Journal of The Korea Convergence Society
Vol. 13. No. 3, pp. 77-89, 2022
e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2022.13.03.077

Cross-Lingual Post-Training (XPT)을 위한 한국어 및 다국어 언어모델 연구

손수현[1], 박찬준[1], 이정섭[1], 심미단[2], 이찬희[3], 박기남[4], 임희석[5*]
[1]고려대학교 컴퓨터학과 석·박사통합과정 [2]경희대학교 소프트웨어융합학과 학생, [3]네이버 연구원,
[4]고려대학교 Human-inspired AI 연구소 연구교수, [5]고려대학교 컴퓨터학과 교수

Korean and Multilingual Language Models Study for Cross-Lingual Post-Training (XPT)

Suhyune Son[1], Chanjun Park[1], Jungseob Lee[1], Midan Shim[2], Chanhee Lee[3], Kinam Park[4], Heuiseok Lim[5*]
[1]Master & Ph. D. Combined Student, Department of Computer Science and Engineering, Korea University
[2]Student, Department of Software Convergence, Kyung Hee University
[3]Research Engineer, Naver Corporation
[4]Research Professor, Human-inspired Computing Research Center, Korea University
[5]Professor, Department of Computer Science and Engineering, Korea University

요 약 대용량의 코퍼스로 학습한 사전학습 언어모델이 다양한 자연어처리 태스크에서 성능 향상에 도움을 주는 것은 많은 연구를 통해 증명되었다. 하지만 자원이 부족한 언어 환경에서 사전학습 언어모델 학습을 위한 대용량의 코퍼스를 구축하는데는 한계가 있다. 이러한 한계를 극복할 수 있는 Cross-lingual Post-Training (XPT) 방법론을 사용하여 비교적 자원이 부족한 한국어에서 해당 방법론의 효율성을 분석한다. XPT 방법론은 자원이 풍부한 영어의 사전학습 언어모델의 파라미터를 필요에 따라 선택적으로 재활용하여 사용하며 두 언어 사이의 관계를 학습하기 위해 적응계층을 사용한다. 이를 통해 관계추출 태스크에서 적은 양의 목표 언어 데이터셋만으로도 원시언어의 사전학습 모델보다 우수한 성능을 보이는 것을 확인한다. 더불어, 국내외 학계와 기업에서 공개한 한국어 사전학습 언어모델 및 한국어 multilingual 사전학습 모델에 대한 조사를 통해 각 모델의 특징을 분석한다

주제어 : 사전학습 언어모델, 전이학습, 한국어 언어모델, 다국어 언어모델, 언어융합

Abstract It has been proven through many previous researches that the pretrained language model with a large corpus helps improve performance in various natural language processing tasks. However, there is a limit to building a large-capacity corpus for training in a language environment where resources are scarce. Using the Cross-lingual Post-Training (XPT) method, we analyze the method's efficiency in Korean, which is a low resource language. XPT selectively reuses the English pretrained language model parameters, which is a high resource and uses an adaptation layer to learn the relationship between the two languages. This confirmed that only a small amount of the target language dataset in the relationship extraction shows better performance than the target pretrained language model. In addition, we analyze the characteristics of each model on the Korean language model and the Korean multilingual model disclosed by domestic and foreign researchers and companies.

Key Words : Pretrained Language Model, Transfer Learning, Korean Language Model, Cross-Lingual Language Model, Language Convergence

*This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).

*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)

Received January 12, 2022 Revised February 3, 2022
Accepted March 20, 2022 Published March 28, 2022

1. 서론

최근 large-scale 데이터를 기반으로 딥러닝 기반 자연언어처리 연구가 학계, 기업 모두에서 활발히 이루어지고 있다. Foundation model[1], GPT-3[2], Prompt tuning[3] 등과 같은 대용량 모델의 등장과 scaling law와 같은 이론적 증명을 통해 대용량 데이터의 효과성에 대한 많은 연구가 이루어지고 있다.

이러한 기술의 뼈대는 방대한 양의 말뭉치를 통해 사전 학습된 Transformer모델이다. Transformer 모델[4]은 sequence to sequence 기반 full attention 모델이며 BERT[5]를 시작으로 RoBERTa[6], XLM[7], BART[8], ELECTRA[9] 등 수많은 변형 모델들의 등장으로 많은 자연어처리 하위 시스템들의 성능을 대폭 향상시켰다. 위와 같은 연구들은 언어모델이라는 분야로 정의할 수 있으며 해당 분야는 최근 가장 활발한 연구가 이루어지고 있다. 현재 진행되고 있는 대부분의 언어모델 연구는 대용량 원시 말뭉치를 요구하는 비지도 학습에 속하며 데이터 확보가 가장 중요한 요소 중 하나이다. 그러나 이러한 언어모델 연구에는 크게 2가지 한계점이 존재한다.

첫째 언어적 한계점이다. 현재 이루어지고 있는 대부분의 언어모델 연구의 대다수는 영어를 기반으로 이루어지고 있다. 위키피디아에는 303종의 언어가 존재하며, 문서의 양이 가장 많은 영어의 경우 6백만건 이상의 문서가 존재한다. 반면, 절반 이상에 해당하는 154종의 언어에는 1만건의 문서도 존재하지 않는다. 한국어의 경우 약 52만건의 문서가 속해 있으며, 24위에 해당한다. 하지만 이는 여전히 영어 문서 대비 8.5%의 해당하는 수치이다.

둘째 모델적 한계점이다. 학습 데이터의 양은 언어모델의 성능에 직결되는 요소이다. 즉 언어 자원의 불균형이 발생할 경우 성능 향상에 있어 장애물이 될 수 있다. 물론 mBERT와 같은 Multilingual 모델을 통해 여러 언어를 동시에 하나의 모델로 학습하는 방법으로 한국어, 스와힐리어와 같은 저자원 언어에서의 성능을 올릴 수 있지만, 이는 학습에 필요한 파라미터와 모델의 크기를 기하급수적으로 증가시켜 실효성과 효율성이 좋지 못하다.

위와 같은 문제를 해결하기 위하여 본 논문은 Cross-lingual Post-Training (XPT)[10] 방법론을 이용하여 최근 뉴럴 심볼릭 연구의 필요성으로 인해 중요성이 강조되는 관계추출 분야에 적용해보려한다. XPT란, 논문에서 제안된 방법으로 오직 영어로만 학습된 Transformer 기반 언어 모델에서 얻은 파라미터 중 재사용 가능한 부분을 선별하여 목표 언어의 모델을 초기화한 후 학습을 진행한다. 즉 별도의 사전학습을 진행하지 않아도 된다는 장점이 존재한다. 또한, 두 언어 간 차이를 학습하는 역할을 하는 적응 층들을 추가하여 이종언어 간 전이 학습을 가능케 한다.

위와 같은 실험과 더불어 본 논문은 학계와 기업에서 공개한 대표적인 한국어 사전학습 언어모델(Pretrained Language Model, PLM)과 전 세계적으로 공개된 한국어 다국어 언어모델(multilingual Pretrained Language Model, multilingual PLM) 모델에 대한 조사를 진행하였으며, 이에 대한 특징을 면밀히 분석하였다.

2. 한국어 PLM 연구

한국어의 경우 Transformer를 기반으로 사전학습된 모델을 오픈하거나 연구한 사례가 존재한다. 그러나 대부분의 연구들이 원본 논문에서 요구한 양을 충족시키지 못하고 있다. KoBERT[1], HanBERT[2], KoELECTRA [11], KcBERT[12], KcELECTRA[13], DistilKoBERT [14], KoBigBird[15], SKT-AI KoGPT2[3], HyperCLOVA [16], SKT KoGPT-Trinity[4], Kakaobrain KoGPT[17], KOBART[5], LG AI - EXAONE[6], KE-T5 [18], ET5[7] 등 다양한 연구들이 존재하며 시간순으로 각 모델들을 설명하며 이는 Fig. 1에 정리되어 있다. 크게 인코더, 디코더 그리고 인코더-디코더 구조로 나누어 구분하였으며 공개되지 않은 것들에 대해서는 ~로 표기하였다. 또, 직접 구축하여 정보를 알 수 없는 토크나이저의 경우 *로 표기하였다.

2.1 기업에서 공개한 한국어 PLM 연구

[2019년] 한국전자통신연구원(Etri)에서 최초의 한국어

1) https://github.com/SKTBrain/KoBERT

2) https://github.com/monologg/HanBert-Transformers

3) https://github.com/SKT-AI/KoGPT2

4) https://huggingface.co/skt/ko-gpt-trinity-1.2B-v0.5

5) https://github.com/SKT-AI/KoBART

6) https://www.lgresearch.ai/blog/view/?seq=170

7) https://aiopen.etri.re.kr/service_dataset.php

PLM인 KorBERT[8]를 2019년 6월에 공개하였다. 이는 Transformer의 Encoder 구조로 이루어진 기존의 BERT 모델을 한국어에 적용하기 위해 한국어 뉴스와 백과사전으로부터 추출한 23GB의 데이터로 학습된 모델이다. 파라미터 크기는 110M이고, Morpheme 및 WordPiece tokenizer[19]를 사용하였으며, vocab의 크기는 30,349(Morphemes), 30,797(WordPiece)이다.

이후 SKT에서 KoBERT를 2019년 10월에 공개하였다. KoBERT는 한국어 위키피디아[9]에서의 5천만개의 문장으로 학습된 모델로, SentencePiece tokenizer[20]를 사용하였다. vocab 크기는 8002이며, 모델의 파라미터 크기는 92M이다.

[2020년] 2020년 1월에 TwoBlock AI에서 공개한 HanBERT는 일반문서 및 특허문서 70GB로 학습되었으며 자체적으로 만든 Moran tokenizer가 사용되었다. vocab 크기는 54,000이며, 모델 파라미터 크기는 128M이다. 이러한 모델들은 BERT와 동일한 Masked Language Model와 Next Sentence Prediction으로 사전학습되었다.

SKT는 2020년 2월에 한국어 버전의 GPT2[21] 모델인 KoGPT2를, 2020년 12월에 한국어 버전의 BART 모델인 KoBART를 공개하였다. KoGPT2은 GPT2와 마찬가지로 transformer decoder 구조를 가지고 있으며, next token prediction을 통해 사전학습되었다. 한국어 위키피디아, 뉴스, 나무위키[10], 네이버 영화 리뷰, 한국어 commonCrawl[11] 등의 다양한 데이터로부터 추출한 152M개의 문장으로 학습되었으며, 기존의 byte pair encoding(BPE)[22] 중에서 character-level의 tokenizer가 사용되었다. Vocab 크기는 51,200이고 현재 125M 파라미터의 base 모델만 공개되었다. KoBART는 BART와 마찬가지로 인코더-디코더 구조를 가지고 있으며 denoising auto encoder 방법으로 사전학습되었다. 한국어 위키피디아, 뉴스 뿐만 아니라 책, 모두의 말뭉치[12], 청와대 국민 청원[13]과 같이 더 다양한 0.27B의 데이터로 학습되었다. KoGPT2와 마찬가지로 character BPE tokenizer를 사용하였으며, vocab크기는 30,000이고 모델 파라미터 크기는 124M이다.

[2021년] 2021년 1월에 공개된 KoreALBERT[23]는 삼성 SDS에서 공개한 모델이다. ALBERT[24]와 같이 Masked Language Model와 Sentence-Order Prediction 방법으로 사전학습 되었다. 한국어 위키피디아 및 나무위키, 뉴스, 책으로부터 얻은 43GB의 데이터로 학습되었으며 SentencePiece tokenizer를 사용하여 분절화한다. vocab 크기는 32,000이고 모델은 12M의 base 모델과 18M의 large 모델이 공개되어 있다.

한국전자기술연구원(KETI)에서는 2021년 4월 KE-T5 즉 한국어와 영어 버전의 Text-to-Text Transfer Transformer(T5)[25] 모델을 공개하였다. 학습에는 한국어와 영어 데이터가 7:3의 비율로 섞인 30GB의 데이터가 사용되었으며, 이 데이터 중 한국어는 KETI에서 확보하고 있는 비정형 말뭉치를 전처리한 데이터와 모두의 말뭉치 중 일부로 이루어져 있다. SentencePiece tokenizer를 사용하였으며, vocab 크기는 64,000이다. 현재 공개된 모델의 파라미터는 247M이고 T5 모델과 같이 mask-fill방식으로 사전학습되었다.

2021년 5월에는 SKT에서 KoGPT-Trinity를 공개하였다. KoGPT-Trinity는 SKT에서 자체 구축한 1.2B의 Ko-DAT dataset으로 학습되었으며, 모델의 크기도 1.2B이다. vocab size는 51,200이며 next token prediction으로 사전학습되었다. 이와 비슷한 시기에 Naver에서 HyperCLOVA를 발표하였다. HyperCLOVA는 뉴스, 카페, 블로그, 지식in, 웹문서, 댓글 등 네이버 내 검색이 허용된 문서와, 모두의 말뭉치, 한국어 위키피디아 등 다양한 문서로부터 추출한 데이터로 학습되었으며, 학습에 사용된 데이터에는 561.8B의 토큰으로 구성되어 있다. 또한 1.3B, 6.9B, 13.0B, 39.0B, 82.0B 등 다양한 크기의 모델이 존재하나 소스와 모델은 공개되어 있지 않다. 또한 KB 국민은행에서 경제/금융 도메인에 특화된 모델인 KB-ALBERT[14]를 2021년 6월에는 발표하였다. 2021년 11월에는 KLUE-BERT[26], KoGPT, ET5 모델이 공개되었다. KLUE-BERT는 한국어

8) https://aiopen.etri.re.kr/service_dataset.php

9) https://ko.wikipedia.org/wiki/

10) https://namu.wiki/

11) https://commoncrawl.org/

12) https://corpus.korean.go.kr/

13) https://www1.president.go.kr/petitions

14)https://github.com/KB-AI-Research/KB-ALBERT

Architecture	Name	Training objective	Tokenizer	Vocab size	Corpus	# Params	Published by
Encoder	KoBERT	MLM,NSP	Sentence-Piece	8002	Korean Wikipedia	92M	SKT
	KorBERT	MLM,NSP	Morpheme, WordPiece	30,349 (Morphemes), 30,797 (WordPiece)	News, Encyclopedia	110M	Etri
	HanBERT	MLM,NSP	Moran	54,000	General document, Patent document	128M	TwoBlock AI
	KRBERT	MLM,NSP	WordPiece	16,424 (Character), 12,367 (Subcharacter)	Korean Wikipedia, News	99M (character), 96M (sub-character)	Seoul National University
	KLUE-BERT	MLM,NSP	Morpheme-based subword	32,000	Modo-Corpus, NamuWiki, CC-100-Kor, News, petition	111M	klue project
	DistilKoBERT	MLM,NSP	Sentence-Piece	30,522	Korean Wikipedia, NamuWiki, News, etc.	27.8M	individual (Jangwon Park)
	KoreALBERT	MLM,SOP	Sentence-Piece	32,000	Korean Wikipedia, Sejong-Corpus, Book Corpus, News, etc.	12M (base), 18M (large)	Samsung SDS
	KALBERT	MLM,SOP	Morpheme, BPE	47,473	Korean Wikipedia, NamuWiki, Web, Book Corpus, News, etc.	~	individual
	KB-ALBERT	MLM,SOP	~	~	~	~	KB Kookmin Bank
	KoELECTRA	RTD	WordPiece	35,000	Korean Wikipedia, NamuWiki, News, Message, Web page, etc.	112M	individual (Jangwon Park)
	KcBERT	MLM,NSP	WordPiece	30,000	Naver New Comments	109M	individual (Junbum Lee)
	KcELECTRA	RTD	WordPiece	30,000	Naver New Comments, Naver New Reply	124M	individual (Junbum Lee)
	KoBigBird	MLM,SOP	WordPiece	32,500	Korean Wikipedia, News, Modo-Corpus, Common Crawl, etc.	113.8M	individual (Jangwon Park)
Decoder	KoGPT2	NTP	Character BPE	51,200	Korean Wikipedia, NamuWiki, News, Movie Review, parsed korean commonCrawl data, etc.	125M	SKT
	KoGPT-Trinity	NTP	*	51,200	Ko-DAT dataset	1.2B	SKT
	KoGPT	NTP	*	64,512	~	6.0B	kakao brain
	HyperCLOVA	NTP	Morpheme- Aware Byte-level BPE	~	Korean Wikipedia, Modu-Corpus, Documents allowed to be searched on Naver (News, Blog, Web document and Comments)	1.3B, 6.9B, 13.0B, 39.0B, 82.0B	naver
Encoder - Decoder	KoBART	DAE	Character BPE	30,000	Korean Wikipedia, News, Book, Modu-Corpus, petition, etc.	124M	SKT
	KE-T5	mask-fill	Sentence-Piece	64,000	Copus, a mixture of Korean (A corpus built on its own and Modu-Corpus) and English.	247M	Korea Electronics Technology Institute (KETI)
	ET5	mask-fill, DAE	Sentence-Piece	45,100	Korean Wikipedia, Newspaper article, broadcast/movie/drama script	60M	Etri
~	EXAONE	~	~	~	~	1.3B, 13B, 39B, 175B	LG-research

Fig. 1. Figure of the summarizing the properties of the Korean PLM. MLM and NSP indicate Masked Language Model and Next Sentence Prediction, respectively. RTD, NTP and DAE indicate Replaced Token Detection, Next Token Prediction and Denoising Auto Encoder, respectively. For items that were not disclosed, it was indicated as ~, and the tokenizer built by itself was indicated as *.

벤치마크 데이터인 KLUE에서 베이스라인으로 사용되었던 모델로, 모두의 말뭉치, CC-100-Kor[15], 나무위키, 뉴스, 청원 등의 문서에서 추출한 63GB의 데이터로 학습되었다. Morpheme-based Subword Tokenizer를 사용하며, vocab size는 32,000이고 모델의 크기는 111M이다.

KoGPT는 kakao brain에서 공개한 모델로, GPT3 모델을 벤치마킹하여 만든 한국어 PLM이다. 200B의 토큰으로 학습된 6B의 초거대모델이며, vocab size는 64,512이다. GPT 모델들과 마찬가지로 transformer decoder구조로 이루어져 있다.

ET5는 T5에서의 mask-fill과 GPT3에서의 Next Token Prediction을 동시에 사전학습한 모델이다. 위키백과, 신문기사, 방송 대본, 영화/드라마 대본 등에서 추출한 136GB의 데이터로 학습하였다. SentencePiece tokenizer를 바탕으로 45,100의 vocab size를 가진다. 모델의 크기는 60M이며, 인코더-디코더 구조를 가진다.

마지막으로 2021년 12월 LG AI research에서 공개한 EXAONE은 텍스트, 음성, 이미지를 바탕으로 학습된 멀티모달(multimodal) 모델로서 이미지에서 텍스트, 텍스트에서 음성 등 자유자재로 변환할 수 있다. 모델의 크기는 1.3B, 13B , 39B, 175B 등으로 초거대 모델이며, 아직 공개되지 않았다.

2.2 학계에서 공개한 한국어 PLM 연구

학계에서는 서울대학교가 지난 2020년 8월 KRBERT[27]를 공개하였다. KRBERT는 기존 BERT 모델과 마찬가지로 MLM와 NSP 방식으로 사전학습된 모델이다. vocab 크기는 16,424 (character)와 12,367 (sub-character)이고 WordPiece tokenizer를 사용했다. 2천만개의 한국어 위키피디아 및 뉴스 문장으로 학습되었으며, 한국어에 특화되고 규모는 작은 BERT 모델을 만들어 공개하였다.

2.3 개인이 공개한 한국어 PLM 연구

개인이 공개한 최초의 한국어 PLM은 2019년 11월에 공개된 KALBERT[16]이다. 한국어 위키피디아, 카이스트 책 말뭉치[17] 및 세종 말뭉치[18] 데이터 6GB로 학습된 모델로, BPE, morph tokenizer가 사용되었다. ALBERT 모델과 마찬가지로 Masked Language Model과 Sentence Order Prediction 방법으로 사전학습되었다.

이후 2020년 1월에는 KoBERT와 동일한 tokenizer 및 사전학습 task를 사용한 DistilKoBERT가 공개되었다. 27.8M의 파라미터 크기를 가진 이 모델은 기존의 BERT 및 한국어기반의 BERT모델들 보다 크기가 작다. 한국어 위키피디아, 나무위키, 뉴스 등으로 이루어진 10GB의 데이터로 학습되었으며, vocab 크기는 30,522이다. 2020년 4월에는 ELECTRA의 한국어 버전인 KoELECTRA가 공개되었다.

기존의 한국어 PLM이 대부분 한국어 위키, 뉴스 기사 등 정제된 데이터를 바탕으로 학습된 모델임에 착안하여, 비교적 정제되지 않고 신조어 및 오탈자를 다량 포함하고 있는 한국어 댓글 (Korean Comment) 데이터로 학습된 한국어 PLM도 등장하였다. 2020년 7월에 공개된 KcBERT와 2021년 4월에 공개된 KcELECTRA이다. 두 모델 다 WordPiece tokenizer를 사용하며, vocab 크기는 30,000이다. KcBERT의 학습에 사용된 데이터의 크기는 12GB이며, BERT와 마찬가지로 사전학습에 Masked Language Model와 Next Sentence Prediction를 적용하였다. KcELECTRA는 KcBERT보다 많은 양 (17.3GB)의 데이터로 학습되었다. Replaced Token Detection으로 사전학습되었으며, 모델의 파라미터 크기는 각각 109M와 124M이다.

2021년 10월에 공개된 KoBigBird는 모두의 말뭉치, 한국어 위키, Common Crawl, 뉴스 데이터 등 70GB의 한국어 데이터로 학습된 BigBird[28] 모델이다. BigBird 모델과 마찬가지로, 기존 PLM 모델들 보다 8배 더 긴 sequence input을 다룰 수 있고 random attention 등 다양한 attention으로 구성된 sparse attention mechanism을 적용한 모델이다. 모두의 말뭉치, 한국어 위키, Common Crawl, 뉴스로부터 추출한 70GB의 데이터로 학습하였으며, WordPiece tokenizer를 사용하였다. vocab 크기는 32,500이며, masked language model와 next sentence prediction로 사전학습 되었다. 공개된 모델의 파라미터 크기는 113.8MB이다.

---

15) http://data.statmt.org/cc-100/

16) https://github.com/MrBananaHuman/KalBert

17) http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus

18) https://github.com/coolengineer/sejong-corpus

Architecture	mPLM	Training objective	Tokenization	Vocab size	Corpus	Prams	#languages	W/ ko
Enc	mBERT (uncased) [5]	MLM, NSP	WordPiece	105K	Wiki	167M	102	✓
	mBERT (cased) [5]	MLM, NSP	WordPiece	120K	Wiki	177M	104	✓
	distil-mBERT (cased) [45]	MLM, NSP	WordPiece	120K	Wiki	134M	104	✓
	XLM-17 [7]	MLM	BPE	200K	Wiki	570M	17	✓
	XLM-100 [7]	MLM, TLM	BPE	200K	Wiki	570M	100	✓
	XLM-R (base) [46]	MLM	SPM	250K	CC	270M	100	✓
	XLM-R (large) [46]	MLM	SPM	250K	CC	550M	100	✓
	mLUKE [30]	MLM, MEP	SPMd	250K	Wiki + CC	585M	24	✓
	infoXLM (base) [32]	MLM, TLM, XICO	SPMd	250K	CC	278M	94	✓
	infoXLM (large) [32]	MLM, TLM, XICO	SPMd	250K	CC	559M	94	✓
	XLM-E [33]	MLM, TLM, TRTD, MRTD	SPMd	250K	CC	278M	100	✓
	XLM-align [34]	MLM, TLM, DWA	SPMd	250K	Wiki + CC	278M	94	✓
	AMBER [47]	MLM, TLM, WA, SA	BPE	120K	Wiki + MT	172M	104	✓
	Unicoder [35]	MLM, TLM, CIWR, CIPC, CIMLM	BPE	250K	CC + MT	270M	104	✓
Enc-Dec	mBART-25 [37]	MLM, MDP	SPM	250K	CC	610M	25	✓
	mBART-50 [38]	MLM, MDP	SPM	250K	CC + Wiki	610M	50	✓
	XNLG [40]	MLM, DAE, XMLM, XAE	BPE	§	Wiki + UN corpus	§	§	χ
	mT5 (small) [41]	MLM-SC	SPM	250K	mC4	300M	101	✓
	mT5 (base) [41]	MLM-SC	SPM	250K	mC4	580M	101	✓
	mT5 (large) [41]	MLM-SC	SPM	250K	mC4	1.2B	101	✓
	mT5 (xl) [41]	MLM-SC	SPM	250K	mC4	3.7B	101	✓
	mT5 (xxl) [41]	MLM-SC	SPM	250K	mC4	13B	101	✓
	ProphetNet-X [48]	MLM, MSP	SPM	250K	Wiki + CC	616M	100	✓
	VECO (small) [42]	IS-MLM, CS-MLM	SPM	250K	CC + OPUS	247M	50	✓
	VECO (large) [42]	IS-MLM, CS-MLM	SPM	250K	CC + OPUS	662M	50	✓
	mT6 [43]	MLM, TLM, TPSC, TSC, PANT	SPM	250K	CC + MultiUN + OPUS + Wiki + IIT Bombay	§	§	✓

Fig. 2. Multilingual PLM with encoder and encoder-decoder structures. § indicates that an element is not specified in the references

3. Multilingual PLM

다국어 사전학습 모델은 여러 언어를 하나의 구조로 극복할 가능성을 보여준다. 이는 여러 언어의 공통 언어적 자질을 학습하여 저자원 언어에 대한 확장 가능성을 제공한다. 단일 저자원 언어 모델을 처음부터 학습하는 것은 언어 데이터의 양 문제로 처음부터 학습하는 것에 문제가 있지만, 다국어 사전학습 모델을 사용하면 고자원 언어를 이용하여 데이터 자원 문제를 어느 정도 해결할 수 있다.

한국어는 저자원 언어에 속한다. 저자원 언어 사전학습 모델을 학습하기 위한 대량의 단일 언어 데이터를 수집하는 것은 비교적 까다롭게 여겨진다. 이러한 이유로, 고자원 언어와 저자원 언어 사이의 언어적 특징(cross-linguality)를 잘 학습할 수 있다면, 고자원 언어의 표현을 바탕으로 한국어 중심의 태스크에서 성능

을 높일 수 있다. 하지만, 다국어 사전학습 모델을 이용한 접근법은 여러 언어의 어휘 사전을 구성해야 하므로, 상당히 큰 임베딩 계층을 요구한다. 이는 파라미터 수를 증가시켜 모델의 크기가 기하급수적으로 커짐을 의미하며, 여러 서비스에서 사용하기에 어려움을 초래한다. 예를 들어, 단일 영어 모델인 BERT의 경우 110M의 파라미터 수를 가지고 있는 반면, 다국어 모델인 mBERT는 BERT에 비해 1.5배 많은 167M의 파라미터를 가지고 있다. 이는 단일 언어 모델이 아닌 다국어 언어 모델을 사용한다면, 대략 1.5배 더 많은 컴퓨팅 리소스를 요구하는 것이다.

다국어 사전학습 모델은 트랜스포머의 디코더 계층 없이 인코더 계층으로 구성된 인코더 구조 모델과 인코더-디코더 구조의 사전학습 모델이 있다. 뿐만 아니라, 디코더 계층으로만 이루어진 빅모델 또한 다국어 모델로 분류할 수 있다. 본 섹션에서는 추후 XPT 모델과 한국어를 포함하는 다국어 사전학습 모델의 성능 비교를 위해 모델의 구조별로 다국어 사전학습 모델에 대해 자세히 분석한다.

Fig. 2는 한국어를 학습에 포함한 다국어 사전학습 모델에 대한 구조 및 학습에 대한 내용을 요약한 것이다. 크게 인코더와 인코더-디코더 구조로 구분하였으며 공개되어 있지 않은 정보들은 §로 표기하였다.

3.1 인코더(Encoder) 구조의 mPLM

인코더 계층으로만 구성된 다국어 사전학습 모델은 저자원 언어의 자연어 이해 (Natural Language Understanding, NLU) 태스크 성능을 높이기 위한 연구가 이루어지고 있다. 초기 다국어 사전학습 모델인 mBERT는 BERT와 동일한 방법으로 다국어 코퍼스를 학습했다[3]. BERT가 사용한 Masked Language Modeling (MLM), Next Sentence Prediction (NSP)등의 학습 방법으로 다국어를 학습하여 저자원 언어에 대한 성능을 높일 수 있는 것이 밝혀졌고, 이후의 연구는 다국어의 표현인 Cross-linguality를 효율적으로 학습하기 위해 다양한 사전학습 방법을 적용해왔다.

이러한 사전학습 방법에는 대표적으로 병렬 데이터셋의 원시 문장과 목표 문장을 연결하여 임베딩을 구성하는 지도학습 방법인 Translation Language Modeling (TLM)[5,29], 엔티티에 마스킹을 하여 이를 예측하는 지도학습 방법인 Masked Entity Prediction (MEP)[30], 다국어에 MLM을 적용하여 마스킹을 예측하는 비지도 학습인 Multilingual Masked Language Modeling (MMLM), 병렬 데이터셋에서의 시퀀스를 섞은 후 이를 복원하는 Alternating Language Modeling (ALM)[31], 원시 언어와 목표 언어의 시퀀스 정보의 유사도를 최대화하는 Cross-lingual Contrastive Learning (XlCO)[32], ELECTRA 사전학습 방법을 모방한 Translation Replaced Token Detection (TRTD), Multilingual Replaced Token Detection (MRTD)[33], 병렬 문장 사이의 단어 정렬을 학습하도록 하는 Cross-lingual Word Recovery (ClWR), 단어 사이 정렬과 정렬의 노이즈를 예측하는 Denoising Word Alignment (DWA)[34], 원시 문장과 목표 문장의 의미가 동일한지 분류하여 문장 사이의 Cross-linguality를 학습 하는 Cross-lingual Paraphrase Classification (CIPC), 다국어 문서에서의 MLM을 수행하여 document-level의 perplexity를 낮추는 Cross-lingual Masked Language Modeling (CIMLM)[35] 등의 다양한 사전학습 방법을 이용하는 다국어 사전학습 모델이 있다.

3.2 디코더(Decoder) 구조의 mPLM

디코더 계층만으로 이루어진 다국어 모델은 아주 많은 파라미터 수를 가지는 모델로, 대표적으로 GPT-3가 있다. GPT-3는 대량의 코퍼스에서 포함된 여러 언어로 딕셔너리를 구성하여 학습하였고, 대략 120개의 언어를 포함한다. 또한, GPT-3는 GPT-2와 같이 별도의 cross-linguality 표현을 학습하는 사전학습 방법 없이 단순히 Casual Language Modeling (CLM)만으로 cross-linguality를 높였다. GPT-3는 대부분 영어 토큰으로 학습됐지만, 데이터셋 내부에 여러 언어 등이 포함되어 번역 등의 태스크도 수행할 수 있어 다국어 모델로 분류된다. GPT-3는 in-context learning으로 추가적인 가중치 업데이트 없이 태스크 설명 (Task description)과 소수의 예시 (Examples) 만으로 추론하는 Few-shot inference로 태스크를 수행한다. 이는 미세조정 과정이 필요하지 않아 일반화 (Generalization)에서 최상의 평가를 받는다.

하지만, GPT-3와 같은 빅모델로 한국어와 같은 저자원 언어에 대한 리소스 문제를 극복하는 것에는 아직 문제점이 존재한다. 1) GPT-3는 175B의 크기로 대중적으로 사용할 수 있는 크기의 모델이 아니며, 2)

GPT-3의 학습 데이터셋 내 120개의 언어 중 92.5%의 단어는 영어인 반면, 학습에 포함된 한국어는 오직 0.01%로 영어에 굉장히 편향되어 있다. 이러한 이유로, 디코더 계층으로만 구성된 다국어 사전학습 모델을 저자원 언어에 사용하기는 아직 다소 무리가 있다[36].

3.3 인코더-디코더 (Encoder-decoder) 구조의 mPLM

인코더에서 cross-linguality를 학습하여 자연어 이해 태스크를 수행하는 것 외에도, 자연어 생성 (Natural Language Generation, NLG) 태스크를 수행하기 위한 인코더-디코더 계층을 모두 가진 다국어 사전학습 모델이 있다. 대표적으로 mBART[37,38] 및 MASS[39] 모델은 저자원 언어의 기계번역 성능을 높이기 위해 다중 언어에 문장에 손상을 준 후 원본 문장을 복원하는 Multilingual Denoising Pretraining (MDP)을 적용하여 다중 언어에 대한 표현력을 학습하였다.

XNLG[40]는 mBART와 유사하게 문장의 일부 토큰을 섞고 일부 토큰을 버리는 방법으로 문장에 손상을 준 뒤 문장의 원본을 복원하는 Denoising Auto-Encoding (DAE), 병렬 코퍼스를 사용하여 원시 문장과 목표 문장에 대한 토큰 예측하는 Cross-lingual MLM (XMLM), 원시 문장과 동일한 언어의 문장을 생성하는 DAE 방식의 Cross-Lingual Auto-Encoding (XAE) 사전학습 접근법을 제안했다. XNLG 모델은 MLM과 DAE 사전학습 태스크를 해결하면서 단일 언어 표현력 (monolingual representation)을 확장하면서, XMLM과 XAE 사전학습 태스크를 통해 다중언어의 cross-lingual 표현력을 확장하여 언어 간 다양한 자연어 생성 태스크의 성능을 높였다.

또한, 모델의 크기를 기하급수적으로 키우고, Span-Corruption Masked-Language Modeling (MLM-SC) 사전학습 태스크를 풀며, 데이터셋의 분포 스무딩을 통해 학습한 mT5[41], 단일언어에 대한 MLM (IS-MLM)으로 자연어 이해 태스크에 대한 성능을 높이고, 다중언어에 대한 MLM (CS-MLM) 태스크를 해결하여 Cross-lingual generation 태스크에 대한 성능을 높인 VECO [42], 병렬 쌍의 문장에 대해 마스킹된 span을 예측하는 Translation Pair Span-Corruption (TPSC), 단일언어의 문장에 대해 마스킹된 span을 예측하는 Translation Span-Corruption (TSC), 부분적으로 문장을 나눈 후 디코딩하는 형식의 학습 방법인 Partially non-autoregressive decoding (PANT)를 적용한 mT6[43] 등의 자연어 생성 태스크에서 높은 성능을 거둔 인코더-디코더 다국어 사전학습 모델이 있다.

4. Cross-lingual Post-Training (XPT) 방법론

본 논문은 Cross-lingual Post-Training (XPT)방법론을 이용하여 관계추출 모델에 적용하고 이를 한국어 PLM과 한국어 multilingual PLM과의 비교 분석을 진행하고자 한다. XPT의 모든 학습과정은 Fig. 3과 같다. 사전학습 과정에서 원시 언어 모델의 파라미터 중 유의미한 파라미터를 선별하여 목표 언어의 모델을 초기화한다. 그 후 적응계층, 임베딩 계층 그리고 출력 계층을 학습한 후, 인코딩 계층을 순차적으로 학습한다. 마지막으로 목표 태스크에 대한 미세조정을 진행한다. 이에 먼저 사전학습 과정을 단계별로 서술하고, 관계추출로의 미세조정 단계를 면밀히 서술한다.

4.1 사전학습

4.1.1 Seed Model 및 언어 선정

본 논문은 XPT 방법론을 이용하기 위한 seed 모델로는 영어 RoBERTa 모델로 선정한다.

Seed 모델의 성능이 좋을수록 전이학습 시에 더 좋은 성능을 보장한다. 즉 어떠한 언어를 사전 학습에 활용할 것인가에 대한 문제는 전이 학습 시의 성능과 직결된다. 우수한 성능은 곧 좋은 품질과 대용량 데이터로 학습한 모델일 가능성이 높다. 이에 본 논문은 데이터의 양도 가장 풍부하면서 모델도 아무런 제약 없이 전면 공개되어 있는 영어 RoBERTa를 기반으로 실험을 진행한다.

4.1.2 학습 파라미터 선정

전이학습시에 모든 파라미터를 재학습 시킬 필요는 없다. 학습에 도움이 되는 파라미터를 굳이 재학습하여 잘 학습된 가중치를 희석시킬 필요가 없으며, 학습이 필요만 부분만 재학습하여 학습의 효율성을 높여야 한다. 즉 학습에 도움이 되는 것들과 되지 않는 것들을 구분하는 작업이 선행되어야 한다.

사전학습된 RoBERTa 구조는 임베딩 계층, 인코딩 계층, 출력 계층의 3가지로 분류해볼 수 있다. 이 중

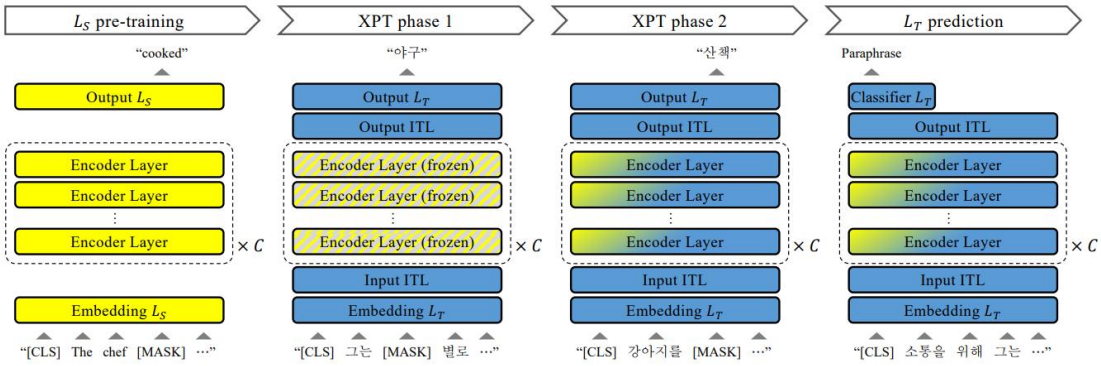


Fig. 3. Training process of XPT[10]. L_s denotes high-resource language L_T denotes low-resource language. ITL (implicit translation layer) refers to the Adpater layer

임베딩 계층과 출력 계층은 vocab에 종속되게 학습이 되며 인코딩 계층은 의미 공간에 투사된 단어 벡터들을 문맥을 바탕으로 맥락화(contextualize)하여 양방향 학습을 진행하게 된다. 즉 인코딩 계층에 파라미터들은 언어간 공유가 가능하며 재활용이 가능하다. 이에 본 논문은 인코딩 계층의 파라미터는 freeze를 진행하고, 임베딩 및 출력 계층의 파라미터는 무작위 값으로 초기화를 진행한다.

4.1.3 적응 계층

단순히 인코딩 계층의 파라미터는 freeze를 진행하고, 임베딩 및 출력 계층의 파라미터는 무작위 값으로 초기화를 진행하여 학습을 진행하는 것은 크게 2가지 문제점을 야기한다.

첫째 대응성의 문제이다. 원시 언어와 목표 언어의 단어들 사이에 1:1 대응 관계가 존재할 가능성은 희박하다. 언어간 단어 사전 구성은 명확히 다를 것이기 때문이다. 둘째로 언어학적 차이성의 문제이다. 두 언어 사이의 문법적 차이에 따라 어순, 의미표현 등이 다를 수 있는데, 단어 임베딩만을 무작위로 초기화 후 새로 학습하는 것만으로는 이러한 차이를 반영하기 한계점이 존재한다. 즉 이러한 차이를 학습시킬 수 있는 별도의 모듈 혹은 계층이 필요하다.

이에 본 논문은 기존 XPT[10] 논문에서 도입한 적응 계층을 추가하여 언어간 차이를 학습하고자 한다. 적응 계층은 인코딩 계층과 구조적으로 완벽히 일치하는 구조이며 입력 계층과 인코딩 계층 사이, 인코딩 계층과 출력 계층 사이 총 두 곳에 추가된다. 구조적으로 완벽히 일치하기에 계산 및 공간 복잡도의 안정성을 보장하여 학습의 효율성과 속도가 상승하게 된다.

4.1.4 전이 학습

본 논문에서 적용하는 전이 학습은 크게 2단계로 진행된다. 1단계에서는 적응 계층과 단어임베딩을 잘 학습시키는 것을 목적으로 한다. 즉 인코딩 계층의 파라미터는 고정시켜 학습에서 배제되며 적응, 임베딩, 출력 계층만 학습에 참여하게 된다. 인코딩 계층을 고정시킴으로써 기존에 영어 모델로 잘 학습된 파라미터를 보호해주는 효과를 볼 수 있다.

2단계에서는 임베딩, 인코딩, 출력, 적응 계층 모두 학습에 참여한다. 이는 인코딩 계층을 고정시킴으로써 남아있던 원시 언어 정보를 희석하는 역할을 하며, 영어 모델을 완벽하게 목표 언어 모델로 변환하는 역할을 수행한다.

4.2 미세조정

관계추출이란, 문장과 두 객체가 주어졌을때 문장내에서 두 객체가 가지는 관계를 분별하는 태스크를 의미한다. 관계추출은 트리플 (객체1, 관계, 객체2) 형태의 정보를 추출할 수 있기 때문에 다양한 자연어처리 응용분야에서 활용되고 있다. 특히, 비정형 문서들에서 유의미한 정보를 추출하는 자동정보추출(Automatic Information Extraction) 분야에서 관계추출은 지식 기반을 구축하는데 중요한 기술로 대두되고 있다. 따라서 본 논문에서는 Cross lingual 방식으로 전이학습한 모델의 미세조성 태스크를 관계추출로 하여 성능 측정 및 비교한다.

데이터의 개수 N 이 정의되어 있을 때 데이터셋 D는 수식 (1)과 같이 표현된다. 각 샘플은 문장 S_n, 두 엔티티 그리고 엔티티 간의 관계로 레이블링되어 있다. 한 문장 내의 토큰의 개수 l이 있을 때, 문장은 $S_n = [w_0^n, w_1^n, \ldots, w_l^n]$로 정의된다. 이때, 문장의 첫 번째 토큰인 w_0^n은 [CLS] 토큰을, 마지막 토큰인 w_l^n은 [SEP]토큰을 의미한다.

$$D = (\mathrm{S}_0, \mathrm{e1}_0, \mathrm{e2}_0, \mathrm{r}), \ldots, (\mathrm{S}_N, \mathrm{e1}_N, \mathrm{e2}_N, \mathrm{r}) \quad (1)$$

S_n을 입력으로 하여 모델의 hidden state vector를 구한다. 그 중 [CLS]에 해당하는 hidden state vector만 사용하여 Linear Layer를 통해 관계의 개수 차원으로 변형한 뒤, softmax를 통해 관계를 예측한다.

5. 실험 및 실험결과

5.1 데이터셋

원시 언어모델의 XPT학습을 위해 한국어 위키피디아를 사용한다. 위키피디아추출기 (Wikiextractor) [44]를 사용하여 문서만 추출하고, 추출한 문서를 문장 단위로 분리하고 SentencePiece tokenizer를 사용하여 토큰나이징한다. 그 결과, 4.19M의 문장을 추출하였으며 이를 학습데이터 4M, 검증데이터 100K 그리고 테스트데이터 88K로 구분하여 학습에 사용한다.

이렇게 학습된 XPT 모델의 관계추출 성능평가를 위해서는 KLUE-RE 데이터셋을 사용한다. KLUE-RE [26] 데이터셋은 32,470개의 학습데이터, 7,765개의 검증데이터 로 구성되어 있으며 테스트데이터셋을 공개하지 않았기 때문에 검증데이터를 테스트데이터로 하여 성능 측정 및 비교한다.

5.2 모델

본 논문에서는 BERT의 단점인 정적 마스킹 문제를 해결한 RoBERTa를 사용한다. RoBERTa-base모델에 한국어 위키피디아 데이터를 사용하여 XPT 학습을 진행한다. 학습된 모델에 대한 미세조정시 KLUE-RE 데이터셋을 사용하여 관계추출에 대한 성능을 측정한다.

해당 모델을 한국어 사전학습 모델과 비교하기 위해 KLUEBERT 모델을 사용하여 관계추출 성능을 측정한다.

또, 기존의 multilingual 모델과의 비교를 위해 mBERT에 대한 관계추출 성능을 함께 측정하여 비교한다.

전이학습은 배치사이즈(batch size) 256, 학습률(learning rate) 1e-4로 총 스탭(totoal step) 60K만큼 학습이 이루어진다. 미세조정에서는 배치사이즈(batch size) 32, 학습률(learning rate) 1e-5로 15에폭(epoch)만큼 학습한다.

5.3 평가 지표

관계추출에 대한 성능 평가 지표(metric)으로는 micro-F1을 사용한다. 미세조정시 사용하는 데이터셋인 KLUE-RE는 클래스별 불균형이 있기 때문에, 클래스별 샘플 수를 고려하는 micro-F1을 평가지표로 사용한다.

5.4 실험결과

XPT 모델 및 비교대상인 기준모델에 대한 KLUE-RE 데이터셋의 관계추출 성능은 Table 1에 있다. 실험 결과, Cross-lingual Post-Training (XPT) 모델의 경우 61.28로, 기존의 multilingual model인 mBERT와 비교하였을때 6.01 향상된 결과를 보인다. 이를 통해 적은 양의 목표 언어 데이터 만으로 대규모의 코퍼스로 학습된 사전학습 모델보다 우수한 성능을 보이는 것을 확인할 수 있다. 또, 한국어 사전학습 모델인 KLUE-BERT와 비교하였을 때 대략 3.85 증가하였다. 이는 방대한 양의 목표 언어 코퍼스로 학습된 모델을 사용하는 것보다, 원시 언어 사전학습 모델의 파라미터를 필요에 따라 재활용하여 학습하는 것이 더 우수하다는 것을 확인할 수 있다. 또한, 적응 계층을 활용하여 언어 간의 차이를 학습하게 하는 것이 학습의 효율성과 속도에 도움을 주고 있음을 알 수 있다. 따라서, 자원이 부족한 언어에서 사전학습 언어모델을 학습하기 위해 대규모의 코퍼스를 구축하는 것보다 자원이 풍부한 언어의 사전학습 모델을 활용하는 것이 더 효율적이라는 것을 알 수 있다.

Table 1. Experimental Results on KLUE-RE dataset

Model	Micro-F1
RoBERTa (with XPT)	61.28
KLUEBERT	57.43
mBERT	55.27

6. 결론

본 논문에서는 대용량의 코퍼스가 부족한 언어 환경에서 자원이 풍부한 원시 언어의 사전학습 모델을 활용하여 모델의 성능을 향상시킬 수 있는 Cross-lingual Post-Training (XPT) 방법론을 사용하여 RoBERTa 모델에 적용한 뒤 관계추출에 미세조정한 결과를 기존의 한국어 언어모델 KLUEBERT, mBERT와 비교하였다. 더불어, 본 논문은 국내외 학계와 기업에서 공개한 한국어 사전학습 언어모델과 한국어 Multilingual 사전학습 언어모델에 대한 분석을 진행하였다. 향후 연구로는 XPT 방법론을 인코더 구조의 모델에만 적용하는 것에서 벗어나 인코더-디코더구조의 모델에 적용하여 범용성을 확인할 예정이다.

REFERENCES

[1] R. Bommasani et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258.*

[2] T. Brown et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165.*

[3] B. Lester, R. Al-Rfou & N. Constant. (2021). The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691.*

[4] A. Vaswani et al. (2017). Attention is all you need. *In Advances in neural information processing systems* (pp. 5998-6008).

[5] J. Devlin, M. Chang, K. Lee & K. Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[6] Y. Liu et al. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

[7] G. Lample & A. Conneau. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291.*

[8] Lewis. M et al. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461.*

[9] K. Clark, M.-T. Luong, Q. V. Le & C. D. Manning. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003. 10555.*

[10] C. Lee, K. Yang, T. Whang, C. Park, A. Matteson & H. Lim. (2021). Exploring the Data Efficiency of Cross-Lingual Post-Training in Pretrained Language Models. *Applied Sciences, 11(5),* 1974.

[11] J. Park. (2020). KoELECTRA: Pretrained ELECTRA model for Korean. https://github.com/monologg/KoELECTRA

[12] J. Lee. (2020). Kcbert: Korean comments bert. *In Annual Conference on Human and Language Technology* (pp. 437-440).

[13] J. Lee. (2021). KcELECTRA: Korean comments ELECTRA. GitHub repository. Opgehaal van https://github.com/Beomi/KcELECTRA

[14] J. Park. (2019). DistilKoBERT: Distillation of KoBERT. GitHub repository. Opgehaal van https://github.com/monologg/DistilKoBERTc

[15] J. Park & D. Kim. (2021). KoBigBird: Pretrained BigBird Model for Korean (Version 1.0.0). doi:10.5281/zenodo.5654154

[16] B. Kim et al. (2021) What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. *arXiv preprint arXiv:2109.04650.*

[17] I. Kim, G. Han, J. Ham & W. Baek. (2021). KoGPT: KakaoBrain Korean(hangul) Generative Pre-trained Transformer. Opgehaal van https://github.com/kakaobrain/kogpt

[18] K. Airc. (2021. Mar). KE-T5: Korean English T5. Opgehaal van. https://github.com/AIRC-KETI/ke-t5

[19] Y. Wu et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609. 08144.*

[20] T. Kudo & J. Richardson. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808. 06226.*

[21] A. Radford et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1(8),* 9.

[22] R. Sennrich, B. Haddow & A. Birch. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508. 07909.*

[23] H. Lee, J. Yoon, B. Hwang, S. Joe, S. Min & Y. Gwon. (2021). KoreALBERT: Pretraining a Lite

BERT Model for Korean Language Understanding. *2020 25th International Conference on Pattern Recognition (ICPR),* 5551-5557. *IEEE.*

[24] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma & R. Soricut. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909. 11942.*

[25] C. Raffel et al. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910. 10683.*

[26] S. Park et al. (2021). KLUE: Korean Language Understanding Evaluation. *arXiv preprint arXiv:2105. 09680.*

[27] S. Lee, H. Jang, Y. Baik, S. Park & H. Shin. (2020). Kr-bert: A small-scale korean-specific language model. *arXiv preprint arXiv:2008. 03979.*

[28] M. Zaheer et al. (2020). Big Bird: Transformers for Longer Sequences. *NeurIPS.*

[29] I. Yamada, K. Washio. H. Shindo & Y. Matsumoto. (2019). Global entity disambiguation with pretrained contextualized embeddings of words and entities. *arXiv preprint arXiv:1909.00426.*

[30] R. Ri, I. Yamada & Y. Tsuruoka. (2021). mLUKE: The Power of Entity Representations in Multilingual Pretrained Language Models. *arXiv preprint arXiv:2110. 08151.*

[31] J. Yang, S. Ma, D. Zhang, S. Wu, Z. Li & M. Zhou. (2020). Alternating language modeling for cross-lingual pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence, 34,* 9386-9393.

[32] Z. Chi et al. (2020). Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007. 07834.*

[33] Z. Chi et al. (2021). Xlm-e: Cross-lingual language model pre-training via electra. *arXiv preprint arXiv:2106. 16138.*

[34] Z. Chi et al. (2021). Improving pretrained cross-lingual language models via self-labeled word alignment. arXiv preprint arXiv:2106.06381.

[35] H. Huang et al. (2019). Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909. 00964.*

[36] B. A. Richards et al. (2019). A deep learning framework for neuro- science. *Nature neuroscience, Vol. 22, No. 11,* pp. 1761-1770.

[37] Y. Liu et al. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics, 8,* 726-742.

[38] Y. Tang et al. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008. 00401.*

[39] K. Song, X. Tan, T. Qin, J. Lu & T.-Y. Liu. (2019). Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905. 02450.*

[40] Z. Chi, L. Dong, F. Wei, W. Wang, X.-L. Mao & H. Huang. (2020). Cross-lingual natural language generation via pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence, 34,* 7570-7577.

[41] L. Xue et al. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010. 11934.*

[42] F. Luo et al. (2020). Veco: Variable encoder-decoder pre-training for cross-lingual understanding and generation. a*rXiv preprint arXiv:2010. 16046.*

[43] Z. Chi et al. (2021). mT6: Multilingual Pretrained Text-to-Text Transformer with Translation Pairs. *arXiv preprint arXiv:2104. 08692.*

[44] G. Attardi. (2015). WikiExtractor. GitHub repository. Opgehaal van. https://github.com/attardi/wikiextractor

[45] V. Sanh, L. Debut, J. Chaumond & T. Wolf. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv, Vol. abs/1910.01108.*

[46] A. Conneau et al. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116.*

[47] J. Hu, M. Johnson, O. Firat, A. Siddhant & G. Neubig. (2020). Explicit alignment objectives for multilingual bidirectional encoders. *arXiv preprint arXiv:2010.07972.*

[48] W. Qi et al. (2021). Prophetnet-x: Large-scale pre-training models for english, chinese, multi-lingual, dialog, and code generation. *arXiv preprint arXiv:2104.08006.*

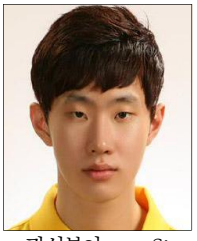
손 수 현(Suhyune Son) [학생회원]

· 2021년 8월 : 이화여자대학교 소프트웨어학부 컴퓨터공학전공 (공학사)
· 2021년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Information Extraction, Relation Extraction
· E-Mail : ssh5131@korea.ac.kr

박 찬 준(Chanjun Park) [학생회원]

· 2019년 2월 : 부산외국어대학교 언어처리창의융합전공 (공학사)
· 2018년 6월 ~ 2019년 7월 : SYS TRAN Research Engineer
· 2019년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Machine Translation, Data-centric AI, Grammar Error Correction, Deep Learning
· E-Mail : bcj1210@naver.com

이 정 섭(Jungseob Lee) [학생회원]



· 2021년 8월 : 동국대학교 정보통신공학전공 (공학사)
· 2021년 10월 ~ 현재 : 고려대학교 Human-Inspired AI 연구소
· 관심분야 : Simultaneous Translation, Dialogue System, Machine Translation, Speech Translation
· E-Mail : cy951011@gmail.com

심 미 단(Midan Shim) [학생회원]

· 2017년 3월 ~ 현재 : 경희대학교 생물학, 소프트웨어융합전공 (이학사, 공학사)
· 관심분야 : Dialogue System, Data Analysis, Numerical Reasoning
· E-Mail : hihello0426@gmail.com

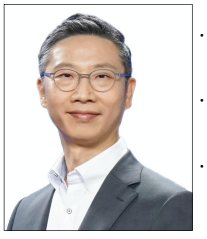
이 찬 희(Chanhee Lee) [정회원]

· 2013년 8월 : 서강대학교 컴퓨터공학심화(학사)
· 2016년 8월 ~ 2021년 8월 : 고려대학교 컴퓨터학과 석박사 통합과정 (공학박사)
· 2021년 11월 ~ 현재 : 네이버 연구원
· 관심분야 : 인공지능, 자연어처리, 전이학습
· E-Mail : chanhee0222@gmail.com

박 기 남(Kinam Park) [정회원]

· 2004년 2월 : 백석대학교 컴퓨터학과 (이학학사)
· 2006년 2월 : 한신대학교 컴퓨터정보학과(이학석사)
· 2011년 8월 : 고려대학교 컴퓨터교육학과(이학박사)
· 2011년 9월 ~ 현재 : 고려대학교 연구교수
· 관심분야 : 인공지능, 인지과학, 스마트교육
· E-Mail : spknn@korea.ac.kr

임 희 석(Heuiseok Lim) [종신회원]



· 1992년 : 고려대학교 컴퓨터학과 (이학학사)
· 1994년 : 고려대학교 컴퓨터학과 (이학석사)
· 1997년 : 고려대학교 컴퓨터학과 (이학박사)
· 2008년 ~ 현재 : 고려대학교 컴퓨터학과 교수
· 관심분야 : 자연어처리, 기계학습, 인공지능
· E-Mail : limhseok@korea.ac.kr