

What Changes Can Large-scale Language Models Bring?

☰ 태그	GPT-3 NL Naver
☰ 주차	1주차

▼ 용어 정리

corpus	말뭉치	
language0specific language	↔ multilingual languages model	
prompt-based learning		
zero-shot learning	이전에 학습된 모델을 사용해 라벨링 되지 않은 새로운 클래스에 대한 분류를 수행	https://velog.io/@nomaday/n-shot-learning
one-shot	하나의 샘플 이미지 만으로 새로운 클래스로 인식	
few-shot	일부 샘플 이미지 만으로 새로운 클래스로 인식	
transferability	어떠한 model에서 만들어진 adversarial examples를 다른 model에 적용했을 때도 비슷한 attack 성능을 보이는 성질	
morpheme	형태소	
Byte Pair Encoding(BPE)	데이터에서 가장 많이 등장한 문자열을 병합해서 데이터를 압축하는 기법	
out-of-vocabulary(OOV)	사전에 존재하지 않는 단어. 기계나 사람이 해석 불가능한 문자 혹은 문자의 집합	

1. Introduction

기존 GPT-3의 한계

1. 영어에만 92.7%가 치중되어 있음: 다른 언어로 모델을 적용시키는 것이 어려움
2. 소규모 혹은 대규모 LM의 분석에만 액세스 가능: 중간 사이즈의 LM 접근이 어려움
3. 입력 값의 역방향 그래디언트를 필요로 하는 고급 prompt-based learning 기법이 아직 대규모 LM 학습기에 대해 실험되지 않았음

이 논문이 기여하는 점

1. **HyperCLOVA**: 560B의 한국어 중심 말뭉치(*corpus) 생성 및 100B의 parameter를 활용한 대규모의 한국어 in-context learning-based LM 소개
2. 비영어 언어들의 *language-specific 토큰화의 영향력 소개
3. 중간 크기의 HyperCLOVA의 zero-shot(32B) 및 few-shot(82B)
4. prompt-based tuning이 최신 model들의 성능을 능가
5. No Code AI의 가능성 제기

2. Previous Work

2.1. Prompt Optimization

- full fine-tuning paradigm보다 *prompt-based learning 방식이 시간 효율성과 공간 복잡도에 있어서 훨씬 효율적임

discrete prompt optimization

- 토큰 공간 자체를 최적화 → *transferability ↑
- limitation
 - 해석 가능성(interpretability) ↓
 - 최적이지 아닐 수 있음(suboptimal)

continuous prompt optimization

- contextualized token space를 fine-tuning 없이 최적화

2.2. Language Models

Language-specific model

- 영어가 아닌 언어의 경우 cost가 높음
- in-context learners에 대한 연구는 몇몇 주요 언어에만 집중되어 있음
- 최근 중국어 말뭉치에 대한 GPT와 유사한 LM 등장

3. Pre-training

3.1. 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

- 최대한 많은 text data를 수집하기 위해 NAver와 외부 소스들을 통해 user-generated content(UGC) 및 외부 파트너에 의해 생성된 content들을 모두 수집
- 최종적으로 수집된 말뭉치는 561B의 토큰들로 구성

3.2. Model and Learning

# Param	n_{layers}	d_{model}	n_{heads}	d_{head}	lr
137M	12	768	16	48	6.0e-4
350M	24	1024	16	64	3.0e-4
760M	24	1536	16	96	2.5e-4
1.3B	24	2048	16	128	2.0e-4
6.9B	32	4096	32	128	1.2e-4
13B	40	5120	40	128	1.0e-4
39B	48	8192	64	128	0.8e-4
82B	64	10240	80	128	0.6e-4

Table 2: Detailed configuration per size of Hyper-CLOVA

- OpenAI의 GPT-3와 같이 transformer decoder 아키텍처 사용
- Goal: 중간 사이즈의 파라미터의 모델의 성능, 가능성 탐구

3.3. Korean Tokenization

교착어로서의 한국어

- 교착어(Agglutinative language): 어근에 접사(-았-, -겠- 등)가 결합해 의미가 변화하는 형태의 언어
- 영어와 달리 명사, 어간, 어미를 기준으로 토큰화 필요
→ **morpheme-aware byte-level *BPE** 사용

Tokenization

1. 자체 형태소 분석기: 띄어쓰기와 형태소를 기반으로 pre-split 수행. 한글 이외의 문자 제외
2. morpheme-aware byte-level BPE(HuggingFace library 이용): 한글 이외의 문자가 단일 바이트 문자들로 표현된 문장 학습

4. Experimental Results

4.1. Experimental Setting

Dataset

- **NSMC**: Naver Movies 영화 리뷰 데이터셋
- **KorQuAD 1.0**: 한국어 버전 기계독해(MRC, Machine Reading Comprehension) 데이터셋
- **AI Hub Korean-English**: 한국어-영어 문장. 뉴스, 정부 웹사이트, 법률서류 등
- **YNAT**(Yonhap News Agency Topic Classification, KLUE-TC): 연합뉴스 기사 제목으로 구성. 토픽 분류 시 이용
- **KLUE-STS**(Korean Language Understanding Evaluation Semantic Textual Similarity): 의미 유사도 분류 시 이용

4.2. In-context Few-shot Learning

	NSMC (Acc)	KorQuAD (EA / F1)		AI Hub (BLEU) Ko→En 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


- model size가 커짐에 따라 각 데이터셋에서 in-context learning 성능이 향상됨
- 한국어 → 영어 번역과 KLUE-STs에서는 baseline보다 낮은 성능을 보임
 - 한국어 → 영어 번역의 낮은 성능의 원인을 자체 말뭉치에서 영어의 낮은 비율 때문으로 추측

4.3. Prompt-based Training

Methods	Acc
Fine-tuning	
mBERT (Devlin et al., 2019)	87.1
w/ 70 data only	57.2
w/ 2K data only	69.9
w/ 4K data only	78.0
BERT (Park et al., 2020)	89.7
RoBERTa (Kang et al., 2020)	91.1
Few-shot	
13B 70-shot	87.9
39B 70-shot	88.0
82B 70-shot	88.2
p-tuning	
137M w/ p-tuning	87.2
w/ 70 data only	60.9
w/ 2K data only	77.9
w/ 4K data only	81.2
13B w/ p-tuning	91.7
w/ 2K data only	89.5
w/ 4K data only	90.7
w/ MLP-encoder	90.3
39B w/ p-tuning	93.0

- NSMC 데이터셋 기반으로 학습(4.1.에서 가장 좋은 성능)

기존 모델 vs Few-shot vs p-tuning

- in-context few-shot learning: 최신 모델의 성능과 유사한 성능
-  p-tuning: 중심 모델의 파라미터 업데이트 없이 기존 모델의 성능 증가

4.4. Effect of Tokenization

	KorQuAD (EA / F1)		AI Hub (BLEU) Ko→En 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

Ours vs Baseline

- 대부분의 데이터셋에서 baseline보다 높은 성능을 보임
- 한국어 → 영어에서는 오히려 더 낮은 성능을 나타냄

Byte-level BPE vs Char-level BPE

- char-level BPE이 byte-level BPE보다 훨씬 더 낮은 성능을 나타냄
- out-of-vocabulary(OOV): 하나의 형태소를 문자별로 split → 해석 불가능한 문자들 생성
⇒ char-level BPE의 낮은 성능

6. Conclusion

1. 다양한 수십억 규모의 한국어 중심 LM, HyperCLOVA 제안
2. 비개발자도 No COde AI paradigm 달성 가능
3. 머신러닝에 익숙하지 않은 사람들도 자신만의 AI 모델을 구현할 수 있는 생태계 조성