

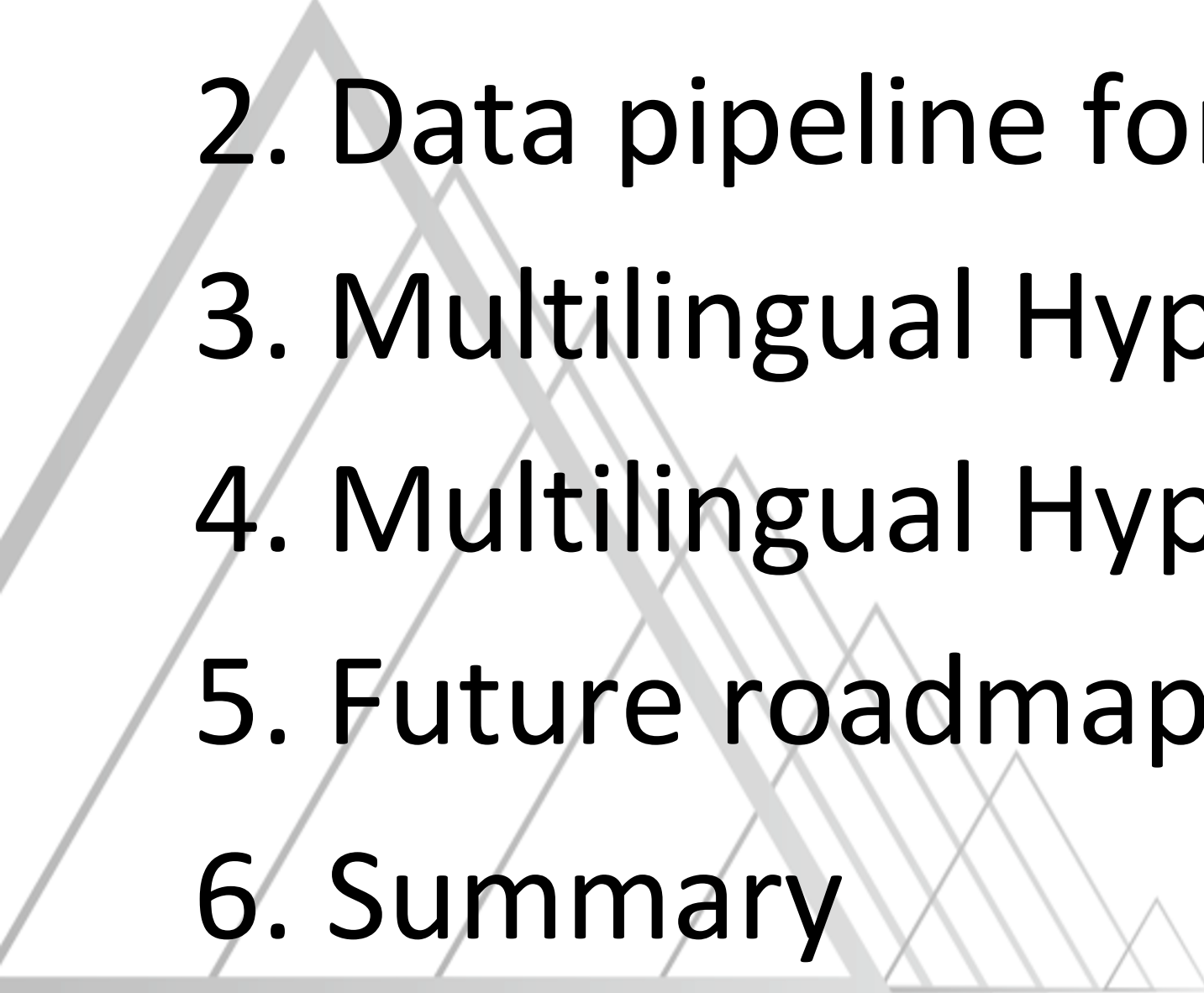



Multilingual HyperCLOVA



신성진, 김보섭 NAVER CLOVA

CONTENTS

- 
1. Why do we need Multilingual HyperCLOVA?
 2. Data pipeline for Multilingual HyperCLOVA
 3. Multilingual HyperCLOVA Model Experiments
 4. Multilingual HyperCLOVA Model Evaluation
 5. Future roadmap
 6. Summary
- 

1. Why do we need Multilingual HyperCLOVA?

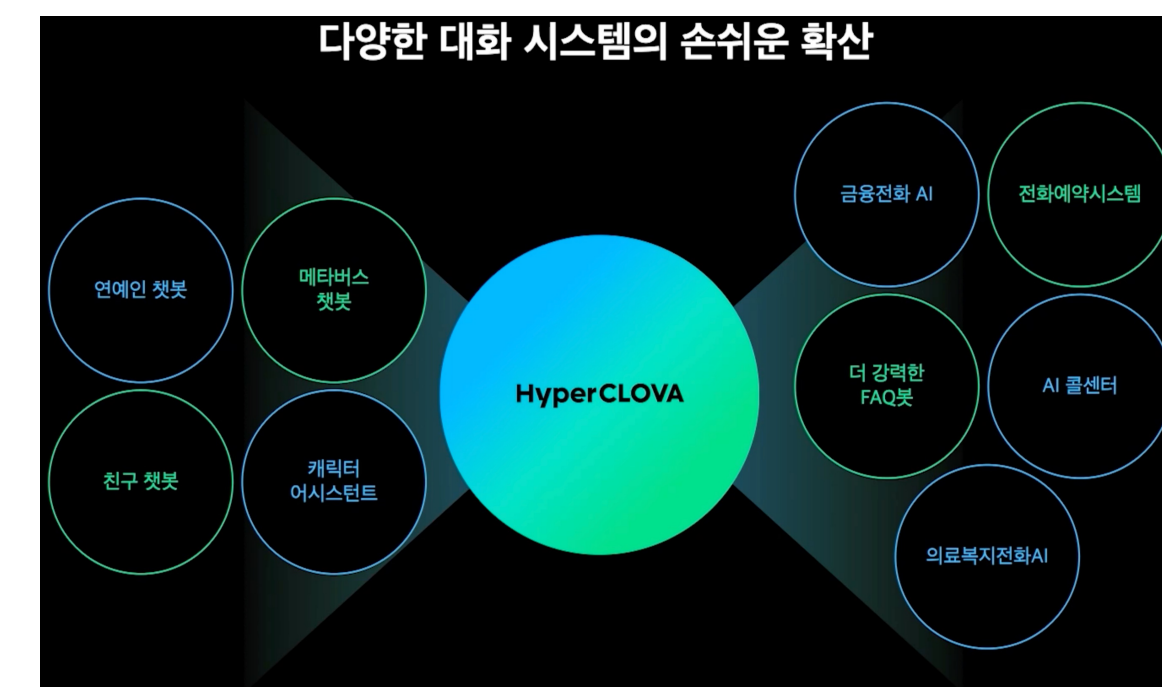
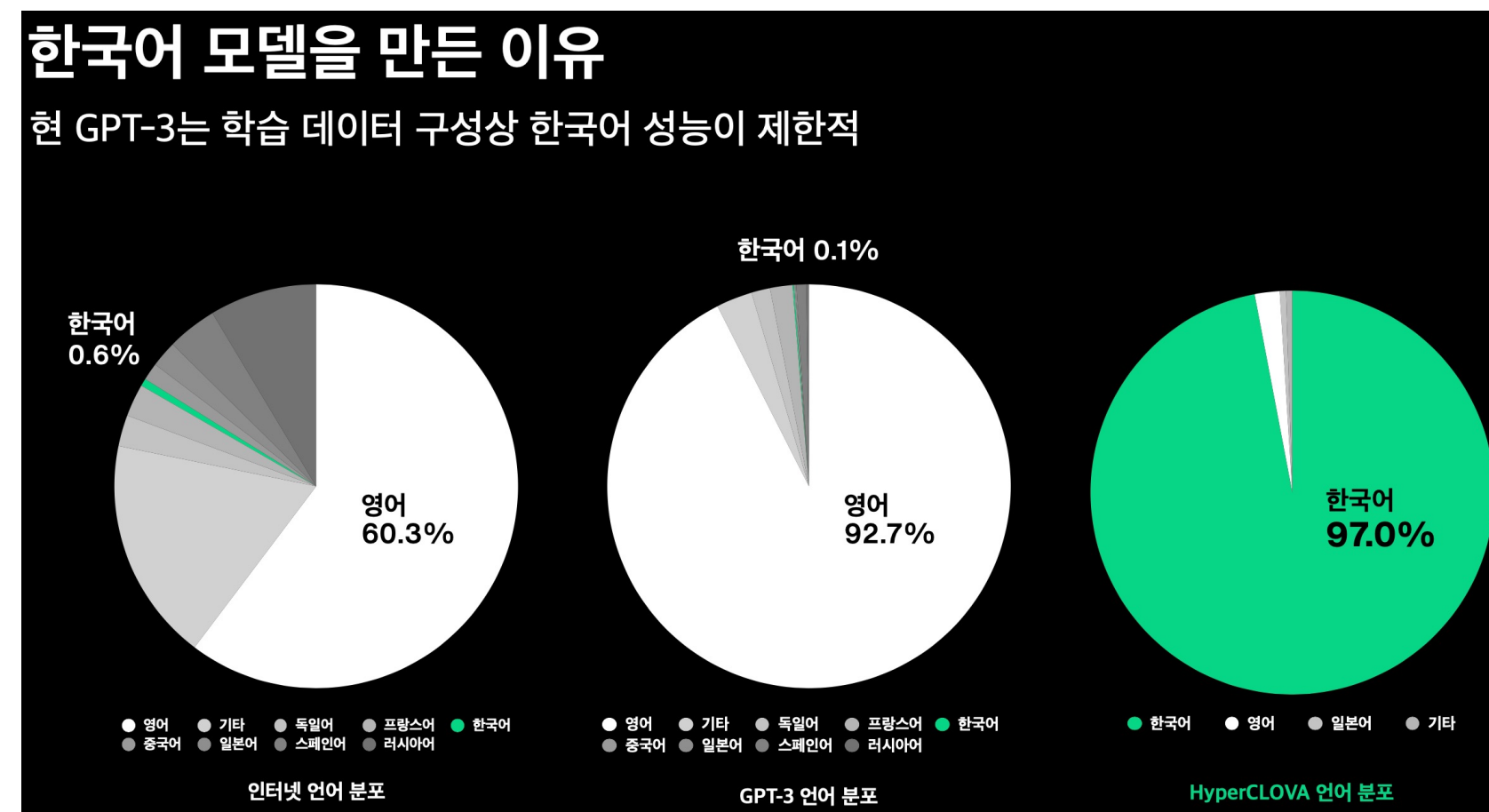
1.1 Overview: HyperCLOVA

HyperCLOVA (1/2)

- 한국어 위주로 구성된 말뭉치로 학습
- 언어모델의 성능을 측정하는 벤치마크 뿐만 아니라, 다양한 어플리케이션에서 초거대 한국어 언어모델의 가능성을 확인

	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

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).



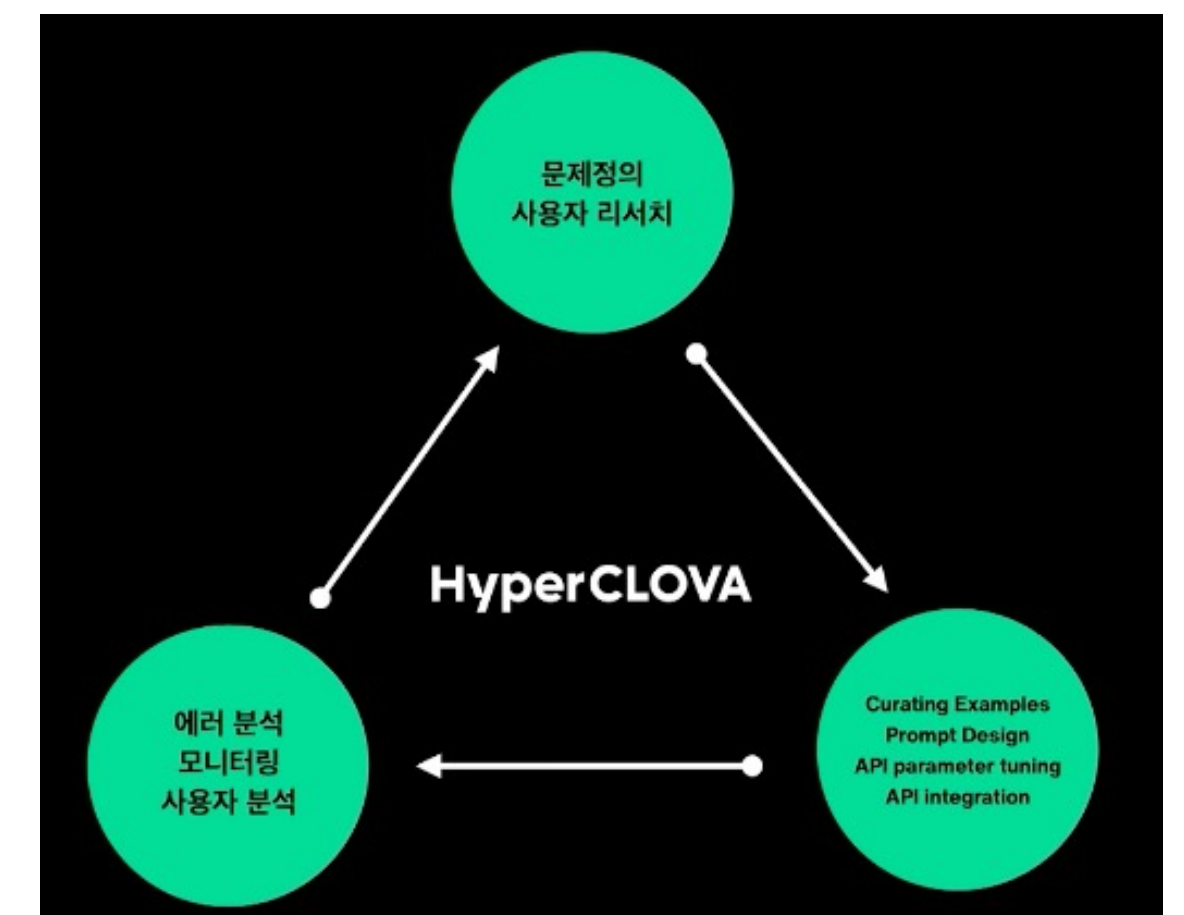
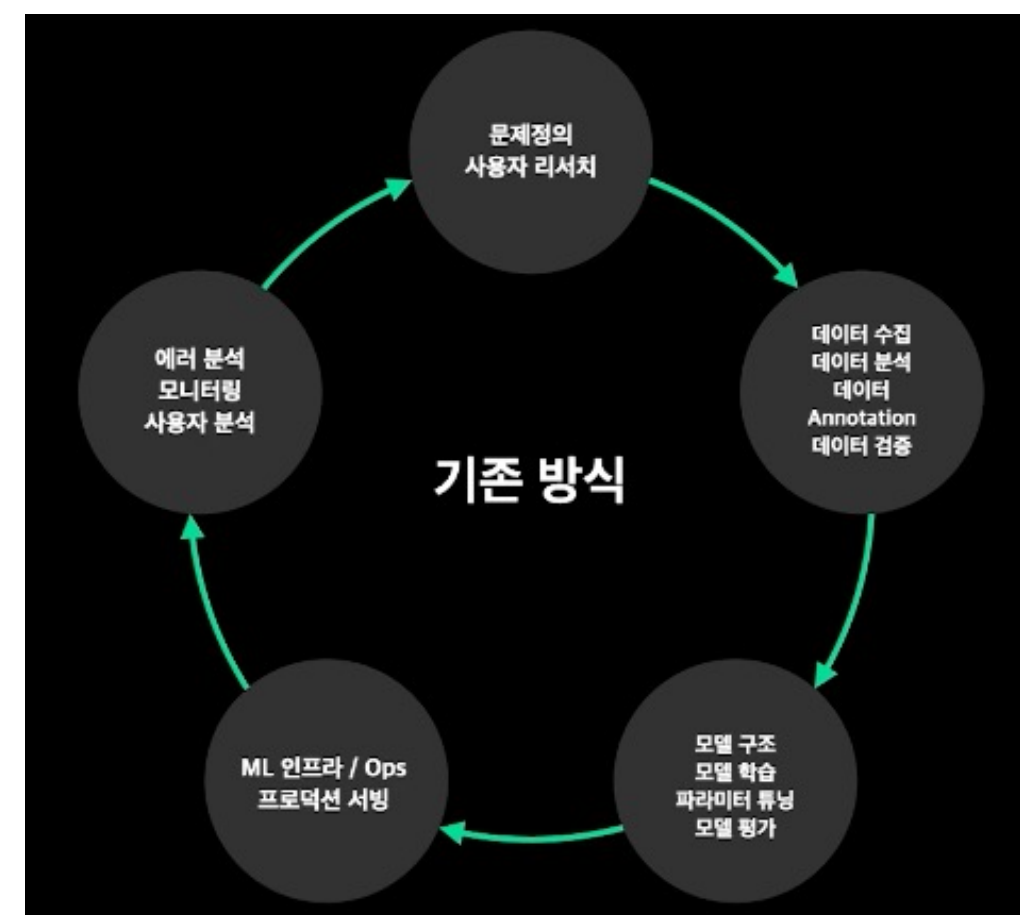
1.1 Overview: HyperCLOVA

HyperCLOVA (2/2)

- 연구 측면에서 GPT3에서 제시하지않은 중간 크기의 초거대 언어모델들에 대한 검증
- 초거대 언어모델을 활용한 “No Code AI Paradigm”을 제시

# 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 HyperCLOVA



1.2 Why Multilingual?

Motivation (1/5)

- HyperCLOVA 모델의 크기가 증가할 수록 성능 증가가 둔화되는 경향을 관찰
- 증가한 파라미터를 효율적으로 활용하지 못하고 있음이 추정되는 상황 → 학습에 활용되는 말뭉치를 다양하게 늘리는 전략이 필요

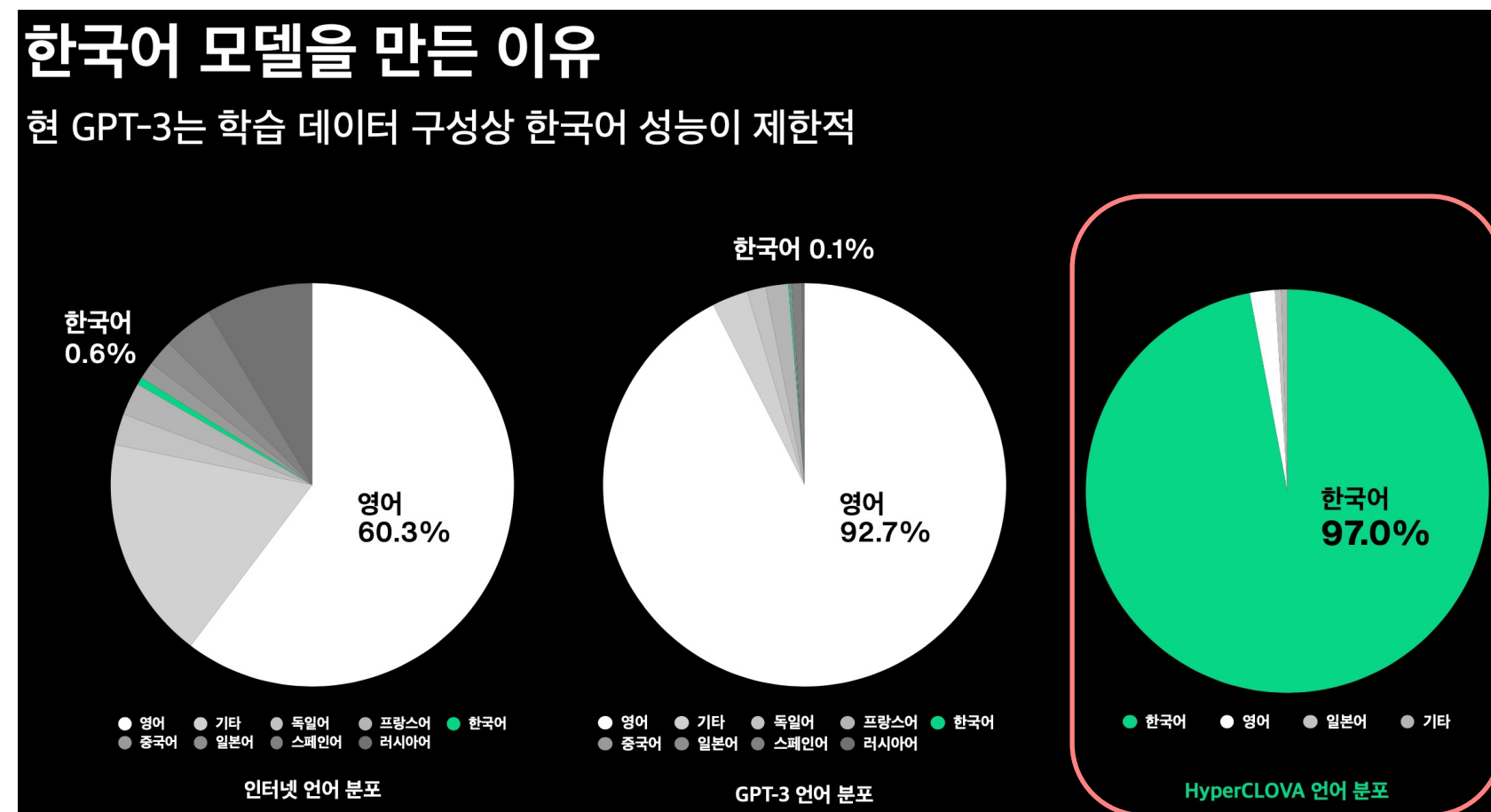
	NSMC (Acc)	KorQuAD (EA / F1)		AI Hub (BLEU) Ko→En En→Ko		YNAT (F1)	KLUE-STS (F1)
Baseline	89.66	74.04	86.66	40.34	40.41	82.64	75.93
137M	73.11	8.87	23.92	0.80	2.78	29.01	59.54
350M	77.55	27.66	46.86	1.44	8.89	33.18	59.45
760M	77.64	45.80	63.99	2.63	16.89	47.45	52.16
1.3B	83.90	55.28	72.98	3.83	20.03	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

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. \(2021\)](#).

1.2 Why Multilingual?

Motivation (2/5)

- "한국어 → 영어" 번역의 성능이 "영어 → 한국어" 번역 성능에 비해 낮음 상황
- 한국어 위주의 말뭉치로만 학습했기때문에, 영어로 말하는 능력이 부족하다고 판단



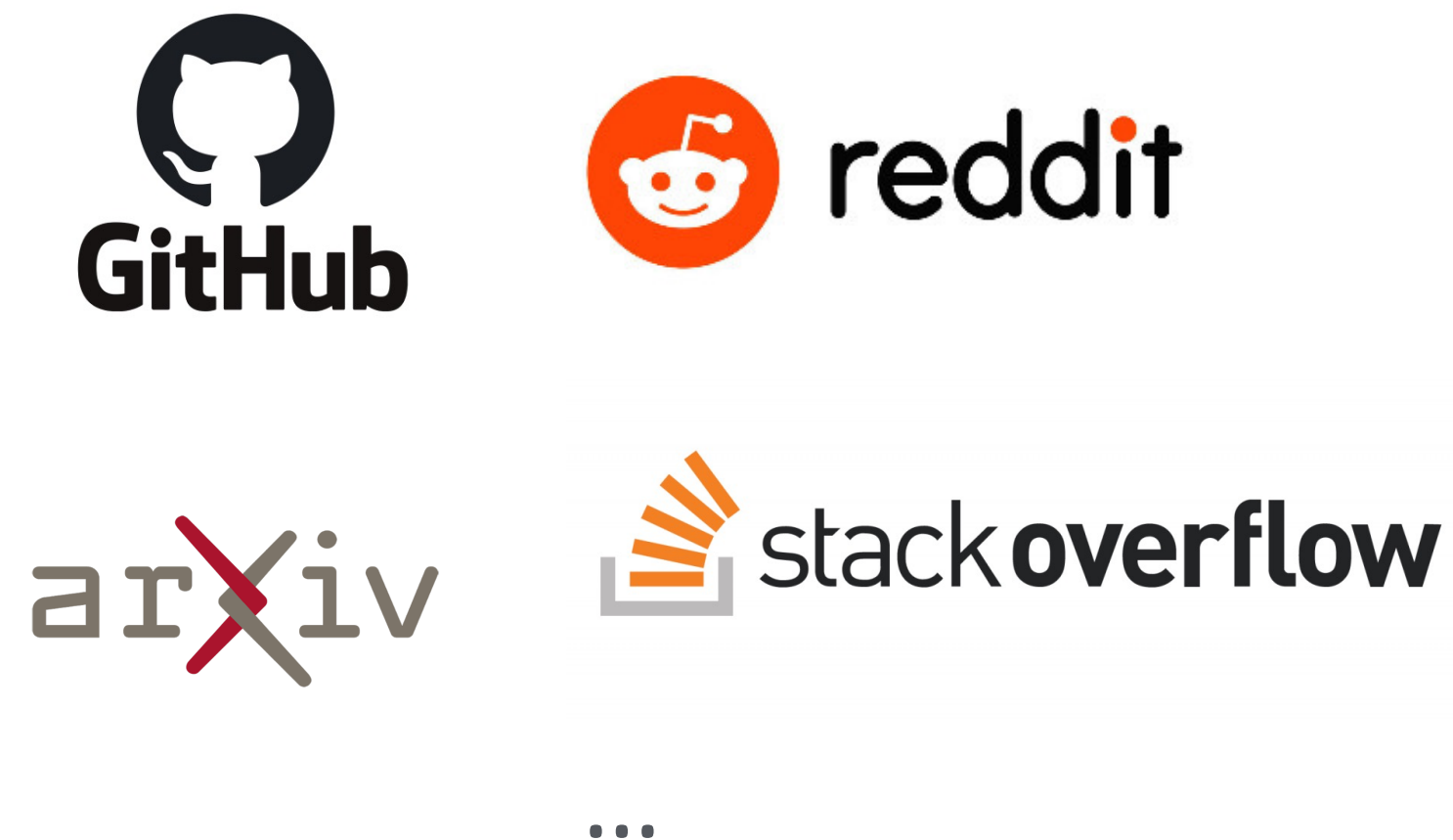
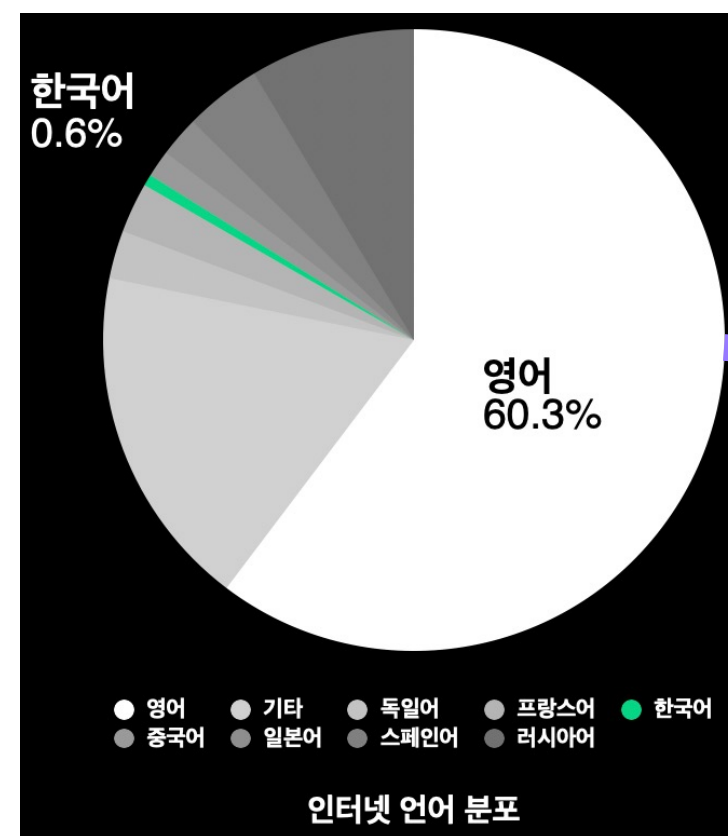
	NSMC (Acc)	KorQuAD (EA / F1)	AI Hub (BLEU)		YNAT (F1)	KLUE-STS (F1)
			Ko→En	En→Ko		
Baseline	89.66	74.04 86.66	40.34	40.41	82.64	75.93
137M	73.11	8.87 23.92	0.80	2.78	29.01	59.54
350M	77.55	27.66 46.86	1.44	8.89	33.18	59.45
760M	77.64	45.80 63.99	2.63	16.89	47.45	52.16
1.3B	83.90	55.28 72.98	3.83	20.03	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).

1.2 Why Multilingual?

Motivation (3/5)

- 타 언어의 말뭉치를 한국어와 말뭉치를 동시에 학습하여 HyperCLOVA를 multilingual HyperCLOVA로 확장
- 지식들은 개별 언어에 파편적으로 존재 → 초거대 언어모델이 지식이 축적된 방식(e.g. 영어)에 상관없이 지식을 이해할 수 있게 만들 수 있을까?



1.2 Why Multilingual?

Motivation (4/5)

- 연구적 측면에서 기존에도 다국어 확장에 대한 많은 도전은 있었음.
- 하지만 서양 언어간으로 연구가 한정되어 있었으며, 지식 전이의 측면이 고려되지 않음. → 본 연구에서는 지식전이의 측면까지 고려

As Is

<monolingual HyperCLOVA - korean>

Q: 나전칠기에 대해서 알려줘
A: 조개껍데기를 얇게 갈아 여러 가지 모양으로 오려낸 것을 말해. 자개라고도 하지. 고려시대부터 사용하기 시작했어. 귀족들이 많이 사용했지. 조선시대에는 더욱 발전해서 왕실이나 상류층에서 주로 사용했어. 우리나라의 대표적인 공예품이야.

<monolingual HyperCLOVA - english>

Q: What is Raden?
A: It's a type of Indonesian food.



To Be

<multilingual HyperCLOVA>

Q: 나전칠기에 대해서 알려줘
A: 조개껍데기를 얇게 갈아 여러 가지 모양으로 오려낸 것을 말해. 자개라고도 하지. 고려시대부터 사용하기 시작했어. 귀족들이 많이 사용했지. 조선시대에는 더욱 발전해서 왕실이나 상류층에서 주로 사용했어. 우리나라의 대표적인 공예품이야.

Q: What is Raden?
A: Say that the shells are thinly ground and cut into various shapes. Also called "mother-of-pearl". It has been used since the Goryeo Dynasty. Used a lot by the nobility. During the Joseon Dynasty, it developed further and was mainly used by the royal family and upper classes. It is a representative craft of Korea.

1.2 Why Multilingual?

Motivation (5/5)

- 서비스 측면에서 아래와 같은 사항을 기대할 수 있음.
 - 1) 한국어 활용 사례들을 다국어로 확장
 - 2) 하나의 통합된 모델만 학습함으로써 전처리 및 학습시간 단축
 - 3) 서빙해야하는 모델의 개수 감소로 모델 Inference 리소스 절약

1.3 Goal

Multilingual HyperCLOVA

- task-specific한 학습없이 다양한 언어의 말뭉치를 사용하는 GPT3 like pre-training으로 언어의 경계를 넘어설 수 있는 지 가능성을 확인하는 것이 목표
- 초기에는 한국어, 영어 bilingual로 가능성 검증을 시작

2.Data pipeline for Multilingual HyperCLOVA

2.1 Data pipeline의 필요성

논문에서 얘기하지 않는 전처리 (1/2)

- GPT3 논문에서는 대략 570GB text data로부터 말뭉치 별로 sampling하여 하나의 데이터셋을 구성, 이를 이용하여 학습을 수행함.

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Table 2.2: Datasets used to train GPT-3. “Weight in training mix” refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

2.1 Data pipeline의 필요성

논문에서 얘기하지 않는 전처리 (2/2)

- 데이터셋을 만드는 프로세스를 두 가지 프로세스로 나눔 → Mix, Serialize
- Mix: 영역별 말뭉치를 섞어서 하나의 학습용 말뭉치로 추출
- Serialize: 학습용 말뭉치를 학습에 사용할 수 있는 데이터셋으로 변형

Korean corpus

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

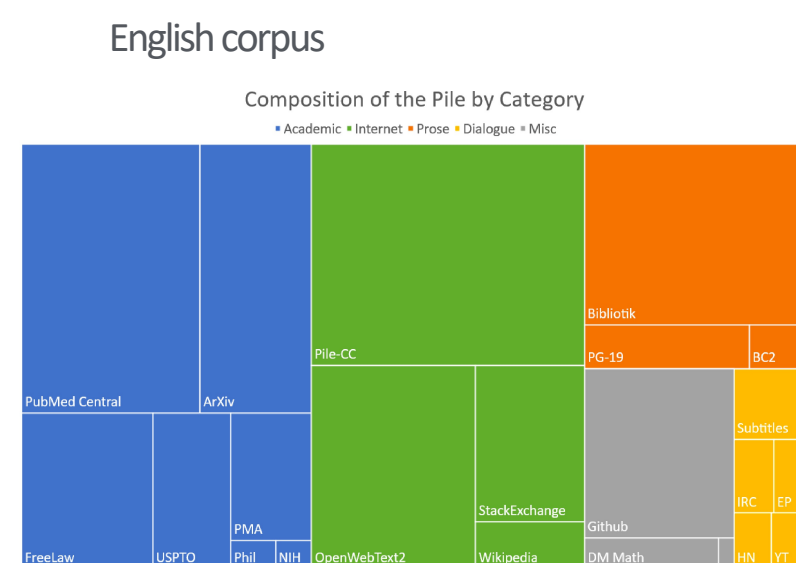


Figure 1: Treemap of Pile components by effective size.

Mix



Serialize



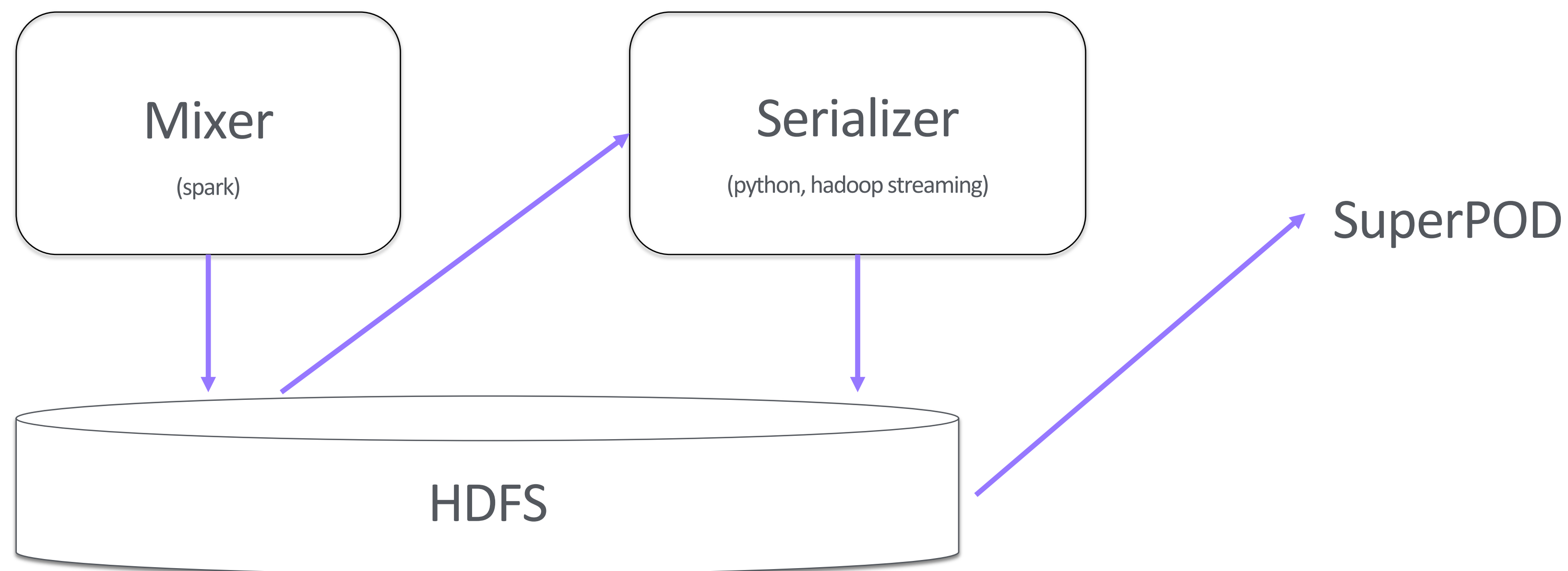
Kim, B., Kim, H., Lee, S. W., Lee, G., Kwak, D., Jeon, D. H., ... & Sung, N. (2021). What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers. arXiv preprint arXiv:2109.04650.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., ... & Leahy, C. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv preprint arXiv:2101.00027.

2.2 Data pipeline 구조

Mixer & Serializer

- Mix, Serialize을 hadoop cluster 상에서 진행하는 것을 고안
- Mixer, Serializer를 고안, 빠른 전처리가 가능하게 됨.



2.2 Data pipeline 구조

Mixer (1/2)

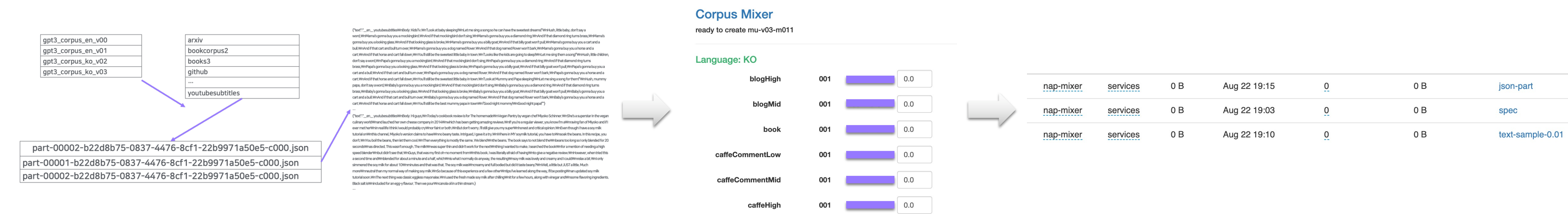
- HDFS 상에 한국어, 영어 말뭉치를 JSONL 형태 chunk 단위로 나누어 적재



2.2 Data pipeline 구조

Mixer (2/2)

- 최소한의 전처리를 수행한 뒤 말뭉치를 mixing하여 학습용 말뭉치를 생성, sampling을 수행한 샘플 말뭉치를 Serializer에 전달

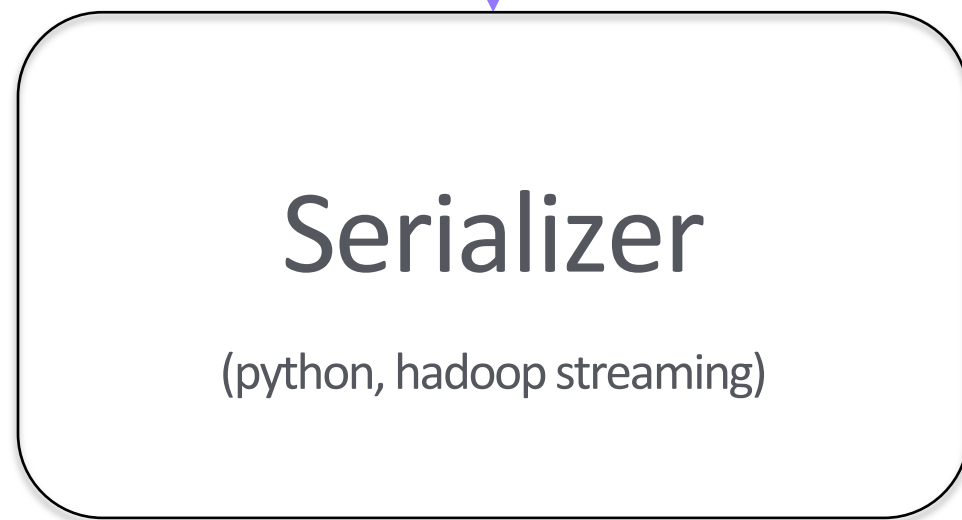


2.2 Data pipeline 구조

Serializer (1/2)

- 샘플 말뭉치를 이용하여 subword tokenizer를 학습
- subword tokenizer로 학습용 말뭉치를 hadoop streaming을 이용하여 serialize

nap-mixer	services	0 B	Aug 22 19:15	0	0 B	json-part
nap-mixer	services	0 B	Aug 22 19:03	0	0 B	spec
nap-mixer	services	0 B	Aug 22 19:10	0	0 B	text-sample-0.01



nap-serializer	services	9.01 GB	Aug 22 20:49	3	256 MB	container_e119_1625191536274_7258769_01_000009_text_document.bin
nap-serializer	services	90.57 MB	Aug 22 20:49	3	256 MB	container_e119_1625191536274_7258769_01_000009_text_document.idx
nap-serializer	services	9.02 GB	Aug 22 20:54	3	256 MB	container_e119_1625191536274_7258769_01_000010_text_document.bin
nap-serializer	services	90.57 MB	Aug 22 20:54	3	256 MB	container_e119_1625191536274_7258769_01_000010_text_document.idx

2.2 Data pipeline 구조

Serializer (2/2)

- subword tokenizer를 학습에는 huggingface의 tokenizers를 사용
- 샘플 말뭉치도 대용량으로 subword-tokenizer 학습에 문제 발생 → 전처리로 해결

20G로 서브워드 토크나이저를 학습할 때, 메모리 부족(Out of Memory)과 같은 이슈들이 발생
→ 전처리를 통해 해결

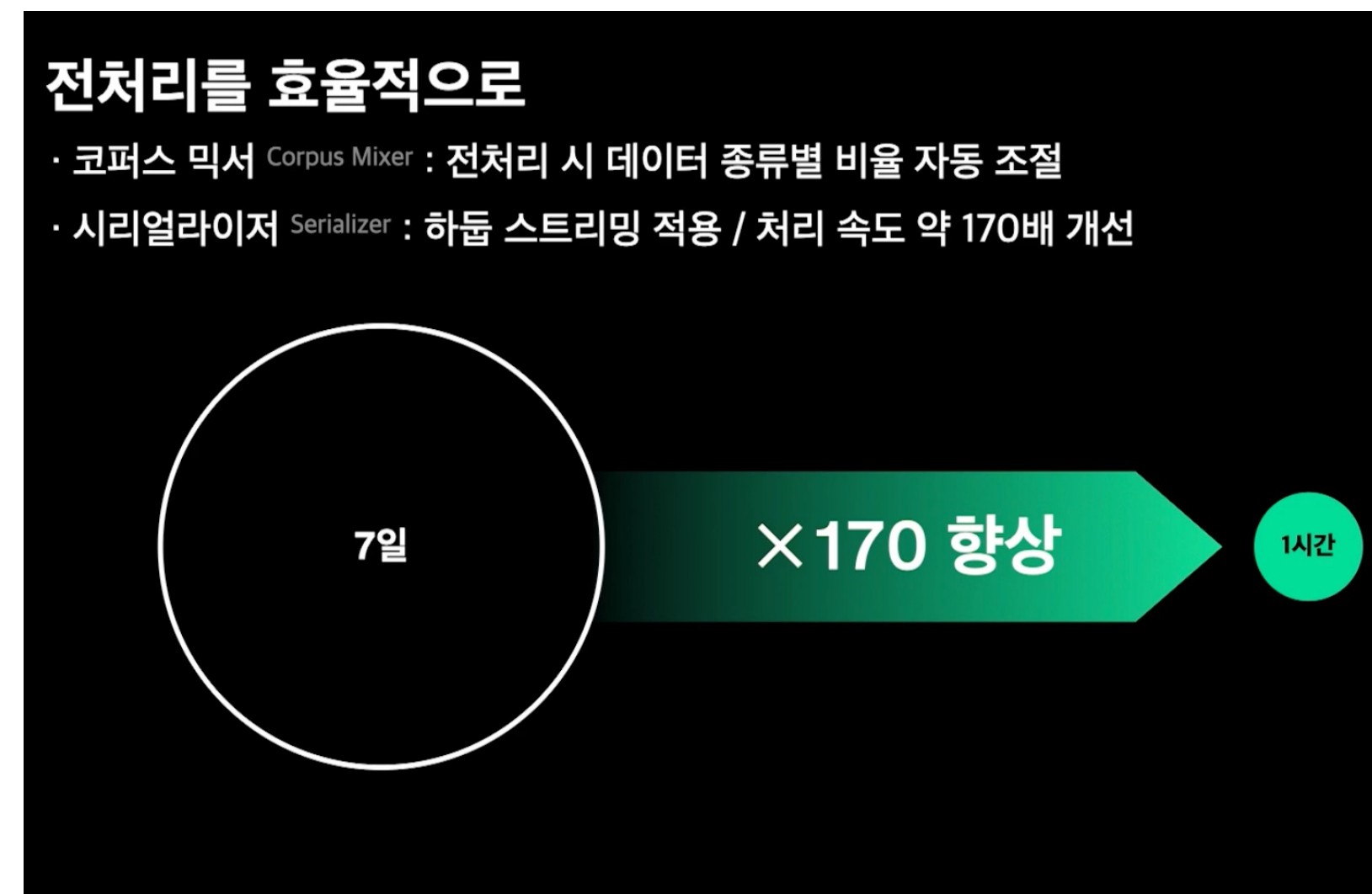
The image shows a GitHub issue comment on the left and a diagram on the right. The comment, from user 'Narsil' on Sep 16, 2020, discusses memory issues with BPE tokenizers and lists four solutions: 1. Adding a PrefixTokenizer to split sentences into groups. 2. Using SentencePiece for languages without spaces. 3. Waiting for a future version (8.9) with a lower memory footprint. 4. Seeking help for bleeding edge solutions. The diagram on the right illustrates three tokenization methods: 'sample_tokenized_by_space_for_100k.txt', 'sample_tokenized_by_space_for_1M.txt', and 'sample_tokenized_by_nltk_for_1M.txt', showing how they break down into individual tokens.

출처 : <https://github.com/huggingface/tokenizers/issues/422#issuecomment-694750093>

2.2 Data pipeline 결과

실험을 위한 토대 마련

- Data pipeline 구성으로 기존 대비 대략 170배 정도 전처리 속도를 개선함.
- 이를 기반으로 다양한 실험을 수행할 수 있었음.





Multilingual HyperCLOVA



신성진, 김보섭 NAVER CLOVA

3. Multilingual HyperCLOVA Model Experiments

3.1 Related Works

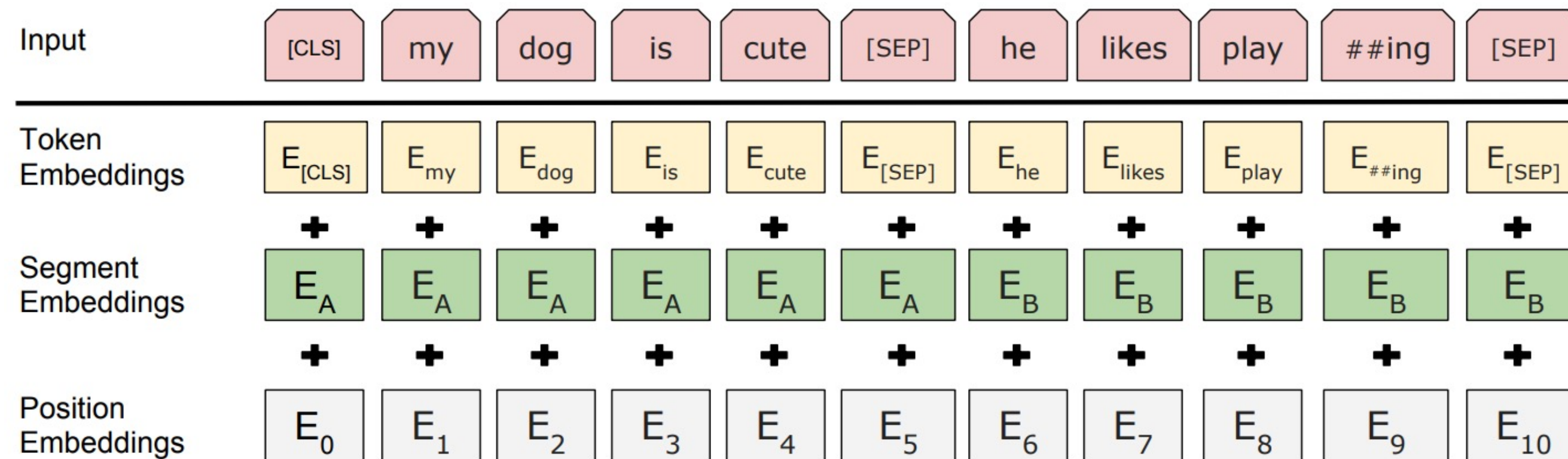
다국어 HyperCLOVA를 위한 첫 걸음

- Google, Facebook 등 전 세계 대상 서비스를 제공하는 회사에서 연구 진행
- 다국어 연구는 기계번역 (NMT)에서 많은 연구
- 최근 Pre-trained 계열에서 관심을 가지기 시작
- HyperCLOVA의 기반이 되는 Language Model 쪽에는 아직 사례가 거의 없음
- 다국어 Pre-trained 모델 소개

3.1 Related Works

mBERT

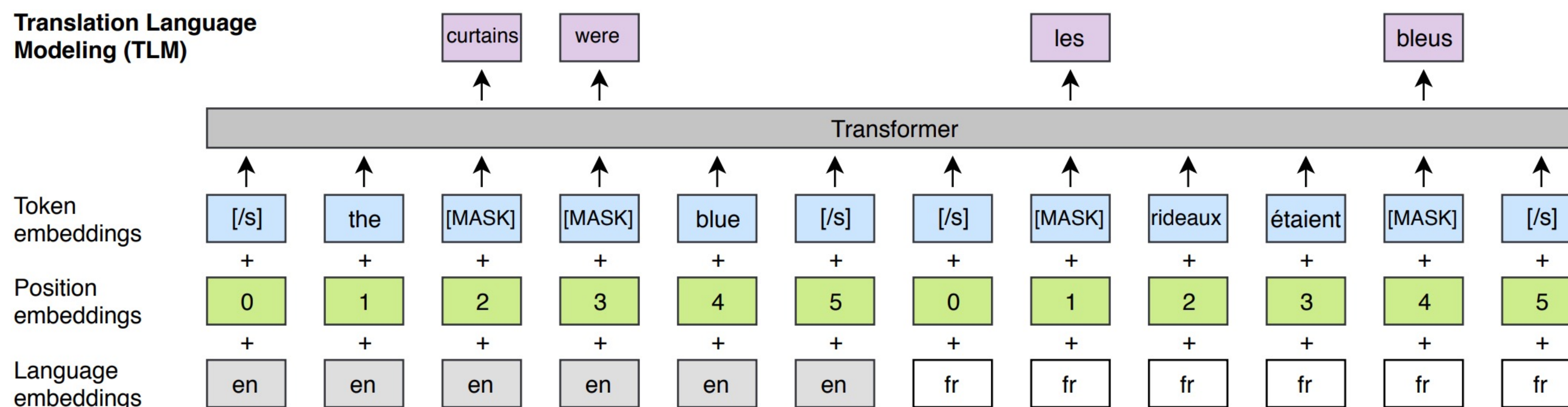
- 기존 Transformer Encoder를 사용한 BERT 구조를 활용
- 104개 언어 데이터를 110K WordPiece Vocab으로 토큰화
- 언어의 선정 기준은 위키피디아 언어에서 가장 많은 데이터 순으로 선정
- 데이터 불균형을 해결하기 위해 각 언어별 비율 조정 진행



3.1 Related Works

XLM

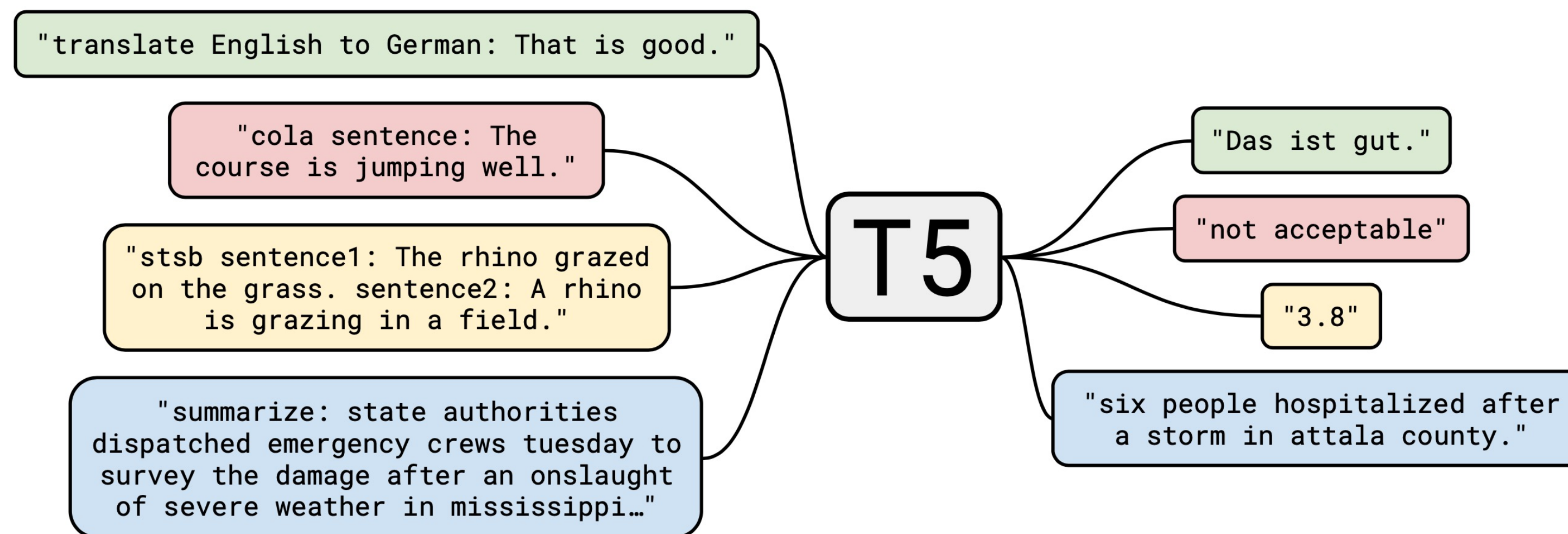
- BERT와 같은 Transformer Encoder 기반의 모델
- Masked Language Model (MLM) + Translation Language Model (TLM)
- TLM을 위한 Parallel 데이터 (번역 데이터) 필요
- 80K Vocab Byte Pair Encoding 활용 토큰화



3.1 Related Works

mT5

- Transformer의 Encoder-Decoder 구조 활용
- Common Crawl 기반의 데이터셋으로 101개 언어 학습
- 250K Vocab SentencePiece 활용 토큰화

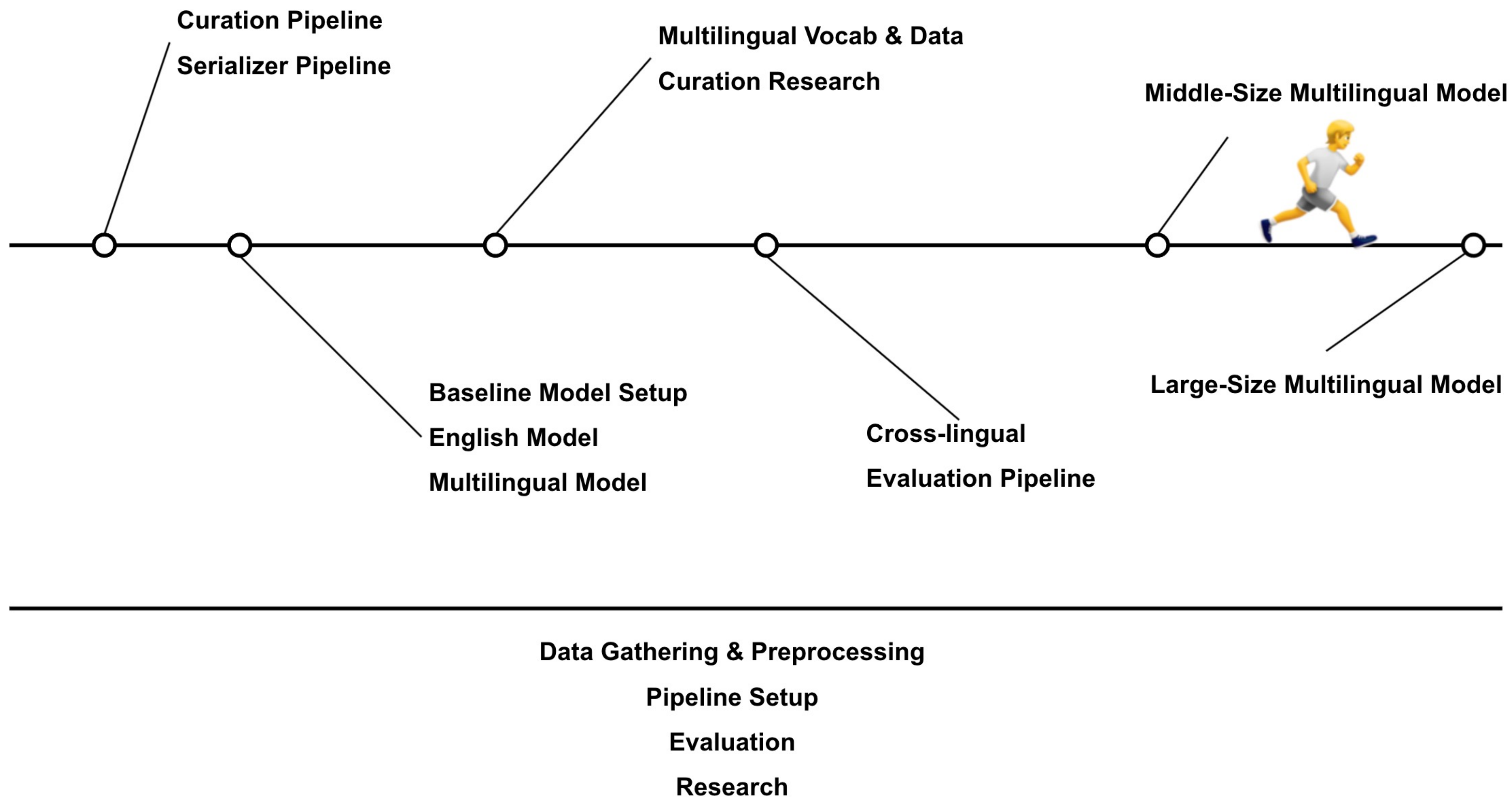


3.1 Related Works

Our Model

- 기존의 연구는 트랜스포머 Encoder 또는 Encoder + Decoder의 양방향 구조
- Multilingual HyperCLOVA는 Decoder Only 구조의 Pre-training 활용
 - 기존 GPT3 모델의 Multi-lingual 확장판
- 한국어 + 영어 모델에 대한 다양한 실험을 통한 Insight 및 언어 확장

3.2 Plan Overview



3.3 Model Preparation

Data Corpus

- 한/영 말뭉치

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

[HyperCLOVA Corpus](#)

Composition of the Pile by Category

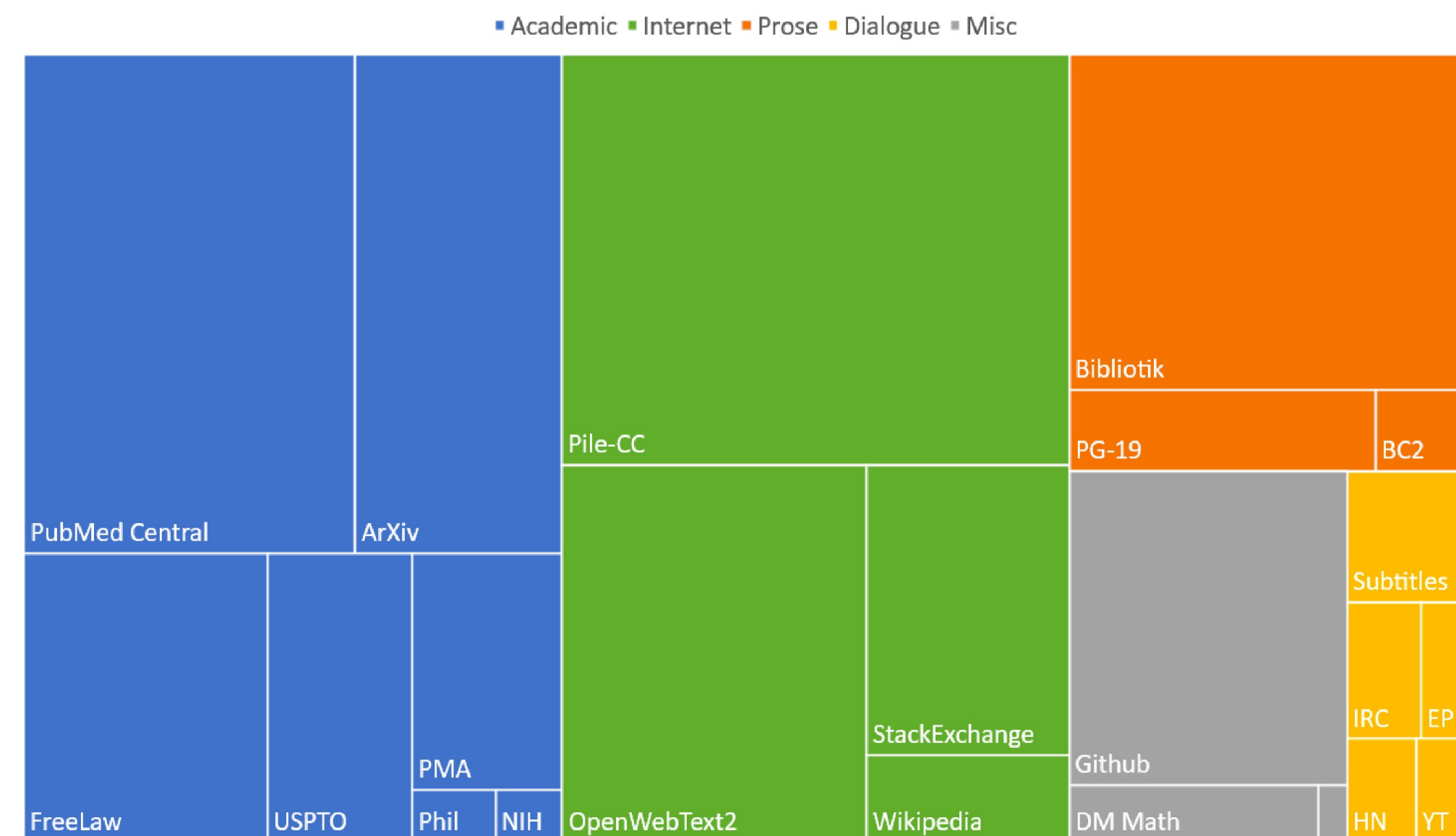


Figure 1: Treemap of Pile components by effective size.

[Pile Corpus](#)

3.3 Model Preparation

Multilingual Tokenization (1/4)

- Monolinugal HyperCLOVA (기존 NAVER AI NOW 발표 내용)
 - morpheme-aware ByteLevelBPETokenizer 사용
- Multilingual HyperCLOVA
 - 한국어와 영어를 함께 pre-tokenize를 space로만 해서 subword-tokenizer 학습
- 50K Vocab BPE 적용

3.3 Model Preparation

Multilingual Tokenization (2/4)

- 한국어 BBPE Vocab

한글 기사 - 토큰 수: 150

[삼, '라만, '상, '에, ' ', '붉은, ' ', '물을, ' ', '들, '이며, ' ', '침, '잠, '하는, ' ', '불, '덩어리, '를, ' ', ' ', '형형색색, '의, ' ', '원형, ' ', '풍선, '들이, ' ', '호위, '하듯, ' ', '에워, '싸고, ' ', '있었다, ' ', ' ', '지, '표면, '을, ' ', '뚫고, ' ', '솟아, '오른, ' ', '수백, '개의, ' ', '바늘, '은, ' ', '천공, ' ', '속, ' ', '존재, '들, '에, ' ', '대항, '이라도, ' ', '하듯, ' ', '벼, '린, ' ', '날, '을, ' ', '곧, '추세, '왔다, ' ', ' ', '해, '넘, '이와, ' ', '열, '기구, ' ', ' ', '그리고, ' ', '천, '불, '천, '탑, '(', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', ')', '이, ' ', '그려, '낸, ' ', '불, '국토, '(', '?', '?', '?', '?', '?', '?', '?', ')', ' ', ' ', '바, '간, '의, ' ', '풍경, '은, ' ', '사진, ' ', '속, ' ', '이미지, '에, ' ', '불과, '했지만, ' ', '보는, ' ', '이의, ' ', '넋, '을, ' ', '앗아, '갔다]

영문 기사 - 토큰 수: 168

['R, 'ou, 'nd, ' ', 'b, 'all, 'o, 'on, 's, ' ', 'of, ' ', 'v, 'ar, 'i, 'ou, 's, ' ', 'co, 'l, 'or, 's, ' ', 'we, 're, ' ', 'su, 'r, 'r, 'ou, 'nd, 'ing, ' ', 'a, ' ', 'b, 'all, ' ', 'of, ' ', 'fi, 're, ' ', ' ', 'H, 'un, 'd, 'red, 's, ' ', 'of, ' ', 'need, 'le, 's, ' ', 'from, ' ', 'the, ' ', 'gr, 'ou, 'nd, ' ', 'had, ' ', 'sh, 'ar, 'p, ' ', 'ed, 'ge, 's, ' ', 'as, ' ', 'if, ' ', 'the, 'y, ' ', 'we, 're, ' ', 'f, 'ight, 'ing, ' ', 'the, ' ', 'be, 'ing, 's, ' ', 'of, ' ', 'the, ' ', 'sky, ' ', ' ', 'T, 'he, ' ', 'photo, ' ', 'of, ' ', 'B, 'ag, 'an, ' ', 'over, 'w, 'he, 'l, 'me, 'd, ' ', 'me, ' ', 'with, ' ', 'the, ' ', 'sun, 'se, 't, ' ', ' ', 'hot, ' ', 'air, ' ', 'b, 'all, 'o, 'on, 's, ' ', ' ', 'th, 'ou, 's, 'and, 's, ' ', 'of, ' ', 'B, 'u, 'd, 'd, 'has, ' ', 'and, ' ', '1, ' ', '0, '0, '0, ' ', 'pa, 'go, 'd, 'as, '!]

3.3 Model Preparation

Multilingual Tokenization (3/4)

- 영어 BPE Vocab

한글 기사 - 토큰 수: 300

[', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '에', '?', '?', '?', '?', '은', '?', '?', '?', '?', '을', '?', '?', '?', '?', '이', '?', '?', '?', '?', '?', '?',
'?', '?', '?', '하', '는', '?', '?', '?', '?', '?', '?', '어', '리', '를', ' ', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '의', '?', '?',
'?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '이', '?', '?', '?', '?', '?', '하', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '고', '있',
'?', '?', '다', ' ', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '을', '?', '?', '?', '고', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?',
'?', '?', '?', '?', '?', '의', '?', '?', '?', '?', '은', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '에', '?', '?',
'?', '?', '?', '이', '?', '?', '?', '?', '?', ' ', '하', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '을', '?', '?', '?', '?', '?', '?', '?',
'?', '?', '?', '다', ' ', ' ', '해', '?', '?', '?', '이', '?', '?', '?', '?', '?', '기', '?', '?', '?', ' ', '?', '?', '?', '리', '고', '?', '?', '?', '?', '?',
'?', '?', '?', '?', '?', '?', '(', '?', '?', '?', '?', '?', '?', '?', '?', '?', ')', '이', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?', '?',
'?', '?', '?', '?', '?', '(', '?', '?', '?', '?', '?', ')', '?', '?', '?', '?', '의', '?', '?', '?', '?', '은', '?', '?', '?', '?', '?', '?',
'?', '?', '이', '?', '?', '?', '지', '에', '?', '?', '?', '?', '?', '?', '?', '?', '지', '?', '?', '?', '?', '?', '?', '는', '이', '의', '?', '?', '?', '을',
'?', '?', '?', '?', '?', '?', '다']

영문 기사 - 토큰 수: 60

['Round', ' balloons', ' of', ' various', ' colors', ' were', ' surrounding', ' a', ' ball', ' of', ' fire', ' ', ' ', 'H', ' undreds', ' of', ' needles', ' from',
' the', ' ground', ' had', ' sharp', ' edges', ' as', ' if', ' they', ' were', ' fighting', ' the', ' beings', ' of', ' the', ' sky', ' ', ' ', 'The', ' photo', ' of',
' B', ' agan', ' overwhelmed', ' me', ' with', ' the', ' sunset', ' ', ' ', ' hot', ' air', ' balloons', ' ', ' ', ' thousands', ' of', ' Budd', ' has', ' and', ' '1', ' ',
'000', ' pag', ' od', ' as', ' '']

3.3 Model Preparation

Multilingual Tokenization (4/4)

- 다국어 BPE Vocab

한글기사 - 토큰 수: 107

['삼', '라', '만', '상에', '붉은', '물을', '들', '이며', '침', '잠', '하는', '불', '덩', '어', '리', '를', ';;', '형', '형', '색', '색의', '원형', '풍선', '들이', '호', '위', '하', '듯', '에', '위', '싸고', '있었다', ';;', '지표', '면을', '뚫고', '솟', '아', '오른', '수백', '개의', '바늘', '은', '천', '공', '속', '존재', '들에', '대항', '이라도', '하', '듯', '벼', '린', '날을', '곧', '추', '세', '웠다', ';;', '해', '◆', '◆', '◆', '이와', '열', '기구', ';;', '그리고', '천', '불', '천', '탑', '(', '◆', '◆', '佛', '◆', '◆', '◆', '◆', ')', '이', '그러', '넌', '불', '국토', '(', '佛', '國', '土', ')', '바', '간의', '풍경', '은', '사진', '속', '이미', '지에', '불과', '했지만', '보는', '이의', '◆', '◆', '을', '앗', '아', '갔다']

영문기사 - 토큰 수: 67

['Round', 'ball', 'oons', 'of', 'various', 'colors', 'were', 'surrounding', 'a', 'ball', 'of', 'fire', ';;', 'H', 'undred', 's', 'of', 'need', 'les', 'from', 'the', 'ground', 'had', 'sharp', 'edges', 'as', 'if', 'they', 'were', 'fighting', 'the', 'beings', 'of', 'the', 'sky', ';;', 'The', 'photo', 'of', 'B', 'agan', 'overwhel', 'med', 'me', 'with', 'the', 'sun', 'set', ';;', 'hot', 'air', 'ball', 'oons', ';;', 'thousands', 'of', 'Budd', 'has', 'and', '1', ';;', '000', 'p', 'ag', 'od', 'as', '!']

3.4 Data Control 모델 실험 설계



Data
Control

Parallel
Corpus

Scaling Law

3.4 Data Control 모델 실험 설계

Data Tagging (1/3)

- CTRL*의 Idea에 착안하여, Style, Content, Task 생성에 도움이 되는 Control Code 적용 (Domain Tagging, Language Tagging)

Reviews Rating: 5.0
I have been using this product for a few years and it is the best thing on my hair. It makes your hair soft but not greasy.
Rating: 5.0
I've used this product before and love it. It's light weight yet keeps my curls defined without making them crunchy!

Reviews Rating: 1.0
I bought this for my daughter and it was a complete waste of money. The case is very flimsy and the screen protector doesn't fit properly.
Rating: 1.0
This product did not work at all. It would charge but when you unplug it, it does nothing. Very disappointed.

Control Code인 Review 점수를 다르게 했을 때, 생성 결과의 차이

3.4 Data Control 모델 실험 설계

Data Tagging (2/3)

- Domain Tagging

컬렉션명	주요 내용	처리 방식
wiki	위키백과	문서\n제목: \${제목}\n본문: \${본문}
nikl-news	모두의말뭉치-뉴스	뉴스\n제목: \${제목}\n본문: \${본문}
nikl-web	모두의말뭉치-웹	웹\n제목: \${제목}\n본문: \${본문}
nikl_dialogue	모두의말뭉치-대화	대화\n본문: \${본문}

3.4 Data Control 모델 실험 설계

Data Tagging (3/3)

- Language Tagging

Strategy	Source sentence	Target sentence
Original	Hello World!	¡Hola Mundo!
T-ENC	--es-- Hello World!	¡Hola Mundo!
T-DEC	Hello World!	--es-- ¡Hola Mundo!
S-ENC-T-ENC	--en-- --es-- Hello World!	¡Hola Mundo!
S-ENC-T-DEC	--en-- Hello World!	--es-- ¡Hola Mundo!

한글: /user/nap-mixer/gpt3_corpus_ko_v02/000_003

> {"text": "_ko_ 문서\n제목: 차수재(HDPE) 보호재로서 토목섬유의 천공(꺾뿔림)하중에 대한 저항효과\n본문: 본 연구는 천공(꺾뿔림) 하중으로부터 차수재(HDPE)를 보호하기 위한 목적으로 설치한 토목섬유의 보호 효과에 대하여 비교 분석하기 위하여 수행되었다. 연구에서는 다양한 모양의 천공(꺾뿔림) 기구를 이용하여 재현한 천공 형식을 이용하여 모형 실험을 수행하였다. 보호재로서 차수재 상부에 설치되는 토목섬유의 천공 저항성은 보호재의 규격, 종류 그리고 천공 형식에 따라 큰 차이를 보인다.

영어: /user/nap-mixer/gpt3_corpus_en_v01/000_002

> {"text": "_en_ arxiv\n Body: ---\naddress: |\n Dipartimento of Mathematics "G.Peano", University of Torino\\\n enrico.bibbona@unito.it\nauthor:\n- Enrico Bibbona\nbibliography:\n- 'bibl.bib'\nntitle: 'Stochastic Chemical Kinetics. Theory and (Mostly) Systems Biological Applications, P. Erdi, G. Lente. Springer (2014)'\n---\n\nThe idea of modeling chemical reaction is rather old, but the topic has gained renewed interest in recent years. The newest experimental techniques make it possible to collect data from networks of reactions that take place at the level of a single cell and to compare experimental evidence with model predictions.

3.4 Data Control 모델 실험 설계

Data Ratio & Curation (1/3)

- 2가지 Question
 - 데이터 Ratio에 대한 영향?
 - 데이터의 Domain의 영향?

3.4 Data Control 모델 실험 설계

Data Ratio & Curation (2/3)

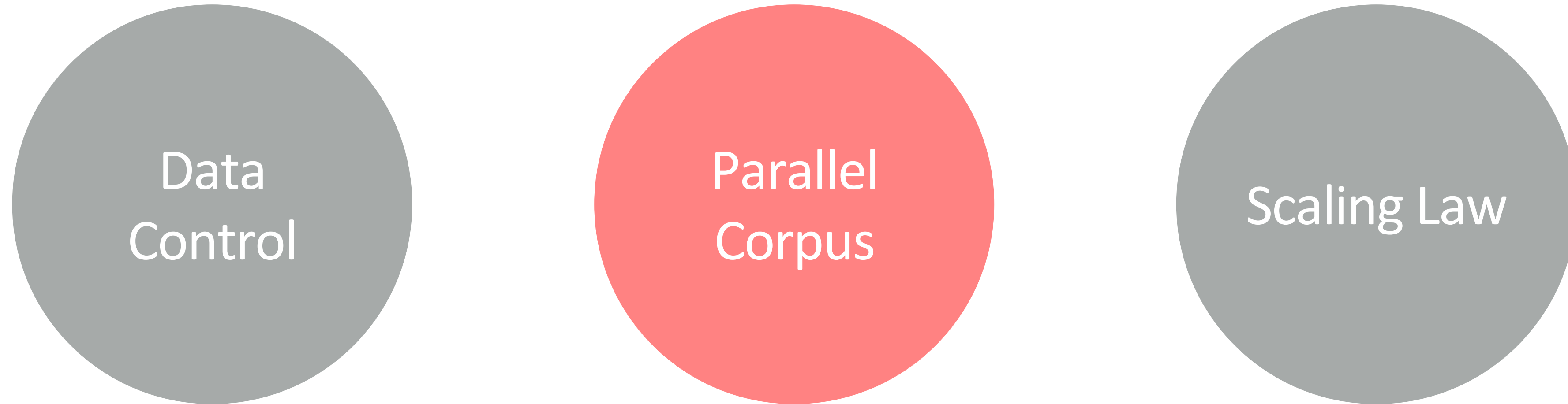
- 데이터의 종류가 다양하다면, 데이터 전체 비율이 미치는 영향
 - Ratio: Full, 1/100, 1/10
 - Random Sampling

3.4 Data Control 모델 실험 설계

Data Ratio & Curation (3/3)

- 데이터의 종류에 따른 영향도 실험 (Data Ablation Study)
 - HyperCLOVA 말뭉치
 - News Only
 - Modu Only
 - 지식데이터 Only

3.5 Parallel Corpus 모델 실험 설계



3.5 Parallel Corpus 모델 실험 설계

Parallel Corpus 소개 및 Hypothesis

- 언어간 전이 성능을 효과적으로 올리기 위한 시도들
 - mBERT (Google), XLM (Facebook), nmT5 (Google)

Aihub 번역 말뭉치를 사용하여 LM 계열인 현재 모델에 적용

스키너가 말한 보상은 대부분 눈으로 볼 수 있는 현물이다.=Skinner's reward is mostly eye-watering.₩n
 심지어 어떤 문제가 발생할 건지도 어느 정도 예측이 가능하다.=Even some problems can be predicted.₩n
 오직 하나님만이 그 이유를 제대로 알 수 있을 겁니다.=Only God will exactly know why.₩n
 중국의 논쟁을 보며 간과해선 안 될 게 기업들의 고충이다.=

3.6 Scaling Law



Data
Control

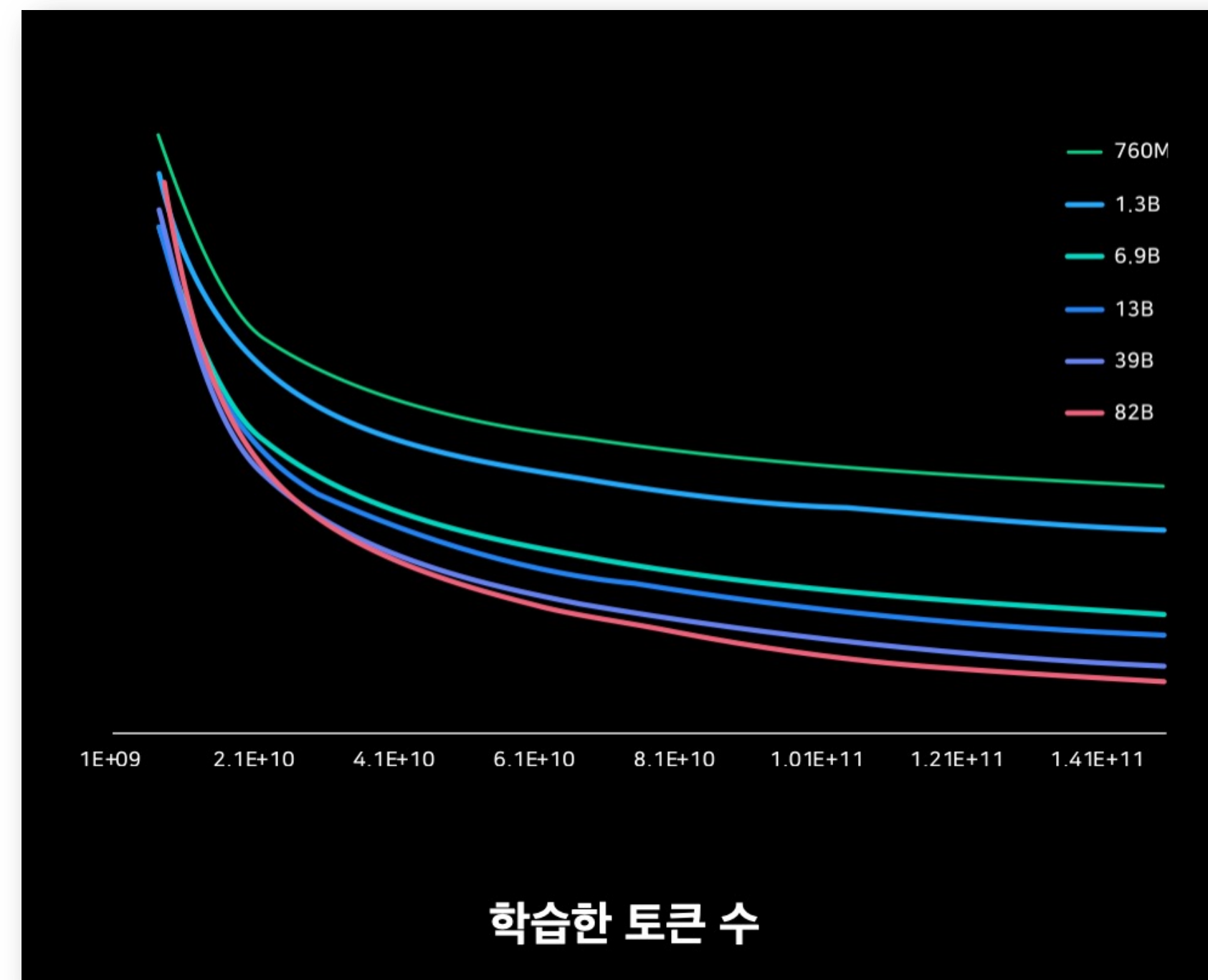
Parallel
Corpus

Scaling Law

3.6 Scaling Law

Scaling Law

- Multilingual LM 계열에서도 모델 사이즈에 따른 성능 증가가 가능할 것인가?



모델크기에 따른 성능 비교 - HyperCLOVA

4. Multilingual HyperCLOVA Model Evaluation

4.1 Model Evaluation 소개

How to evaluate? (1/2)

- Multilingual HyperCLOVA도 HyperCLOVA가 평가된 방식인 Few-shot learning 기반의 evaluation을 수행

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```

1 Translate English to French: ← task description
2 cheese => ..... ← prompt
  
```

One-shot

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

```

1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
  
```

Few-shot

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

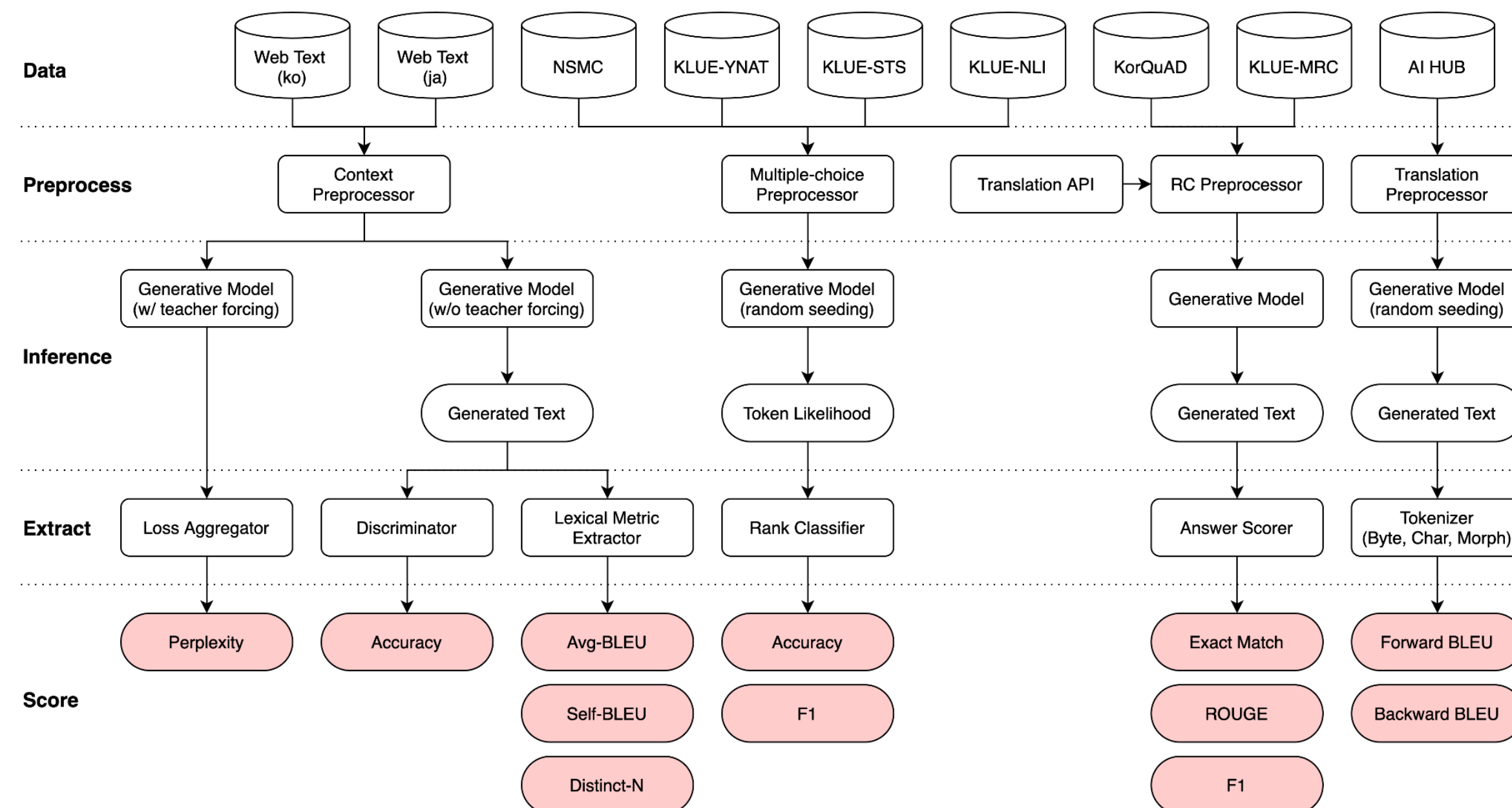
```

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

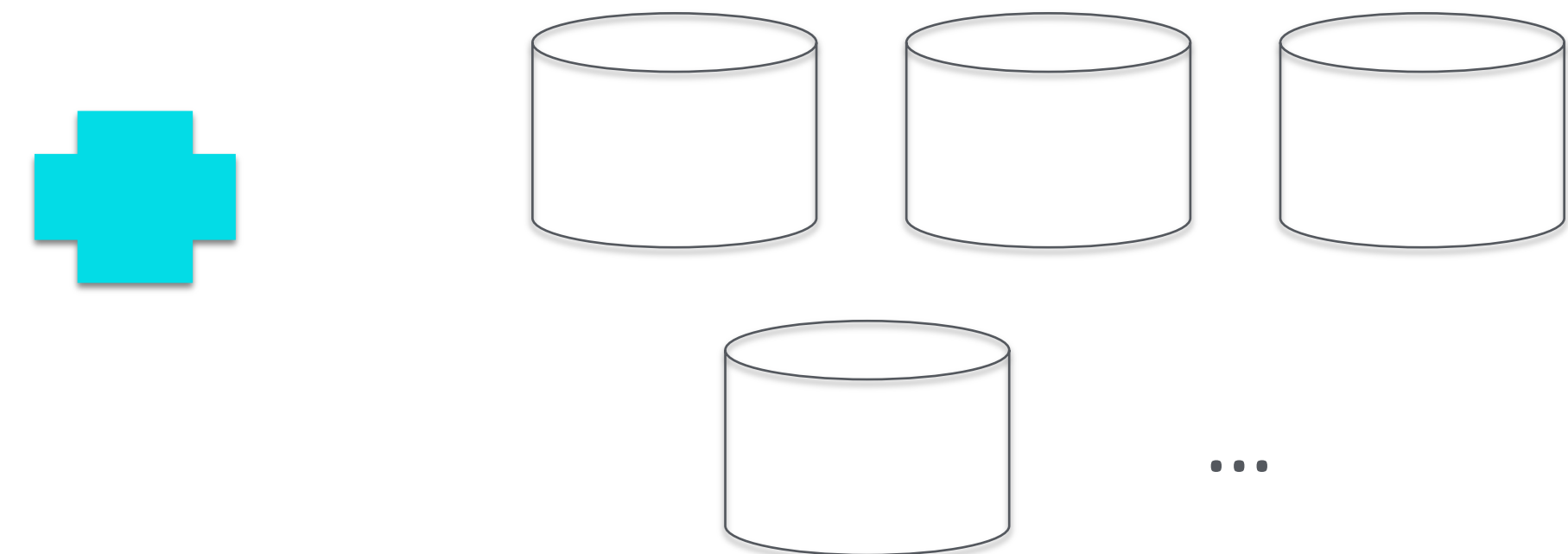
4.1 Model Evaluation 소개

How to evaluate? (2/2)

- HyperCLOVA에서 활용한 Few-shot learning 기반의 evaluation pipeline에 평가 데이터를 추가하여 Multilingual HyperCLOVA를 평가하는 데 활용



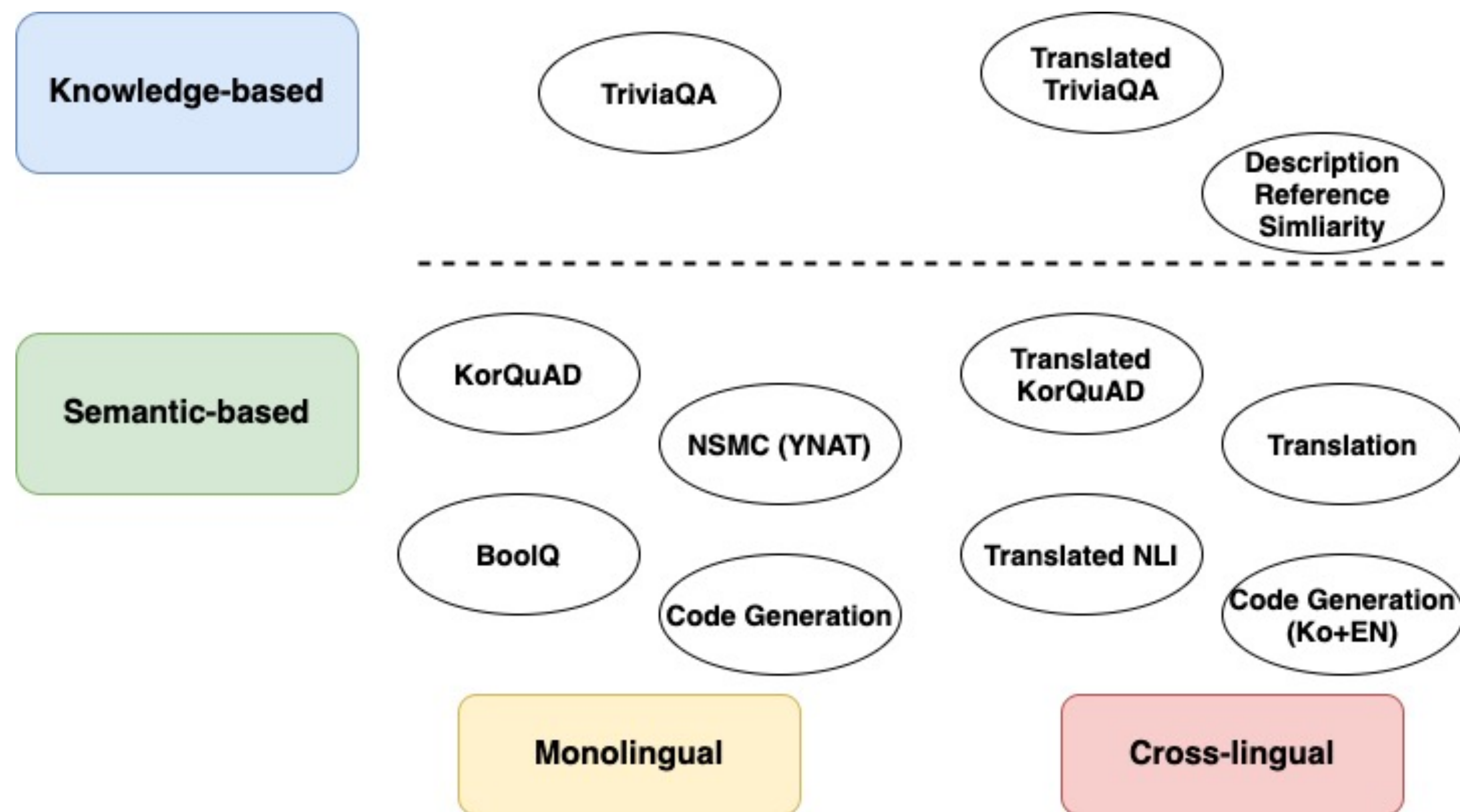
Multilingual HyperCLOVA를 평가하기 위한 특화된 데이터셋들



4.1 Model Evaluation 소개

Monolingual & Cross-lingual 평가 제작

- 다국어 모델의 성능 측정을 위한 평가 Dataset 제작 (진행 중)



4.2 Cross-lingual 평가 소개

Translated Trivia QA

- Factual Knowledge Transfer 측정을 위한 TriviaQA 활용 데이터셋

QuestionID: “wh_396”,

Question: “창세기에 기록된 성경의 첫 네 단어는 무엇입니까?”

Answer: “In THE BEGINNING GOD”

QuestionID: “dpql_3600”,

Question: “어벤져스에 린다 토르슨이 연기한 캐릭터는?”

Answer: ____ **“Tara King”** ____ **[생성 Part]**

4.2 Cross-Lingual 평가 소개

Translated NLI

- XNLI의 데이터와 KorNLI 데이터의 Pair를 기반으로 premise와 hypothesis 구성
 - KOR-ENG & ENG-KOR

Premise 한, hypothesis 영 (or premise: 영, hypothesis: 한)

말을 탄 사람이 고장난 비행기 위로 뛰어오른다. A person is training his horse for a competition.:1
나이 든 남자가 커피숍의 작은 테이블에 오렌지 주스를 들고 앉아 있고 밝은 색 셔츠를 입은
직원들은 뒤에서 미소를 짓고 있다. A boy flips a burger.: 2 [생성 Part]

4.2 Cross-Lingual 평가 소개

Translated KorQuAD & SQuAD

- 주어진 본문에 대하여, 다국어로 변형하여 질문하는 Task
 - 한글 지문: 영어 질문 - 영어 답변
 - 영어 지문: 한글 질문 - 한글 답변

[지문] 평일·휴일 모두 낮 시간대는 구로사키 역 앞 - 지쿠호 노가타 사이의 전선 열차와 구로사키 역 앞 - 지쿠호 나카마 사이의 구간 열차가 12분 간격으로 교대로, 평일 아침의 러쉬 시간대는 전선 열차와 구로사키 역 앞 - 지쿠호 나카마·구스바시 사이의 구간 열차가 5 - 10분 간격으로 거의 교대로, 평일 저녁의 러쉬 시간대는 전선 열차와 구로사키 에키마에 - 구스바시간의 구간 열차가 7 - 10분 간격으로 교대로 운행되고 있다.

Q: What is the interval between the Chikuho Electric Railway lines during the daytime?

A: 12 minutes [생성 Part]

4.2 Cross-Lingual 평가 소개

Description Reference Similarity (DRS)

- 생성된 문장과 실제 문장을 비교하여 평가

Q: 남녀노소

A: Men or women who are either young or old.

생성결과: ____ [생성 Part]

Q: 동고동락

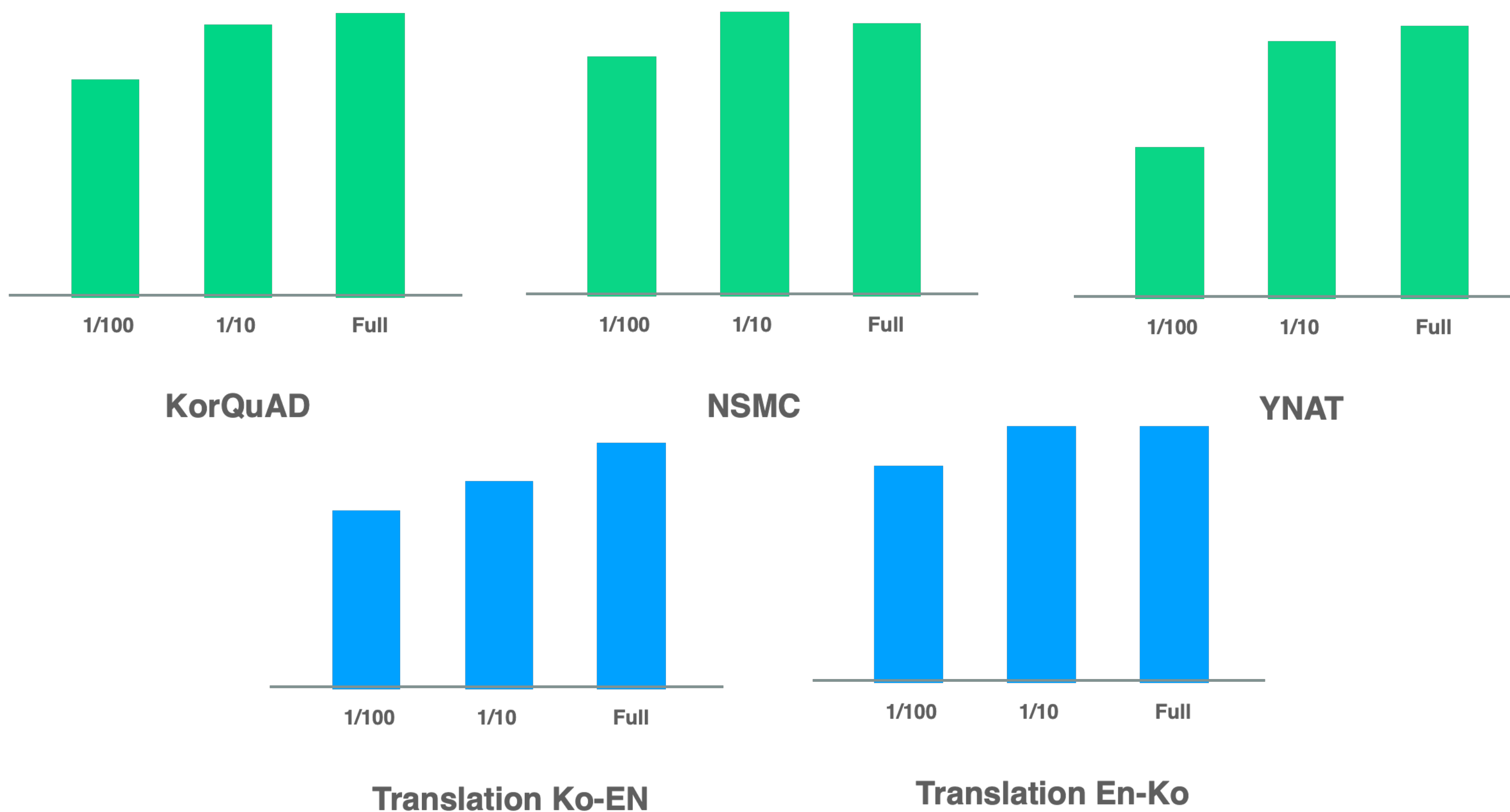
A: The state of sharing one's pains and joys.

생성결과: ____ [생성 Part]

4.3 Experiments Results

Data Ratio & Curation (1/2)

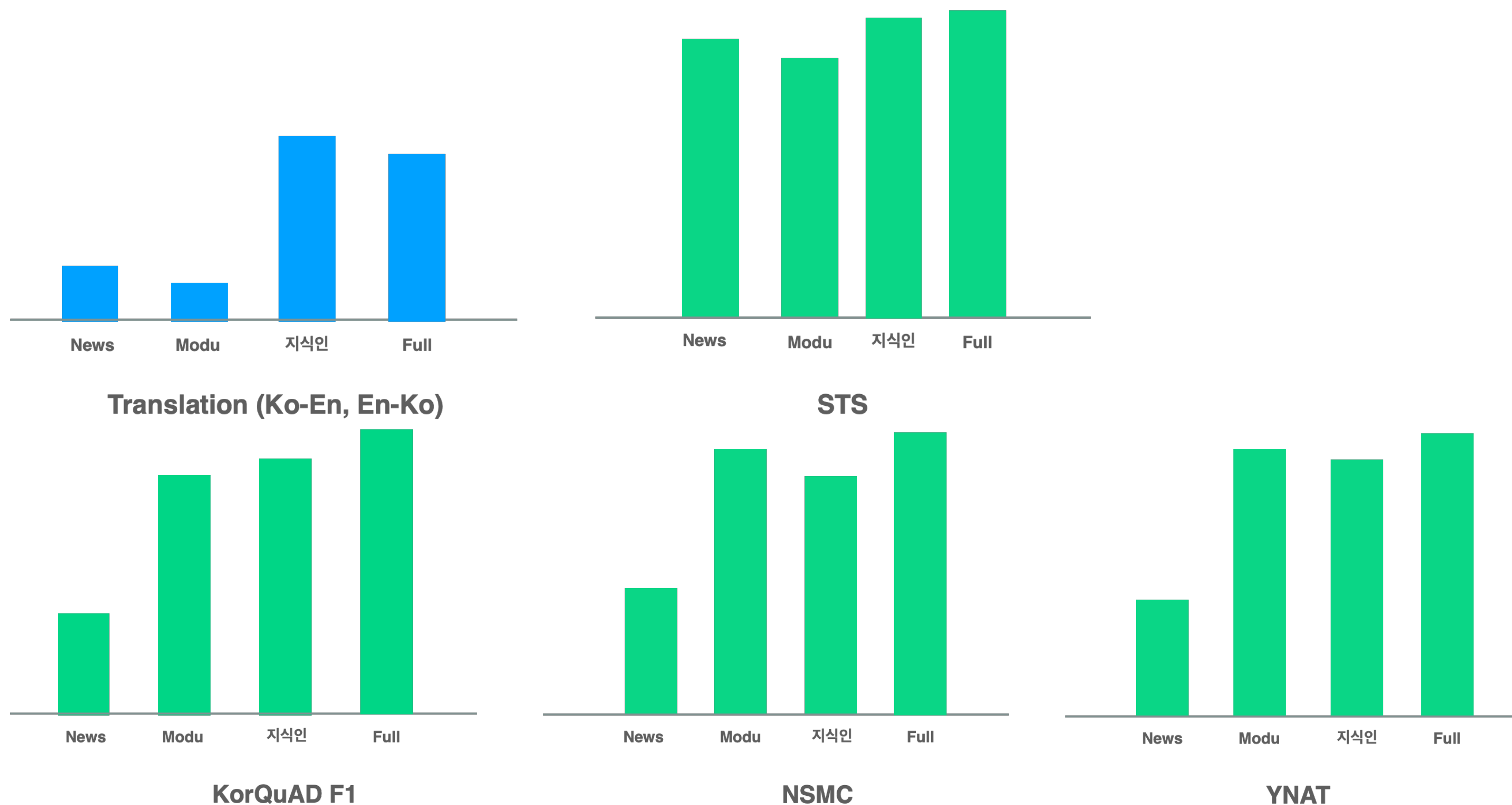
- Ratio 별 성능 측정



4.3 Experiments Results

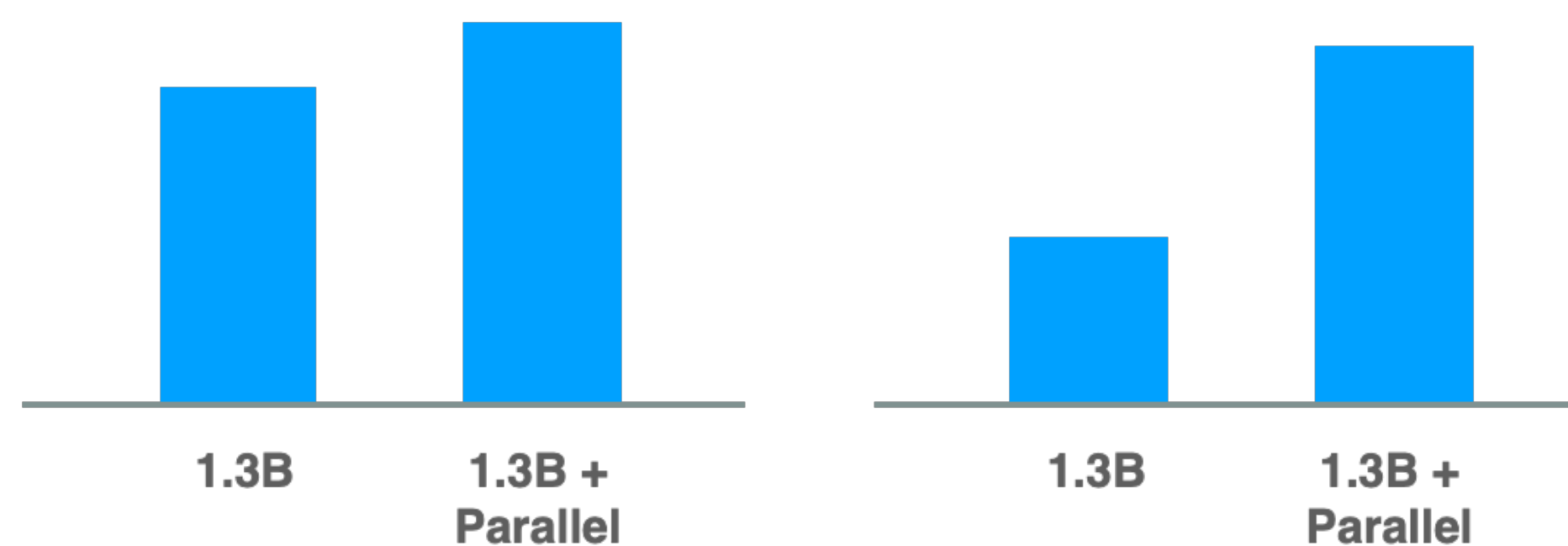
Data Ratio & Curation (2/2)

- Curation 별 성능 측정



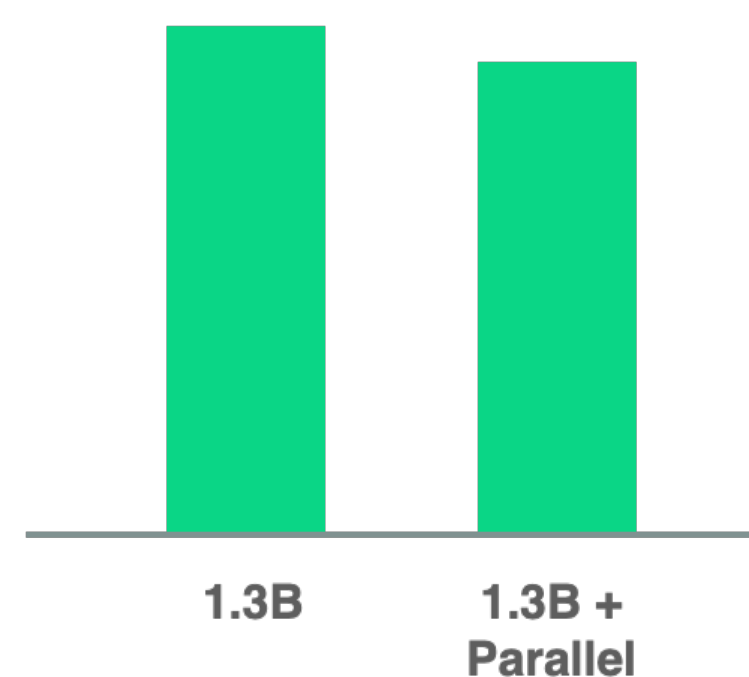
4.3 Experiments Results

Parallel-Corpus

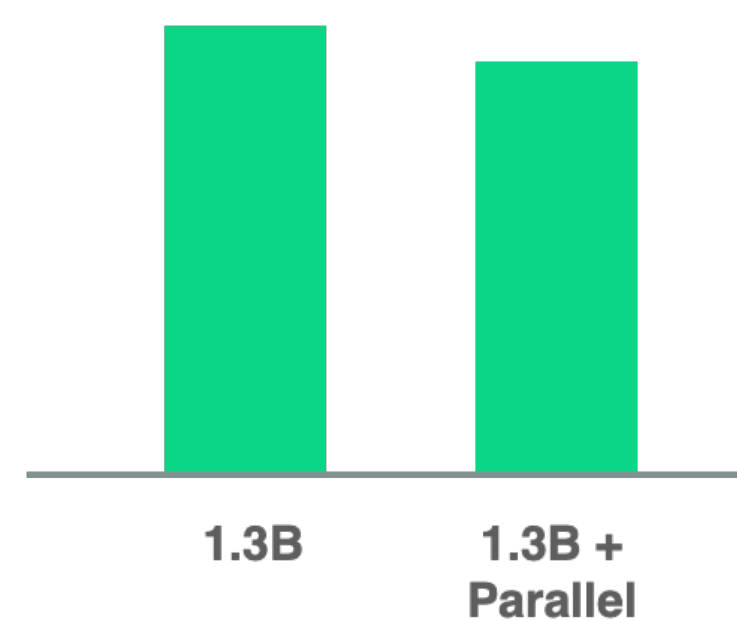


Translated KorQuAD

DRS



KorQuAD F1

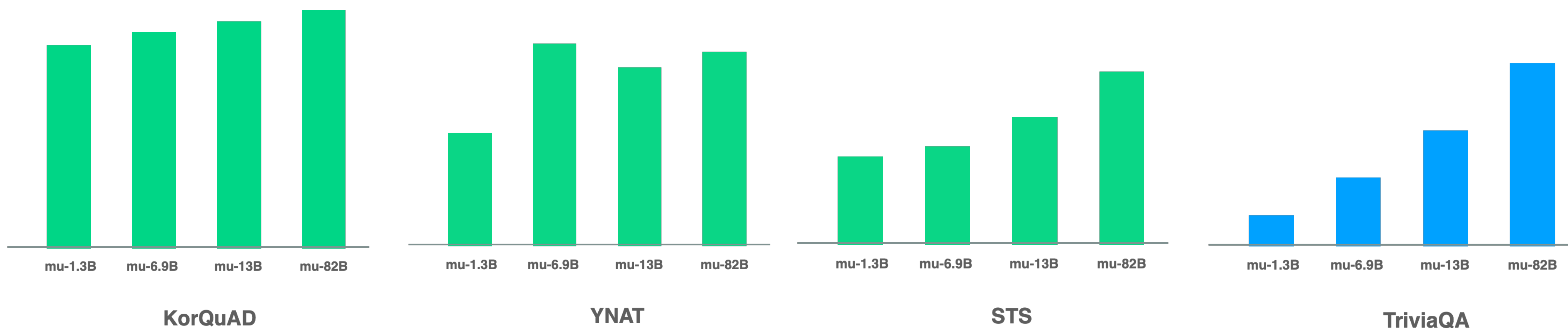


YNAT

4.3 Experiments Results

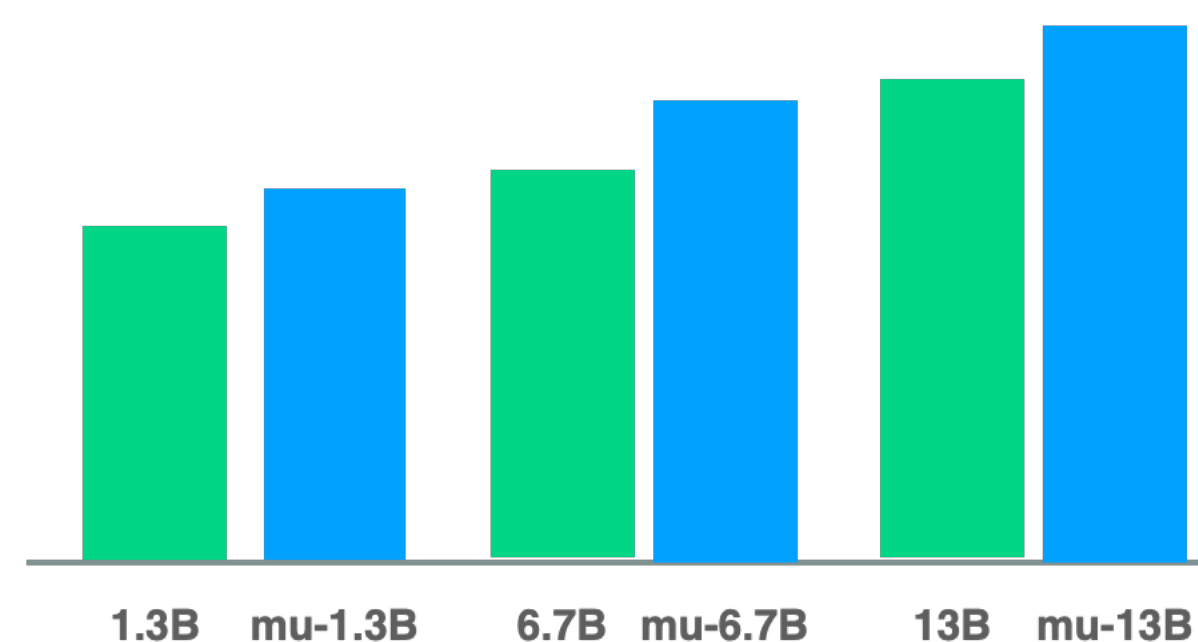
Scaling Law

- 사이즈가 커질수록 성능 증가 경향성

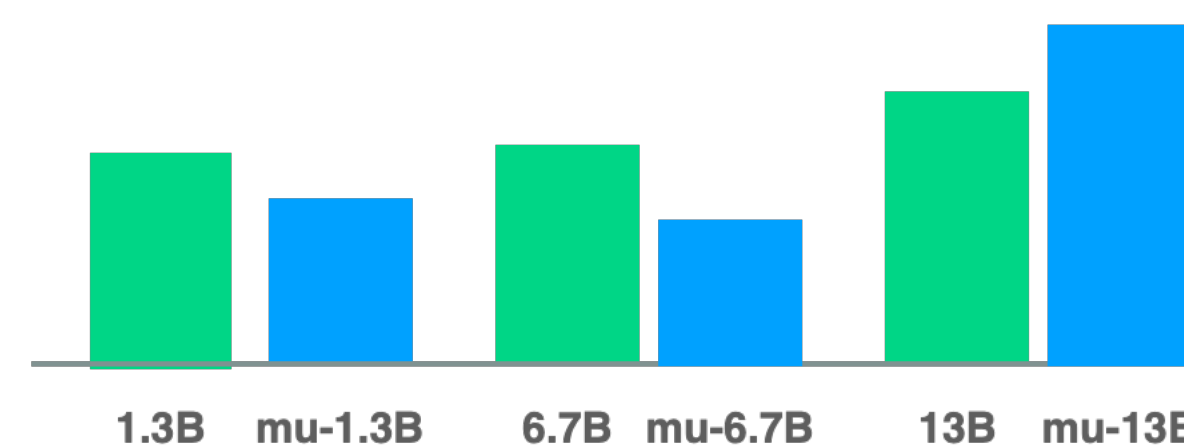


4.3 Experiment Results

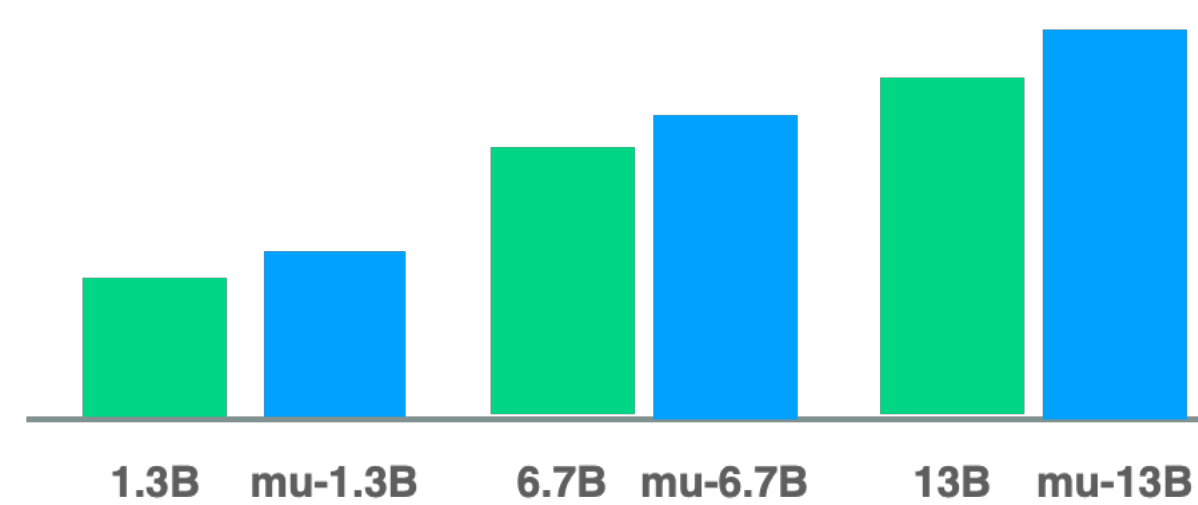
Cross-lingual



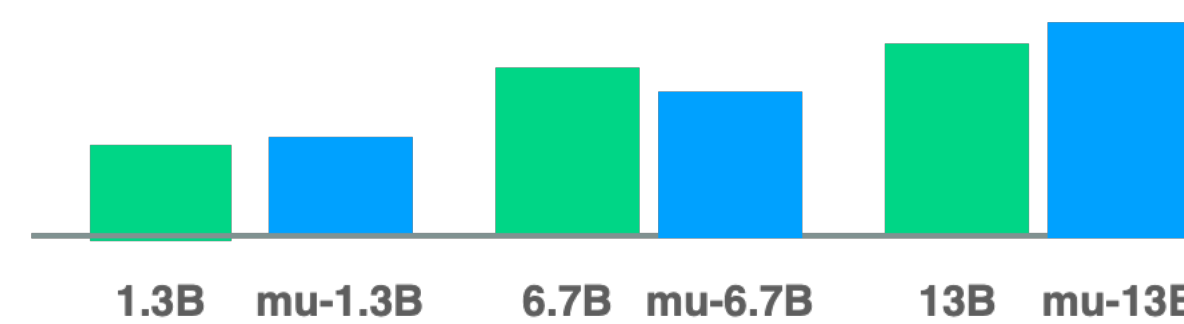
Translated KorQuAD



Translated NLI



Translated TriviaQA



Translation (Ko-En, En-Ko)

4.4 Applications

Code Switching

Q: What does idiom '개과천선' means?

A: A change of character. A change of character that makes a person more or less like a new person.

GT: Mending one's ways. The state of regretting one's faulty past and starting to behave better.

--

Q: What does idiom '격세지감' means?

A: A change in the level of a subject or a situation that is not as dramatic as it was in the past.

GT: Being amazed at how much things have changed in a short time. An impression of being in a completely different world due to much progress and change in a relatively short period of time.

--

Q: What does idiom '결자해지' means?

A: The right thing to do. The right thing to do is to do it.

GT: One's untying a knot one has tied. Whoever creates the problem solves it.

--

4.4 Applications

동의어 생성

아래의 한국어 단어의 영어 동의어를 찾으세요.

주장하다: insist, maintain, assert

먹다: eat, have, take

떠나다: leave, start off, depart

쉬다:

생성 Prompt

아래의 한국어 단어의 영어 동의어를 찾으세요.

주장하다: insist, maintain, assert

먹다: eat, have, take

떠나다: leave, start off, depart

쉬다: rest, stop, pause, lay down

듣다: listen, hear, overhear

보다: see, look at, watch, observe

쓰다: write, compose, draw up, make a note of

알다: know, understand, realize, recognize

말하다: talk, speak, say, tell

가다: go, come, walk, get on the bus, drive to work

오다: come, arrive, show up, appear, happen, occur

읽다: read

생성 결과

4.4 Applications

Code Generation

생성 Prompt

```
def is_max(a: int, b: int, c: int) :
    """
    입력받은 세개의 숫자중 가장 큰 숫자를 리턴하는 파이썬 프로그램을 만들어줘.
    >>> is_max(3,2,1)
    3
    >>> is_max(1,1,1)
    1
    """
```

생성 결과

```
def is_max(a: int, b: int, c: int) :
    """
    입력받은 세개의 숫자중 가장 큰 숫자를 리턴하는 파이썬 프로그램을 만들어줘.
    >>> is_max(3,2,1)
    3
    >>> is_max(1,1,1)
    1
    """
    return max(a,b,c)
```

5. Future roadmap

5.1 Pre-training 방법론 개선

직접적인 지식 전이 방법 검토

- Code Switching 등의 Data Augmentation 접근법

1998년 1월 네이버 정식 서비스가 시작되었다. 네이버는 '항해하다'라는 뜻의 *Navigate*와 '-하는 사람'의 뜻을 가진 접미사 *-er*이 만나 탄생한 이름이다.[2] 이는 '정보의 바다라고 일컫는 인터넷을 항해하는 사람'을 의미한다. 1999년 6월 네이버컴(주)가 설립, '네이버' 정식 서비스와 어린이 전용 포털인 '쥬니어네이버' 서비스가 시작되었다. 2000년 5월 '네이버 뉴스' 서비스가 시작되었다.

Google was founded on September 4, 1998, by Larry Page and Sergey Brin while they were Ph.D. students at Stanford University in California. Together they own about 14% of its publicly-listed shares and control 56% of the stockholder voting power through super-voting stock. The company went public via an initial public offering (IPO) in 2004. In 2015, Google was reorganized as a wholly-owned subsidiary of Alphabet Inc..



일부 sentence code를 switch

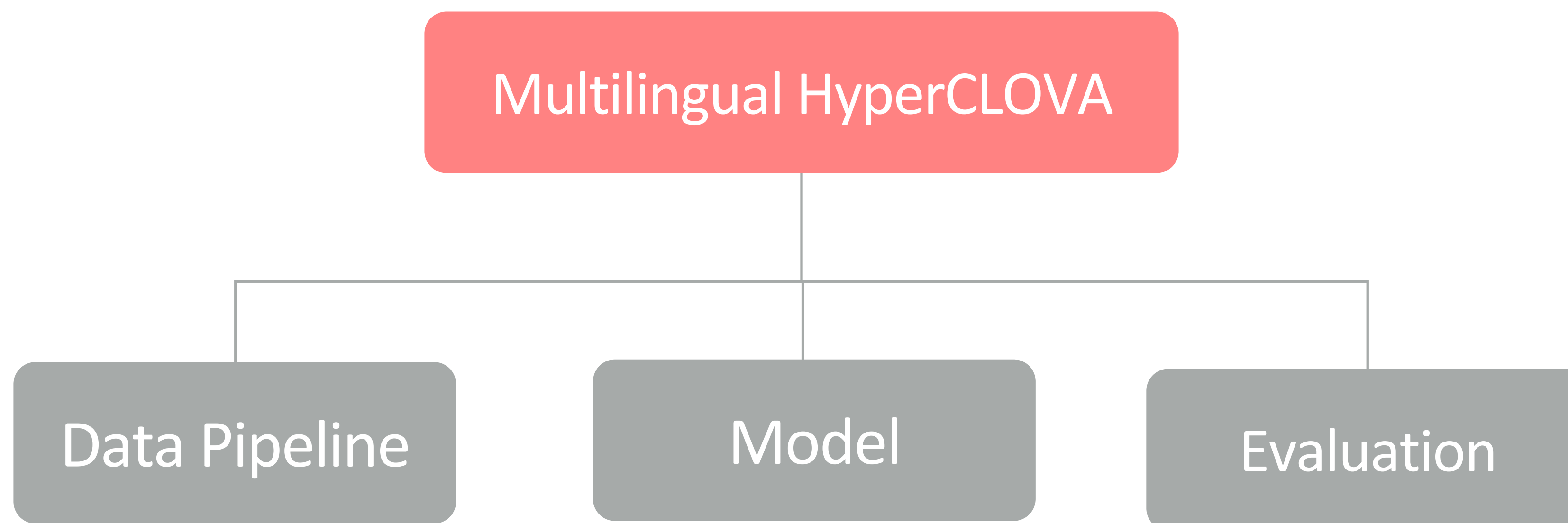
1998년 1월 네이버 정식 서비스가 시작되었다. 네이버는 '항해하다'라는 뜻의 *Navigate*와 '-하는 사람'의 뜻을 가진 접미사 *-er*이 만나 탄생한 이름이다.[2] It means 'a person who navigates the Internet, which is called the sea of information'. 1999년 6월 네이버컴(주)가 설립, '네이버' 정식 서비스와 어린이 전용 포털인 '쥬니어네이버' 서비스가 시작되었다. 2000년 5월 '네이버 뉴스' 서비스가 시작되었다.

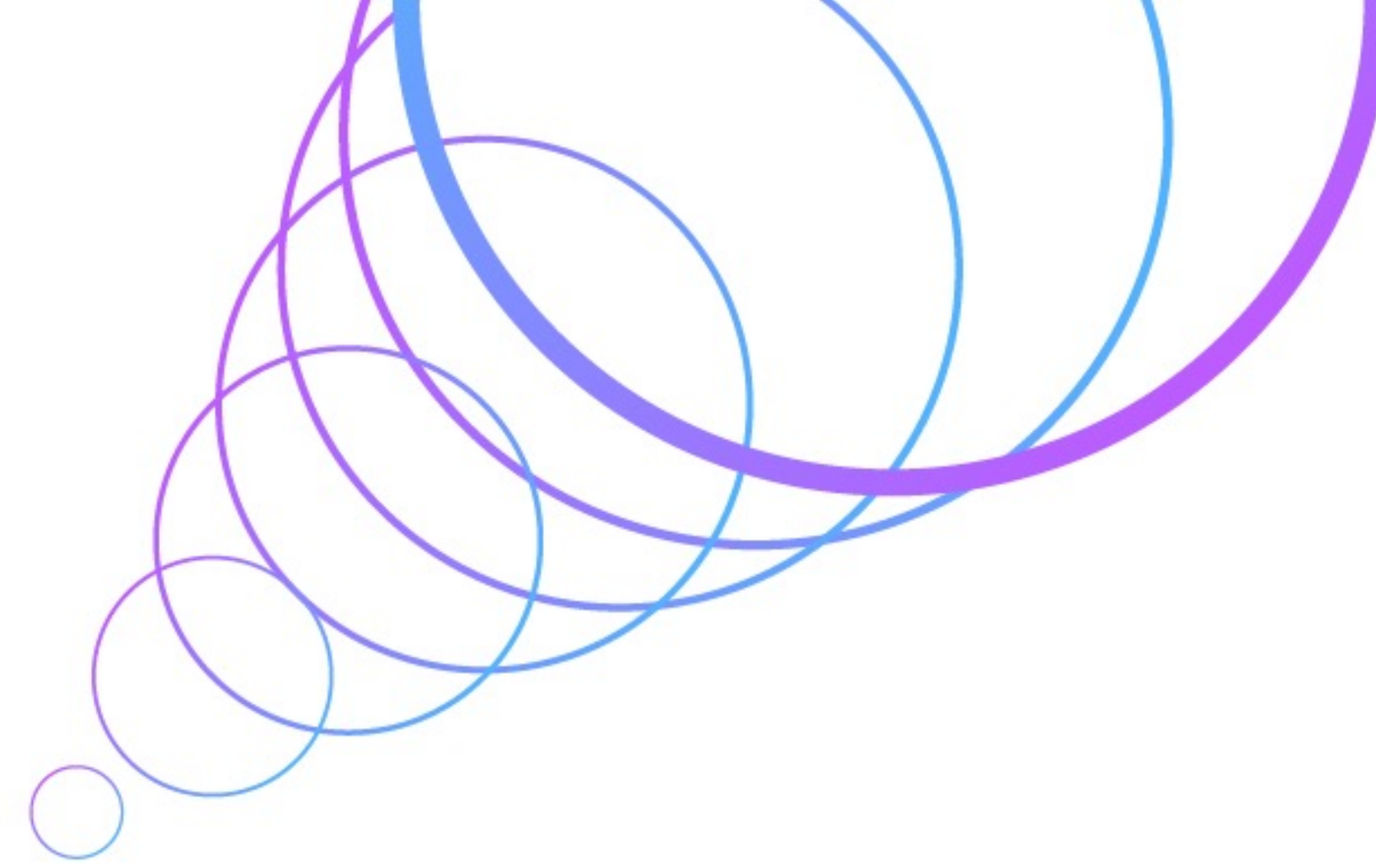
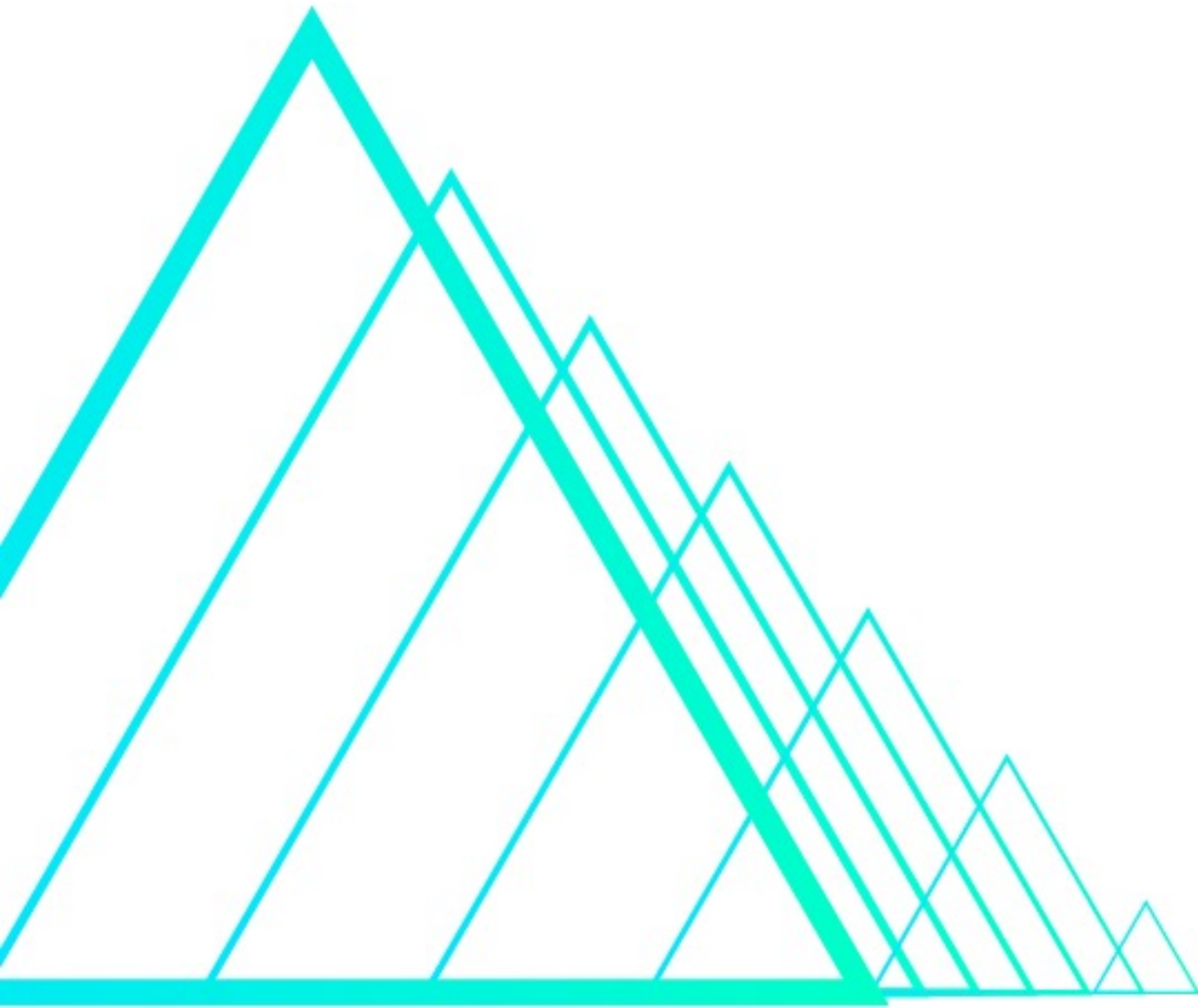
Google was founded on September 4, 1998, by Larry Page and Sergey Brin while they were Ph.D. students at Stanford University in California. Together they own about 14% of its publicly-listed shares and control 56% of the stockholder voting power through super-voting stock. 회사는 2004년에 기업공개(IPO)를 통해 공개되었습니다. 2015년에 Google은 Alphabet Inc.의 전액 출자 자회사로 재편되었습니다.

6. Summary

6.1 Summary

Beyond HyperCLOVA





Thank You

