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

NAVER

CLOVA

Boseop Kim*, HyoungSeok Kim*, Sang-Woo Lee*, Gichang Lee, Donghyun Kwak, Dong Hyeon Jeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, Nako Sung

NAVER CLOVA, AI Lab, Search

Main Contribution

- Introducing **HyperCLOVA**, a large-scale Korean in-context learning-based LM with 82B parameters, by constructing a large Korean-centric corpus of 560B tokens.
- Discovering the effect of language-specific tokenization on large-scale in-context LMs for training corpus of non-English languages.
- We explore the zero-shot and few-shot capabilities of mid-size HyperCLOVA with 39B and 82B parameters and find that prompt-based tuning can enhance the performances, outperforming state-of-the-art models on downstream tasks when backward gradients of inputs are available.
- We argue the possibility of realizing No Code AI by designing and applying HyperCLOVA Studio to our in-house applications. We will release HyperCLOVA Studio with input gradients, output filters, and knowledge injection.

HyperCLOVA: Korean Hyperscale LM

- 82B Transformer decoder with sparse Transformer.
- 560B tokens of Korean-centric corpus.
- Korean: 97%, English: 1%, Japanese: 1%, etc: 1%.
- 13.4 days on 150B tokens among 560B tokens training of 82B model with 1,120 A100 GPUs.
- Korean-specific tokenization: morpheme-aware byte-level BPE.

Name	Description	Tokens (B)
Blog	Blog corpus	273.6
Café	Online community corpus	83.3
News	News corpus	73.8
Comments	Crwaled comments	41.1
KiN	Korean Social QnA websites	27.3
Modu	Collection of five Korean datasets	6.0
WikiEn, WikiJP	Foreign wikipedia	5.2
Others	Others corpus	51.5
Total		561.8

Table 1. Description of pre-training corpus for HyperCLOVA

In-context Learning

In context few-shot learning

	NSMC (SC)	KorQuAD (MRC)		AI Hub (Translation)		YNAT (TC)	KLUE-STS
				Ko → En	En → Ko		
Metrics	Acc	EM	F1	BLEU	BLEU	F1	F1
Baselines	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	50.89
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 2. Results of in-context learning tasks. Baseline refers to BERT-base or Transformer-base. 137M ~ 82B denotes the size of the corresponding model.

Ablation study on tokenization

	KorQuAD (MRC)		AI Hub (Translation)		YNAT (TC)	KLUE-STS
Methods	EM	F1	Ko->En	En->Ko	F1	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 3. Effects of tokenization approaches. HyperCLOVA-1.3B is used for evaluation. Our morpheme-aware byte-level BPE performs well in most cases.

Effects of P-tuning

NSMC

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 w/ 70-shot	87.9
39B w/ 70-shot	88.0
82B w/ 70-shot	88.2

Methods	Acc
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/ 70 data only	89.5
w/ 2K data only	90.7
w/ 4K data only	90.3
39B w/ p-tuning	**93.0**

Table 4. Comparison results of p-tuning with fine-tuned LMs and in-context few-shot learning on NSMC.

Query modification task

Example 1:
User: Play IU's track
AI Speaker: I am playing the track.
User: How old?
Modified query: How old is IU?

Example 2:
User: Who invented airplane?
AI Speaker: Wright brothers did.
User: What is the younger's name?
Modified query: What is the younger one's name of Wright brothers?

Sizes	Few-shots	P-tuning	BLEU
13B	0-shot	X	36.15
		O	**58.04**
	3-shot	X	45.64
		O	**68.65**
39B	0-shot	X	47.72
		O	**73.80**
	3-shot	X	65.76
		O	**71.19**

Table 5. Data examples and experimental results of query modification task.

HyperCLOVA Studio & API

- We made HyperCLOVA Studio and API, which is the GUI and CUI interface of HyperCLOVA.
- Various functionality exists including input gradients, output filters, and knowledge injection.

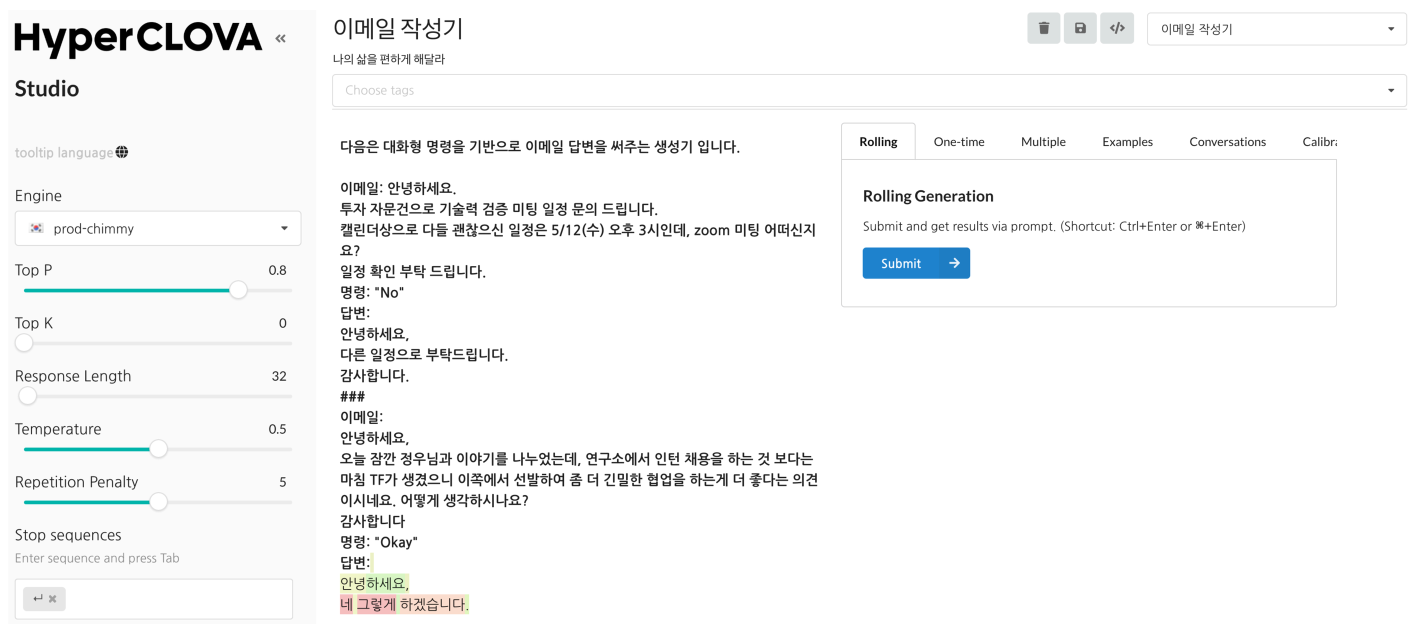

Figure 3. An example interface of HyperCLOVA Studio.

Discussion on No/Low Code AI

- For making AI products, many types of AI experts are required to make an AI product.
- It also makes huge communication overhead between them.
- With GUI and CUI interface of LM, one user can quickly do a problem definition, curating a few examples, and error monitoring altogether.
- GUI interface can also be used by people who do not know AI well.
- We show in-house usage of HyperCLOVA Studio, showing the effectiveness of using in-context learner language model.

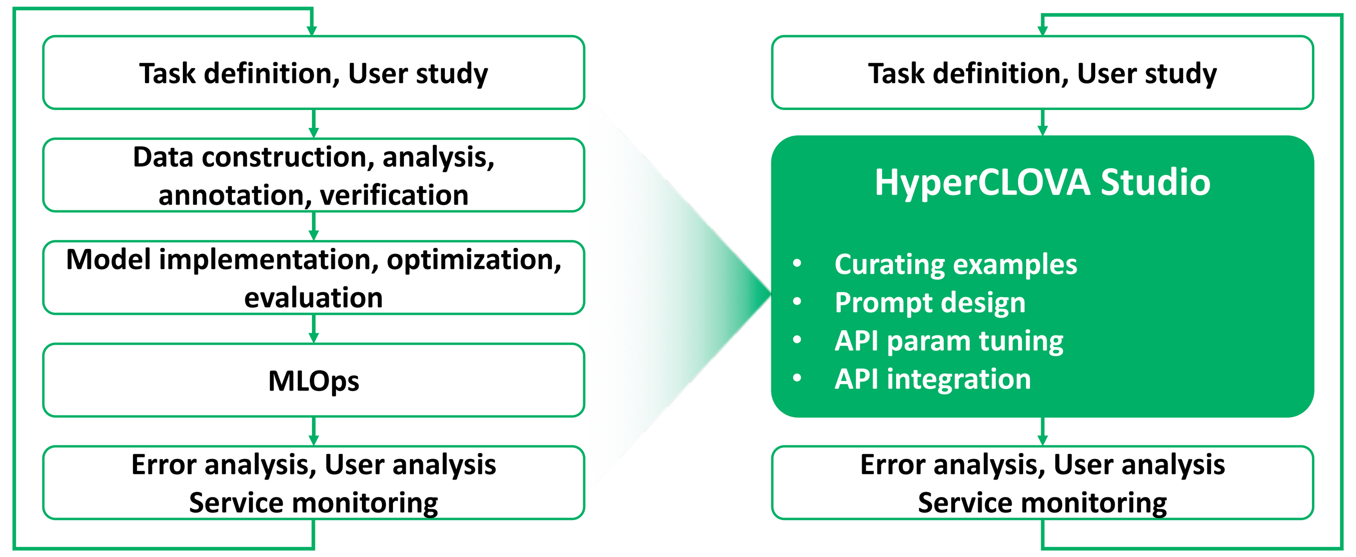

Figure 4. No/Low Code Ai paradigm in HyperCLOVA Studio.

Example In-house Applications using HyperCLOVA Studio

Zero-shot transfer data augmentation

intent: Reservation inquiry

#1: Will it be reserved for a room?
#2: If you don't have a seat, it's okay to sit outside.
#3: Is it possible for a group seat?
#4: What is the most common menu for dinner?
#5: You want to make a reservation for the weekend, but do you have a lot of customers?
#6: Do I have to order by number of people?

Zero-shot (Acc)

	Number of augmented samples (k)				
n	5(1)	10(2)	15(3)	25(5)	125(3)
0(0)	$60.8_{9.3}$	$68.9_{4.0}$	$71.9_{2.7}$	$74.8_{2.5}$	$78.0_{2.3}$

Few-shot (Acc)

	Number of original samples (n)				
k	1(1)	2(1)	3(1)	4(1)	5(1)
0(0)	$26.8_{6.0}$	$52.0_{4.9}$	$64.7_{5.2}$	$76.5_{4.4}$	$83.0_{3.0}$
25(5)	$79.2_{2.5}$	$81.2_{2.5}$	$82.6_{2.6}$	$83.4_{1.9}$	$84.3_{2.0}$
125(30)	$80.7_{2.2}$	$82.7_{1.9}$	$83.7_{2.1}$	$86.3_{1.5}$	$87.2_{1.7}$

Figure 1. Zero-shot transfer data augmentation task. in 20-class classification of zero-shot learning, the performance reaches near 80.

Event title generation task

tag: Toggle Bar Necklace, Half and Half Chain Necklace, Cubic Earrings, Gemstone Earrings, Drop Earrings, One Touch Ring Earrings, Chain Silver Ring, Onyx Earrings, Pearl Earrings, Heart Earrings
time: December 19th

Title: Jewelry for you who shines brightly

Comparison	BLEU	Win	Lose	Tie
mT5 vs. GT	13.28	0.311	**0.433**	0.256
HyperCLOVA vs. mT5	N/A	**0.456**	0.350	0.194
GT vs. HyperCLOVA	5.66	0.311	0.333	0.356

Figure 2. Task of generating advertisement event titles. It takes less than 10 minutes of designers' effort to make a prompt of few-shot examples to use HyperCLOVA.

The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), Online and in the Barceló Bávaro Convention Centre, Punta Cana, Dominican Republic