Long-KE-T5: 긴 맥락 파악이 가능한 한국어-영어 언어 모델 구축

김산°, 장진예, 정민영, 신사임
한국전자기술연구원, 인공지능연구센터
{kimsan0622, jinyea.jang, minyoung.jung, sishin}@keti.re.kr

Long-KE-T5: Korean-English Language model for Long Sequences

San Kim°, Jinyea Jang, Minyoung Jeung, Saim Shin
Korea Electronics Technology Institute, Artificial Intelligence Research Center

요 약

이 논문에서는 7,400만개의 한국어, 영어 문서를 활용하여 최대 4,096개의 토큰을 입력으로하고 최대 1,024개의 토큰을 생성할 수 있도록 학습한 언어모델인 Long-KE-T5를 소개한다. Long-KE-T5는 문서에서 대표성이 높은 문장을 생성하도록 학습되었으며, 학습에 사용한 문서의 길이가 길기 때문에 긴 문맥이 필요한 태스크에 활용할 수 있다. Long-KE-T5는 다양한 한국어 벤치마크에서 높은 성능을 보였으며, 사전학습 모델링 방법이 텍스트 요약과 유사하기 때문에 문서 요약 태스크에서 기존 모델 대비 높은 성능을 보였다.

주제어: 언어 모델, 생성 언어 모델, 긴 문맥 입력, KE-T5

1. 서론

BERT[1], GPT[2]와 같은 사전학습 언어 모델이 나온 뒤로 여러 기관에서 한국어 처리를 위하여 한국어 언어 모델들을 구축하여 공개하였다. 언어 모델 공개 초기에는 모델을 최대 512 토큰 길이의 문서로 학습하여 공개하였으나, 긴 문맥 이해에 대한 필요성이 늘어나며 최대 입력 토큰 길이가 점점 늘어난 모델들이 공개되고 있다. KoBigBird[3]와 같이 최대 입력 토큰 길이가 4,096로 늘린 한국어 모델들이 공개되었으나 인코더만 사용하여 학습되었기 때문에 생성 작업이 불가능한 한계점이 있다. 이러한 제한사항을 해결하기 위하여 본 논문에서는 최대 입력 토큰 길이가 4,096인 인코더-디코더 기반 언어 모델인 Long-KE-T5[1]의 구축과정을 소개한다. 또한 다양한 한국어 자연어처리 벤치마크를 활용하여 정량적으로 모델의 성능을 평가한다.

2. 관련연구

Transformer의 주의집중(Attention) 연산과정에서 메모리가 토큰 길이의 제곱에 비례하여 사용되기 때문에 최대 입력 토큰 길이를 늘리면 학습시 자원이 많이 필요하다. 따라서 토큰 길이 증가에 따른 메모리 사용량 및 연산량을 줄이기 위하여 다양한 Sparse 주의집중 알고리즘들이 개발되었다. Longformer[4]의 같은 경우, 윈도우 주의집중(Window attention, Local attention)과 전역 주의집중(Global attention)을 조합하여 토큰 길이 대비 메모리 사용량을 낮췄다. BigBird[5]는 Longformer의 아이디어에 랜덤 주의집중(Random attention)을 추가하여 Longformer보다 메모리 사용량은 늘어나지만 문맥 파악 능력을 높였다.

Long-T5[6]의 경우 윈도우 주의집중에 일정 토큰 길이마다 평균 벡터를 구하고 평균 벡터들을 모든 쿼리 벡터들이 주의집중하는 단기 전역 주의집중(Transient global attention)을 통하여 메모리 사용량을 낮추었다. 입력 토큰 길이가 N이라고 할때, 전역주의집중이 N개의 벡터에 주의집중하는 것과 달리 단기 전역 주의 집중은 K개의 벡터마다 평균 벡터를 구하고 N/K개의 평균 벡터에 주의집중을 한다. 따라서 Long-T5는 윈도우 주의 집중과 단기 전역 주의집중을 사용하기 때문에 윈도우 크기가 r이라고 할때, 쿼리 벡터당 r + N/K개의 벡터에 주의집중 연산을 수행한다.

3. Long-KE-T5

언어 모델링 학습을 위하여 3,605 만개의 한국어 문서와 3,816개의 영어 문서를 수집하였다. 이렇게 수집된 문서들을 문서별로 문장 단위로 자른 뒤, ROUGE[7] 점수를 이용하여 20%의 문장을 출력 텍스트로, 나머지 텍스트를 입력 텍스트로 지정하였다. 특정 반복 단어로 인해 점수가 높게 계산되는을 방지하기 위하여, ROUGE 점수 계산할때 N-그램 갯수를 가중치로 사용하지 않고 1로 고정하였다. 문장 선택시, 이전에 선택한 문장에 인접한 문장 중 ROUGE 점수가 높은 쪽으로 확장하는 것이 아닌 독립적으로 ROUGE 점수를 계산하여 20%의 문장을 선택하였다.

데이터 전처리 단계에서 데이터를 입력 텍스트의 토큰 길이로 정렬한뒤, 표 1과 같이 토큰 길이에 따라 배치크기를 다르게 하여 배치 단위로 데이터를 구성하였다. 토큰 길이에 따라 배치 크기를 다르게하면 학습 속도를 향상시킬 수 있으며, 각 스텝당

[1] https://github.com/AIRC-KETI/long-ke-t5,

비교적 비슷한 양의 토큰을 학습 시킬 수 있다. 사전학습모델 학습은 데이터셋을 총 5번 반복하여 학습하였으며 Adafactor로 최적화 하였고, A100 80G GPU 8개를 이용하여 학습하였다.

표 1. 토큰 길이별 배치 크기 (단위: 개)

토큰 길이	배치 크기
128 이하	448
512 이하	96
1024 이하	48
2048 이하	22
4096 이하	8

4. 자연어 벤치마크 결과

다운스트림 태스크로는 한국어 요약 벤치마크들과 한영/영한 번역 벤치마크들을 활용하여 실험을 진행하였다. 모든 실험은 기존 인코더-디코더 기반 모델이며 최대 입력 토큰 길이가 512인 KE-T5[8]과 비교실험을 진행하였으며 KE-T5와 Long-KE-T5 모두 small 모델로 실험을 진행하였다. 모든 다운스트림 태스크는 데이터셋을 3번 반복 학습 후 성능을 측정하였으며, 생성시에는 빔서치를 활용하였고 빔 크기는 4로 설정하였다.

4.1 한국어 요약 벤치마크

한국어 요약 벤치마크로는 AIHub의 논문자료 요약[2], 문서요약 텍스트[3], 요약문 및 레포트 생성 데이터[4]를 활용하였다. 논문자료 요약 데이터셋에서는 섹션 요약을 제외하고 "학술논문-전체요약"과 "특허명세서-전체요약"을 실험하였고, 문서요약 텍스트 데이터셋에서는 법률, 잡지, 신문기사 데이터를 활용하여 요약 성능을 측정하였다. 표 2는 각 요약 벤치마크에서 KE-T5와 Long-KE-T5의 성능을 보여주며, R-1, R-2, R-L은 각각 Rouge 1 gram, Rouge 2 gram, Rouge LCS(Longest Common Subsequence)를 의미한다. 또한 BLEU[9]의 경우 BLEU 1-4 gram의 평균 점수이다. 측정결과 모든 성능지표에서 KE-T5보다 Long-KE-T5의 성능이 높게 나타났으며, 신문기사와 논문을 제외한 나머지 데이터들에서 성능차이가 컸다. 이는 Long-KE-T5가 KE-T5보다 긴 문맥을 볼 수 있어서 성능이 크게 높아진 것으로 보인다. 신문기사와 같은 경우 512 토큰 길이로도 전체 기사 내용을 볼 수 있어서 성능차이가 크지 않으며, AIHub 논문자료 요약의 논문 요약 데이터의 경우 데이터를 분석한 결과 대부분의 경우 요약 텍스트가 원문의 초반부와 매우 유사한 특성을 가지고 있어 성능차이가 크지 않다.

표 2. 요약 벤치마크 성능

한국어 요약 벤치마크						
	논문 - 전체			특허 - 전체		
	R-1	R-2	R-L	R-1	R-2	R-L
KE-T5	19.91	9.30	19.74	12.92	8.02	12.72
Long-KE-T5	**21.46**	**10.29**	**21.28**	**41.02**	**34.43**	**40.66**
	문서 - 법률			문서 - 잡지		
	R-1	R-2	R-L	R-1	R-2	R-L
KE-T5	1.76	0.41	1.69	7.97	1.43	7.94
Long-KE-T5	**7.29**	**1.36**	**7.26**	**25.26**	**8.87**	**24.96**
	문서 - 신문기사			레포트		
	R-1	R-2	R-L	R-1	R-2	R-L
KE-T5	45.84	20.24	43.90	10.72	1.75	10.63
Long-KE-T5	**48.54**	**22.48**	**45.83**	**15.09**	**3.28**	**14.94**

표 3. 번역 벤치마크 성능

영한 번역 벤치마크						
	식품			기술과학		
	BLEU	R-1	R-2	BLEU	R-1	R-2
KE-T5	**23.94**	**50.72**	**27.67**	**29.98**	**56.77**	**31.88**
Long-KE-T5	21.16	47.95	25.21	29.38	56.06	31.29
	사회과학			구어체		
	BLEU	R-1	R-2	BLEU	R-1	R-2
KE-T5	**12.81**	**20.19**	**7.88**	**28.35**	16.13	5.70
Long-KE-T5	11.52	19.79	7.56	29.24	**16.21**	5.70
한영 번역 벤치마크						
	식품			기술과학		
	BLEU	R-1	R-2	BLEU	R-1	R-2
KE-T5	19.40	55.86	38.39	35.49	66.84	50.77
Long-KE-T5	**25.23**	**61.85**	**45.63**	**43.95**	**73.54**	**58.97**
	사회과학			구어체		
	BLEU	R-1	R-2	BLEU	R-1	R-2
KE-T5	27.61	60.47	40.30	41.92	68.92	49.30
Long-KE-T5	**34.61**	**66.74**	**48.54**	**47.15**	**72.40**	**53.91**

4.2 한영/영한 번역 벤치마크

한영/영한 번역 벤치마크로는 AIHub의 한국어-영어 번역 말뭉치 기술과학[5], 사회과학 분야 데이터[6]와 전문분야 영-한 · 중-한 번역 말뭉치 식품 분야 데이터[7], 일상생활 및 구어체 한-영

[2] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=90

[3] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=97

[4] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=582

[5] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=124

[6] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=125

[7] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71262

번역 병렬 말뭉치 데이터[8]를 활용하였다. 표 3는 각 벤치마크에서 측정한 영한, 한영 번역 성능을 보여주며, 번역 태스크이기 때문에 BLEU 점수와 Rouge 점수를 사용하여 성능을 측정하였다. 영한 번역의 경우에는 거의 모든 벤치마크에서 KE-T5가 Long-KE-T5보다 높은 성능을 보였으나 성능차이가 작다. 하지만 한영 번역의 경우에는 모든 벤치마크에서 Long-KE-T5가 KE-T5보다 큰 차이로 높은 성능을 보였다.

5. 결론

이 논문에서는 긴 문맥을 볼 수 있는 Long-KE-T5의 구축 과정과 방법을 소개했다. Long-KE-T5는 4,096개의 토큰을 입력받을 수 있으며, 1,024개 토큰 길이만큼 문장을 생성할 수 있다. 긴 문맥을 볼 수 있기 때문에 긴 문맥 파악이 필요한 요약 태스크에서 성능이 높다는 것을 실험을 통해 확인하였다.

감사의 글

이 논문은 2023년도 정부 (과학기술정보통신부)의 재원으로 정보통신기획평가원(No. 2022-0-00320)의 지원을 받아 수행된 연구임

참고문헌

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Jun. 2019. [Online]. Available: https://aclanthology.org/N19-1423

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Vol. 33, pp. 1877–1901, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[3] J. Park and D. Kim, "Kobigbird: Pretrained bigbird model for korean," Nov. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5654154

[4] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv:2004.05150*, 2020.

[5] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Vol. 33, pp. 17 283–17 297, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf

[6] M. Guo, J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, and Y. Yang, "LongT5: Efficient text-to-text transformer for long sequences," *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 724–736, Jul. 2022. [Online]. Available: https://aclanthology.org/2022.findings-naacl.55

[7] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, pp. 74–81, Jul. 2004. [Online]. Available: https://aclanthology.org/W04-1013

[8] S. Kim, J. Y. Jang, M. Jung, and S. Shin, "A model of cross-lingual knowledge-grounded response generation for open-domain dialogue systems," *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 352–365, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.33

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[8] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71265