[week1] Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers

What Changes Can Large-scale Language Models Bring?
Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers

Abstract

💡 한국어 코퍼스를 이용한 GPT-3 기반 학습 진행 → HyperCLOVA

한국어 downstream task에 대한 few-shot learning & zero-sot learning 부분 SOTA 달성

No Code AI에 대한 패러다임 제시

Introduction

GPT-3를 활용한 in-context learning이 주목 받고 있음

그러나 practical한 문제점 존재

→ GPT-3의 학습 코퍼스의 92.7%가 영어 (다른 언어에 대한 고려 부족)

→ 접근 가능한 모델 사이즈가 13B , 175B (그 사이는 없음)

→ 발전된 prompt-based method가 필요함

- HyperCLOVA는?

  1) 560B의 토큰으로 이루어진 한국어 코퍼스를 이용해 학습시킨 LM

  2) 비영어권 코퍼스에 대한 language-specific tokenization 사용

  3) prompt-based tuning을 통해서 zero-shot, few-shot learning에 대한 성능 향상

  4) in-house 어플리케이션을 통해 No Code AI 제공 (HyperCLOVA Studio)

Pre-training

- **Data Description**

Name	Description	Tokens
Blog	Blog corpus	273.6B
Cafe	Online community corpus	83.3B
News	News corpus	73.8B
Comments	Crawled comments	41.1B
KiN	Korean QnA website	27.3B
Modu	Collection of five datasets	6.0B
WikiEn, WikiJp	Foreign wikipedia	5.2B
Others	Other corpus	51.5B
Total		561.8B

Table 1: Descriptions of corpus for HyperCLOVA

기존 GPT-3는 학습 코퍼스에 한국어 비중이 적기 때문에 한국어 LM의 학습을 위해 직접 코퍼스 생성

총 561B의 토큰으로 구성된 한국어 코퍼스를 pre-train을 위해 런덤하게 샘플링

pre-train에는 openAI의 GPT-3 구조 차용

- **Model and Learning**

  openAI의 GPT-3가 사용한 transformer의 decoder 아키텍처 이용

  mid-size의 파라미터 개수 사용 (39B, 82B)

  +) test 시에는 HyperCLOVA 코퍼스에 포함되지 않은 encyclopedia코퍼스를 이용해서 loss 측정

- **Korean tokenization**

  한국어는 agglutinative language (교착어)

  → 동사나 형용사 어근 뒤에 명사가 따라오는 형태

  따라서 영어와는 다른 형태의 토큰화가 필요하다

  ⇒ morpheme-aware byte-level BPE 이용 (BPE에 morpheme 분석이 추가)

  (morpheme analyzer로 문장을 pre-split하고 그다음에 토큰화 - HuggingFace의 토크나이징 라이브러리 사용)

Experimental Results

- **In-context Few-shot learning**

	NSMC (Acc)	KorQuAD (EA / F1)		AI Hub (BLEU) Ko→En	AI Hub (BLEU) En→Ko	YNAT (F1)	KLUE-STS (F1)
Baseline	89.66	74.04	86.66	40.34	40.41	82.64	75.93
137M	73.11	8.87	23.92	0.80	2.78	29.01	59.54
350M	77.55	27.66	46.86	1.44	8.89	33.18	59.45
760M	77.64	45.80	63.99	2.63	16.89	47.45	52.16
1.3B	83.90	55.28	72.98	3.83	20.03	58.67	60.89
6.9B	83.78	61.21	78.78	7.09	27.93	67.48	59.27
13B	87.86	66.04	82.12	7.91	27.82	67.85	60.00
39B	87.95	67.29	83.80	9.19	31.04	71.41	61.59
82B	88.16	69.27	84.85	10.37	31.83	72.66	65.14

Table 3: Results of in-context few-shot tasks on question answering, machine translation, topic classification, and semantic similarity per model size. As baselines, we report the results of BERT-base for NSMC and KorQuAD, and Transformer for AI Hub from Park et al. (2020). mBERT is used for KLUE-YNAT and KLUE-STS from Park et al. (2021).

→ 6가지 task에 대한 few-shot learning 결과

⇒ real-word application에 대해서는 더 좋은 결과를 보임을 확인 가능

+) prompt 튜닝이 더해진다면 성능 향상 가능할 것

- **Prompt-based Tuning**

input side에 p-tuning 적용

→ input 쿼리의 퀄리티를 향상시키는 것에 도움을 줌을 확인 가능

- **Effect of Tokenization**

	KorQuAD (EA / F1)		AI Hub (BLEU) Ko→En	AI Hub (BLEU) En→Ko	YNAT (F1)	KLUE-STS (F1)
Ours	**55.28**	**72.98**	3.83	**20.03**	**58.67**	**60.89**
byte-level BPE	51.26	70.34	**4.61**	19.95	48.32	60.45
char-level BPE	45.41	66.10	3.62	16.73	23.94	59.83

Table 6: Effects of tokenization approaches on three tasks. HyperCLOVA-1.3B is used for evaluation.

AI hub (ko→en) 제외하고는 모두 제일 좋은 성능을 보임

+) char-level BPE는 OOV를 만들어낸다 (다른 2가지는 그렇지 않음) → 그래서 성능 안 좋은 것

Discussion on Industrial Impacts

- **HyperCLOVA Studio**

  1) GUI 인터페이스 제공

  2) API call으로 output 쉽게 얻을 수 있는 API end point 제공

- **Rapidly Prototyping Chatbots with Personalities**

  HyperCLOVA를 이용했을 때 특정 캐릭터의 페르소나를 가진 챗봇을 쉽게 디자인 가능했음

- **Zero-shot Transfer Data Augmentation**

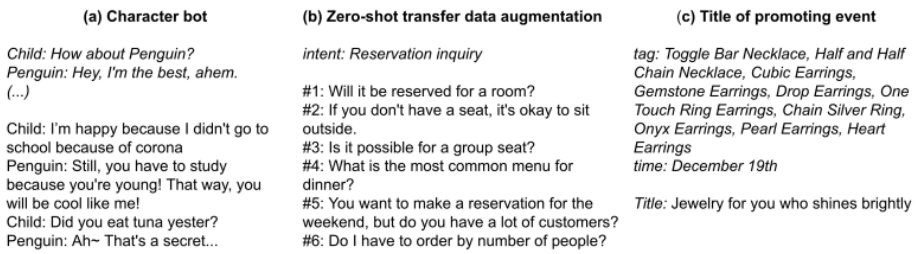

Figure 1: Examples generated by HyperCLOVA with the prompts under three different tasks. Italic implies given prompts and non-italic corresponds to generated outputs. The examples are translated into English.

→ 사용자의 의도에 맞춰 발화를 생성해야함

⇒ in-context zero-shot transfer data aumentation의 형태로 이 문제를 formulate

: source domain classes와 그에 해당하는 예시를 프롬프트에 제공

(이때 소스 도메인 클래스는 target과 다름)

- **No/Low Code AI Paradigm**

  HyperCLOVA에 프롬프트 형식으로 주면 pipeline에 따라서 한 스텝으로 모델 구현 가능

Conclusion

82B 파라미터의 한국어 LM, HyperCLOVA 제안

→ in-context zero-shot & few-shot에 대한 SOTA 달성

+) 비개발자들이 쉽게 AI-backend product를 만들 수 있도록 하는 HyperCLOVA studio 제공

→ 한국에서 머신러닝에 친숙하지않은 사람들이 AI 모델을 만들 수 있도록돕는 시스템 구축하는 것이 목표