KE-T5: 한국어-영어 대용량 텍스트를 활용한 이중언어 사전학습기반 대형 언어모델 구축

신사임○, 김산○, 서현태
한국전자기술연구원 인공지능연구센터
{sishin, kimsan0622, dchs504}@keti.re.kr

Construction of bilingually pre-trained language model from large-scaled Korean and English corpus

Saim Shin○, San Kim○, Hyeon-Tae Seo
Korea Electronics Technology Institute Artificial Intelligence research Center

요 약

본 논문은 한국어와 영어 코퍼스 93GB를 활용하여 구축한 대형 사전학습기반 언어모델인 KE-T5를 소개한다. KE-T5는 한국어와 영어 어휘 64,000개를 포함하는 대규모의 언어모델로 다양한 한국어처리와 한국어와 영어를 모두 포함하는 번역 등의 복합언어 태스크에서도 높은 성능을 기대할 수 있다. KE-T5의 활용은 대규모의 언어모델을 기반으로 영어 수준의 복잡한 언어처리 태스크에 대한 연구들을 본격적으로 시작할 수 있는 기반을 마련하였다.

주제어: KE-T5, T5, 사전학습, 언어모델

1. 서론

최근 자연어처리 연구는 사전학습 기반 언어모델(Pre-trained language model)을 통한 언어처리 모델의 성능 고도화가 빠르게 진행되고 있다. 사전학습 기반의 언어모델이란, 대규모의 비정형 텍스트를 깊은 트랜스포머 (Transformer) 모델링 방식을 통하여 어휘 간의 범용적인 의미와 그 차이를 학습시킨 대형 신경망 학습 모델링 방식을 말한다. 해외 인공지능 기술 연구를 선도하는 기업들이 경쟁적으로 다양한 사전학습기반 언어모델 알고리즘을 연구하여 발표하고, 학습된 모델의 규모를 주기적으로 대형화하여 공개하고 있다. 그러나, 외국 대기업들의 관심 언어가 아닌 비영어권 언어의 경우에는 공개되는 모델의 구축 규모가 영어모델에 비교하여 크지 않기 때문에, 해당 언어를 위한 다양한 언어처리 성능의 고도화에 어려움이 있다.

본 논문에서는 다양한 한국어 언어처리 태스크를 범용화하고 고도화하는데 이바지할 수 있는 대형 언어모델인 KE-T5 (Korean-English T5)의 구축과정을 설명하고, 공개된 모델의 성능을 소개한다.

2. 관련 연구

2.1. 사전학습기반 언어모델 구축 알고리즘

사전학습기만 언어모델은 대용량 텍스트 데이터로부터 가상 (Proxy task)를 통한 자기 지도 (Self-supervised) 학습으로 범용적 의미 표현을 사전학습 (Pre-training)하고, 다양한 응용 태스크에 활용하는 언어처리를 위한 모델이다 [1]. 대형 언어모델 구축을 위한 사전학습 알고리즘은 크게 언어이해를 위한 task들에서 높은 성능을 보인다고 알려진 인코더 (Encoder) 기반 모델링 방식, 생성을 위한 디코더 (Decoder) 기반 모델링 방식, 그리고 인코더와 디코더를 모두 포함하는 sequence to sequence 구조 기반의 모델링 방식으로 구분할 수 있다. 대표적인 모델링 알고리즘으로는 인코더 모델링을 위한 XLNet [2], BERT (Bidirectional Encoder Representations from Transformers) [3], RoBERTa (A Robustly Optimized BERT Pretraining Approach) [4], ALBERT (A Lite BERT) [5], ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [6], SpanBERT와 디코더 모델링 알고리즘인 GPT [7, 8], 그리고 인코더와 디코더 모델을 포함하는 BART [9], T5 (Text-To-Text Transfer Transformer)등이 발표되어 활용되고 있다 [10].

2.2. 한국어를 위한 사전학습기반 언어모델

국내 학계에서는 기존 알고리즘의 한계를 보완하는 기술변화가 매우 빠르게 진행 중이며, 산업계에는 높은 성능과 학습데이터 구축비용 절감으로 상용화에 대한 기대가 높다. 인코더 모델로는 ETRI의 KorBERT[1], SKT의 KoBERT[2], 투블럭AI의 HanBERT[3], 삼성전자의 KoELECTRA[4], 금융도메인 데이터 처리 지원을 위한 KB-ALBERT[5]가 공개되고 기술이 이전되며 다양한 연구와 시스템에서 활용되고

1) https://aiopen.etri.re.kr/service_dataset.php
2) https://github.com/SKTBrain/KoBERT
3) https://github.com/monologg/HanBert-Transformers
4) https://github.com/monologg/KoELECTRA
5) https://github.com/KB-Bank-AI/KB-ALBERT-KO

있다. 한국어 디코더 모델은 SKT에서 공개한 KoGPT2[6]가 널릴 활용되고 있고, 인디코더 모델의 경우 네이버와 SKT에서 구축되어 공개한 T5 기반 한국어 언어모델[7]이 있다.

그러나, 기존의 공개된 한국어 언어모델의 경우는 구축된 코퍼스의 규모와 모델 파라미터 수가 크지 않아서, 적은 어휘 수와 어휘 별 참조하는 콘텍스트 규모로 인해 같은 알고리즘으로 구축되어 공개된 영어모델과 비교하면 기대 성능이 높지 않다. 또한, 중소기업이나 학교 및 비영리 연구기관의 경우는 구축을 위한 고비용의 한국어 코퍼스 수집비용과 연산환경 확보 비용으로 양질의 대규모 모델의 구축과 활용에 어려움을 겪고 있다.

3. KE-T5 구축 방법

KE-T5는 구글에서 개발하여 발표한 T5 알고리즘을 적용하여 한국어와 영어 코퍼스를 이용하여 구축한 범용적인 한국어 사전학습 모델이다. KE-T5는 한국어 및 영어 약 64,000개의 어휘를 포함한 모델로 총 92.92 GB의 한국어와 영어 코퍼스를 모델 규모 및 활용 목적에 따라 다양하게 선택하여 성능을 고도화할 수 있도록 다양한 모델을 구축하여 공개하였다. 공개된 모델의 규모 또한 Small, Base, Large로 다양하다.

종류		확보한 코퍼스 규모					정규화 코퍼스 규모	합계
한국어	자체 확보	70 (3억 7천만 문장)					72	92.92
	NIKL[8]	신문	문어	구어	메신저	웹		
		15.6	9.59	7.2	0.638	4.1		
영어	Real-News[9]	21						

표 1 KE-T5 구축에 활용한 코퍼스 상세 (단위: GB)

모델 구축을 위한 한국어 비정형 텍스트의 전처리 과정에는 정규 표현식을 기반으로 문서에 포함된 태그, 욕설과 비속어, 문맥과 연관 없는 각주, 수식 표현 등을 삭제하였다. 패턴 기반 자동화된 전처리 과정 뒤에는 반복적인 임의 추출을 통한 검토 과정을 통해 텍스트 전처리를 수행하였다. 전처리 과정에 활용한 tokenizer는 구글의 sentencepiece 기법을 활용하여 영어와 한국어 어휘를 한꺼번에 처리하였다 [11].

모델의 학습을 위한 mini-batch 크기는 256으로 수행하였으며, 각각의 모델 크기 별 학습한 step 수는 Small 모델은 60만에서 5백50만 번, Base 모델은 60만에서 2백30만 번, Large 모델은 60만에서 2백 2십만 번의 학습을 통해 모델을 구축하였다. 모델의 학습은 구글의 TPU (Tensor Processing Unit) 클라우드에서 수행하였으며, 8대의 TPU v3를 기반으로 모델 학습에 필요한 시간은 가장 대규모 모델의 경우 2달을 소요하였다.

6) https://github.com/SKT-AI/KoGPT2
7) https://github.com/seujung/kolang-t5-base
8) https://corpus.korean.go.kr/
9) https://github.com/rowanz/grover/tree/master/realnews

4. 한국어 언어처리 기반 성능 검증

4.1. 평가 환경

구축된 언어모델의 성능 검증을 위해 언어모델에 다양한 downstream 모델을 추가 학습하여, KE-T5의 성능을 간접적으로 평가하였다. 평가에 활용한 학습데이터는 다음과 같다.

- KLUE: 관계추출 (RE: Relational Extraction), 자연어 추론 (NLI: Natural Language Inference), 토픽 인식 [12]
- NIKL: 문장 연관성 (CoLA: Corpus of Linguistic Acceptability), 이진 의도분류 (NSMC), 유사 질문 인식, 요약, 자연어 추론, 유사텍스트 인식 (STS), 증오언설 인식
- KorQuAD 1.0[10]
- TED 번역 데이터: 영-한, 한-영

4.2. 언어이해 분야 처리 성능

명시된 평가 데이터셋에서 언어이해 task들을 위한 학습데이터들을 적용하여 finetuning을 수행한 KE-T5 기반 downstream 모델의 성능평가 결과는 다음과 같다.

task / size	CoLA	NSMC	유사질문인식	
	Mattew's [13]	정확도	F1 정확도	정확도
small	-3.72	87.9	87.9	91.5
base	12.51	88.95	**93.7**	91.49
large	**13.31**	**89.7**	89.74	**92.52**
task / size	NLI	STS		증오언설인식
	정확도	Pearson [14]	Spearman [15]	정확도
small	73.41	78.19	77.9	60.65
base	78.67	80.02	79.73	**64.14**
large	79.76	83.65	83.25	62.82

표 2 모두의 말뭉치 기반 downstream 모델 성능

task / size	RE	NLI	토픽 인식
	F1 정확도	정확도	정확도
base	**73.45**	85	85.58
large	69.59	**89.43**	**86.42**

표 3 KLUE 기반 downstream 모델 성능

기존의 인코더 기반의 언어모델링 알고리즘과 달리, KE-T5에서 활용한 T5 모델링 방식은 전이학습 파라다임을 기반으로 하나의 모델로 다수의 task 학습과 처리가 가능하도록 설계되었기 때문에, 모든 평가 결과가 하나의 모델로 다양한 task를 수행한 성능이다. 범용적인 하나의 모델 구축이 가능하면서도, 다양한 task에서 최고 수준의 성능을 보임을 알 수 있다.

4.3. 언어 표현 분야 처리 성능

수행한 성능평가 중 인코더를 포함한 언어 표현 관련

10) https://korquad.github.io/KorQuad%201.0/

task를 통한 성능은 다음과 같다.

task / size	KorQuAD	
	EM[16]	F1 정확도
small	88.39	87.9
base	91.11	88.95
large	**92.06**	**89.7**

표 4 발췌식 QA를 통한 downstream 모델 성능

모델 크기 기반 성능 추이를 살펴보면, base 쿠기의 모델에서 일부 태스크가 최고 성능을 보이고 있다. 이는, large 크기 모델의 구축 시에는 모델의 규모에서 최대한 학습할 수 있는 충분한 규모의 비정형 코퍼스의 확보가 더 좋은 성능의 대형 모델 구축에 필수요건이라는 것을 알 수 있다.

task / size	영-한		한-영	
	Rouge-1[17]	Rouge-2[17]	Rouge-1	Rouge-2
small	10.02	2.07	39.19	19.78
base	**12.03**	2.81	44.12	19.76
large	11.45	**2.96**	**44.52**	**20.21**
task / size	summary		topic	
	Rouge-1	Rouge-2	Rouge-1	Rouge-2
small	38.85	18.65	48.79	32.61
base	**40.86**	19.58	50.71	35.43
large	40.54	**20.04**	**55.52**	**37.72**

표 5 TED 및 NIKL 기반 downstream 모델 성능

5. 결론

본 논문은 한국어 코퍼스를 기반으로 구축한 T5 범용 언어모델인 KE-T5를 소개하였다. 본 모델은 githurb를 통해 apache 라이센스로 사전학습 모델과 downstream 모델이 공개되었다.[11] KE-T5는 현재까지 공개된 한국어 언어모델 중 가장 큰 어휘수와 모델 규모로, 이를 통한 한국어 처리 성능은 다양한 한국어 task에서 안정적이면서도 가장 높은 성능을 기록하였다.

감사의 글

이 논문은 2021년도 정부 (과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (S1601-209-1034, 정서적 안정을 위한 인공지능기반 공감 서비스 기술 개발).

참고문헌

[1] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, Matthieu Cord, "Boosting Few-Shot Visual Learning With Self-Supervision", IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[2] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. L, "XLNet: Generalized Autoregressive Pretraining for Language Understanding", arXiv, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv, 2018.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, arXiv, 2019.

[5] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut, “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations”, arXiv, 2019.

[6] Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning, “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators”, International Conference on Learning Representation (ICLR) 2020.

[7] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy, “SpanBERT: Improving Pre-training by Representing and Predicting Spans”, Transactions of the Association for Computational Linguistics (TACL), 2020.

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, “Language Models are Unsupervised Multitask Learners”, 2018.

[9] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”, arXiv, 2019.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”, Journal of Machine Learning Research (JMLR), 2020.

[11] Rico Sennrich, Barry Haddow, Alexandra Birch, “Neural Machine Translation of Rare Words with Subword Units”, Annual meeting of the Association for Computational Linguistics, 2016.

[12] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, Kyunghyun Cho, “KLUE: Korean Language Understanding Evaluation”, arXiv, 2021.

11) https://github.com/AIRC-KETI/ke-t5

[13] Matthews, B. W., "Comparison of the predicted and observed secondary structure of T4 phage lysozyme", Biochimica et Biophysica Acta (BBA) - Protein Structure, 405 (2): 442-451, 1975.
[14] Pearson, Karl, "Notes on regression and inheritance in the case of two parents". Proceedings of the Royal Society of London. 58: 240-242, 1985.
[15] Myers, Jerome L, Well, Arnold D, “Research Design and Statistical Analysis”, Lawrence Erlbaum, 2003.
[16] P. Rajpurkar, et al, “Squad: 100,000+ questions for machine comprehension of text”, arXiv, 2016.
[17] Lin, Chin-Yew, “ROUGE: a Package for Automatic Evaluation of Summaries”, Workshop on Text Summarization Branches Out (WAS), 2004.