #1	gold label	62.889	41.756
Prompt #1 (same)	random label	56.064	32.577
Prompt #1 (same)	zero-shot	41.300	11.552

Prompt #1	아래 지문을 읽고 물음에 답을 유추하시오. \n {context} \n 질문: {question} 답변: {answer}
Prompt #2	질의응답 태스크 수행: \n {context} \n 질문: {question} \n 답변: {answer}
Prompt #3	{context} \n {question}에 대해 답이 뭔지 아니? \n {answer} \n

Experimental Results

Prompt-based Learning for Arbitrary Text Style Transfer

Goal of Experiment

앞선 실험에서 한국어 모델+태스크에도 어느정도의 프롬프트 기반 퓨샷 학습의 성능이 나옴을 보였다.

본인은 이를 few-shot learning의 장점을 살려 직접 k=4개의 데이터 쌍을 구축했을 때 모델이 새로운 태스크를 수행할 수 있는지 검증하고자 텍스트 기반 이미지 생성 태스크에서 대화형 구어체 텍스트 입력을 이미지 생성을 잘 하지 못하는 단점을 퓨샷 학습을 통해 텍스트를 적합한 입력으로 변환하여 더 나은 이미지를 생성하도록 하는 Textchat2prompt 모델을 구성하였다.

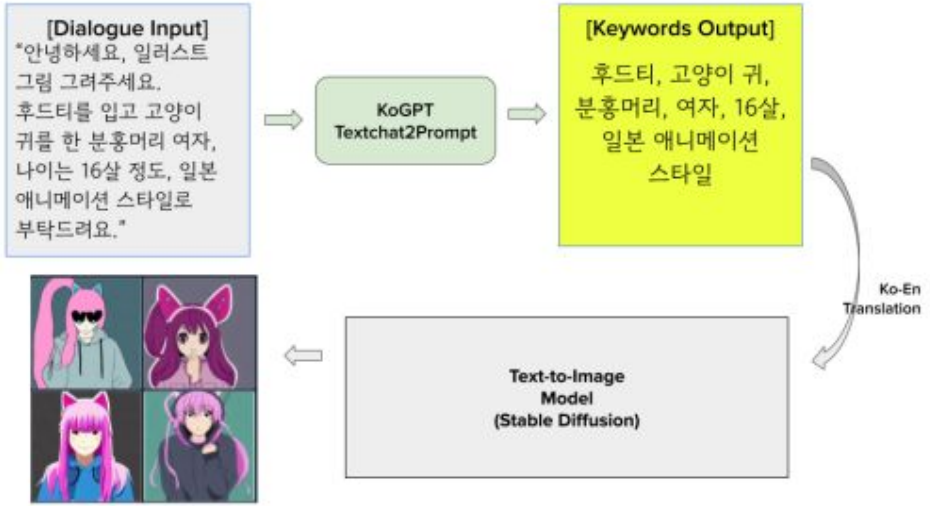


Fig. 11: Overview of the entire process of configuring Arbitrary Text Style Transfer as an interactive text-based image generation modeling

Experimental Results

Prompt-based Learning for Arbitrary Text Style Transfer

Setting

- Main Task: (Input= Dialogue Text) →[Few-shot Text style Transfer] → (Output= Command-type Text) → (Better) Text to Image Generation
- Used Pre-trained Language Model: KoGPT 6.16B
- Used Training Strategy: Prompt-based Few-shot Learning
- The number of shots: k=4

- Few-shot data: manually generated data pairs (same as Section 3.7)
- Used Text-to-Image Generation Model: Stability AI's Stable Diffusion [42]
- Evaluation Metric: Qualitative evaluation(정성적 평가방식) (based on qualitative evaluation of image generation because there is a limitation that there is no suitable metric because it is a custom task)

Experimental Results

Prompt-based Learning for Arbitrary Text Style Transfer

Used Prompt and Data that I made.

Used Prompt	입력:{input}\n 출력:{output}\n\n
Paired Data Example	**input:**웹툰 표지로 일러스트를 쓰고싶은데 그림 그려줘. 웹툰은 좀비 던전과 보물에 얽힌 비밀들에 관한 내용이고, 내가 필요 한 일러스트는 좀비 던전을 지나는 칼을 든 금발의 5살 여자 아이야. **output:**일러스트레이션, 웹툰, 좀비 던전, 칼을 든, 금발, 5살, 여자 아이

Fig. 12: Prompt text and few-shot data example for prompt-based arbitrary text style transfer (Textchat2prompt)

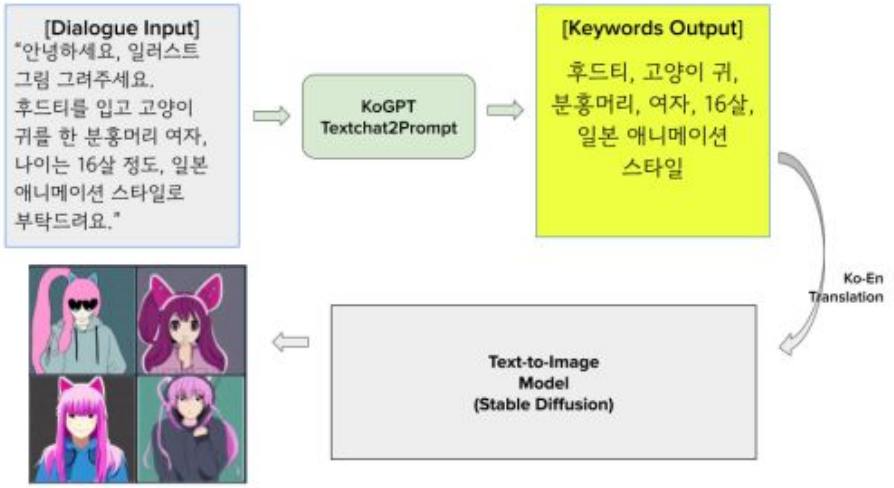


Fig. 11: Overview of the entire process of configuring Arbitrary Text Style Transfer as an interactive text-based image generation modeling

Experimental Results

Prompt-based Learning for Arbitrary Text Style Transfer

실험 결과

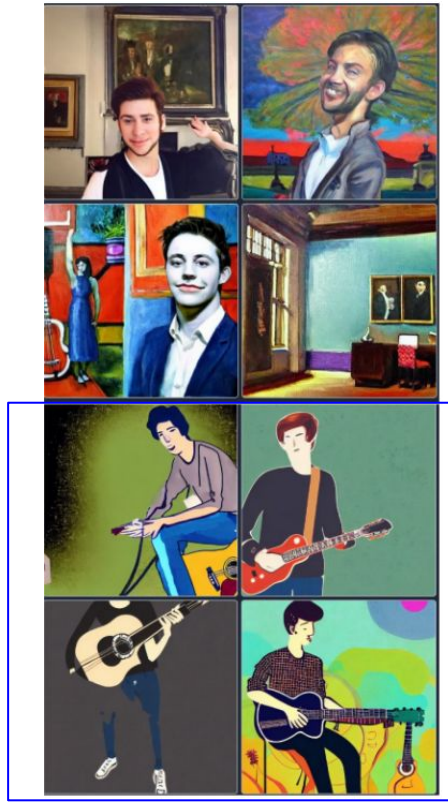
Results: Arbitrary Text Style Transfer, Dialogue to Prompt input

	Text
Example #1 Input Dialogue Text	"그림을 그려주세요. 음 기타를 든 잘생기고 젊은 가수로 부탁하고, 배경은 대학교 축제로 해주세요!"
Example #1 style-changed output	**대학교 축제, 기타를, 든, 잘생기고, 젊은, 가수, 배경**
Example #2 Input Dialogue Text	안녕하세요, 일러스트 그림 그려주세요. 후드티를 입고 고양이 귀를 한 분홍머리 여자, 나이는 16살 정도, 일본 애니메이션 스타일로 부탁드려요.
Example #2 style-changed output	**후드티, 고양이 귀, 분홍머리, 여자, 16살, 일본 애니메이션 스타일**

Fig. 14: Results: Arbitrary Text Style Transfer, Dialogue to Prompt input

Used Prompt	입력:{input}\n 출력:{output}\n\n
Paired Data Example	input:웹툰 표지로 일러스트를 쓰고싶은데 그림 그려줘. 웹툰은 좀비 던전과 보물에 얽힌 비밀들에 관한 내용이고, 내가 필요 한 일러스트는 좀비 던전을 지나는 칼을 든 금발의 5살 여자 아이야. output:일러스트레이션, 웹툰, 좀비 던전, 칼을 든, 금발, 5살, 여자 아이

Fig. 12: Prompt text and few-shot data example for prompt-based arbitrary text style transfer (Textchat2prompt)



better!

5 Conclusions

프롬프트 기반 (퓨샷) 학습 method를 사용하여 모델을 gradient 학습시키지 않고, 학습시키는 방식을 한국어 GPT 모델과 한국어 기반 자연어 태스크에 적용하는 실험을 적용했다.

또한 다양한 양적인 실험을 통해 태스크마다 프롬프트의 적절한 선택, 퓨샷의 샷 개수, 레이블의 정확성 여부 등이 프롬프트 기반 퓨샷 학습의 성능에 영향이 있음을 도출하였다.

또한 이 방식을 Text-to-Image generation의 대화형 구어체(textchat)의 낮은 이미지 생성 성능을 개선하기 위한 새로운 텍스트 스타일 변환 태스크에 적용해보며, 프롬프트 기반 퓨샷 학습의 실용성을 어느정도 입증하였다.

결론적으로 한국어 초대형 언어 모델인 KoGPT 모델이 프롬프트 기반 학습을 통해 적절하게 사용될 수 있기 적절한 모델임을 증명하였다. 이는 한국어 초대형 언어 모델을 통해 조금 더 이러한 다양한 연구가 진행될 수 있음을 시사한다.

Future work로는 Prompt를 직접 고안하는 것이 아닌 모델 기반의 자동 고안 방식, 프롬프트 튜닝 등을 고려할 수 있다.

향후 다양한 한국어 NLP 연구자들에게 언급될 수 있는 연구 논문 성과로 이어질 수 있다면 좋을 것이다.

Thank You