PALM 기반 한국어 T5 개선: 기계독해 및 텍스트 요약으로의 응용

박은환, 나승훈, 임준호, 김태형, 최윤수, 장두성
전북대학교, 한국전자통신연구원, KT

- References

[1]. Sequence-to-Sequence Learning with Neural Networks [Sutskever et al, 14’]

[2]. Neural Machine Translation by Jointly Learning to Align and Translate [Bahdanau et al, 16’]

[3]. Get To The Point: Summarization with Pointer-Generator Networks [See et al, 17’]

[4]. Multi-Style Generative Reading Comprehension [Nishida et al, 19’]

[5]. BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension [Lewis et al, 20’]

[6]. MASS: Masked Sequence-to-Sequence Pre-Training for Language Generation [Song et al, 19’]

[7]. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

[8]. PALM: Pre-Training an Autoencoding & Autoregressive Language model for Context-Conditioned Generation [Bi et al, 20’]

[9]. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding [Devlin et al, 19’]

[10]. Improving Language Understanding by Generative Pre-Training [Radford et al, 18’]

[11]. Language Models Are Unsupervised Multitask Learners [Radford et al, 19’]

[12]. BART를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의미역 결정 [홍승연 et al, 20’]

[13]. RoBERTa를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존파싱 [민진우 et al,19’]

[14]. KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋 [임승영 et al, 18’]

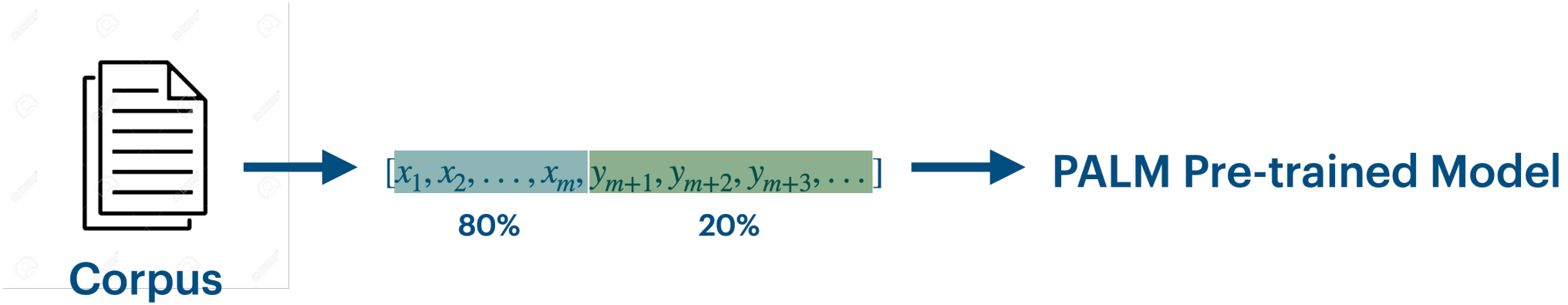
[15]. BERT를 이용한 한국어 기계 독해 [이동헌 et al, 19’]

[16]. SpanBERT를 이용한 한국어 자연어처리: 기계 독해, 개체 연결, 의존 파싱 [박은환 et al, 21’]

[17]. ROUGE: A Package for Automatic Evaluation of Summaries [Lin et al, 21’]

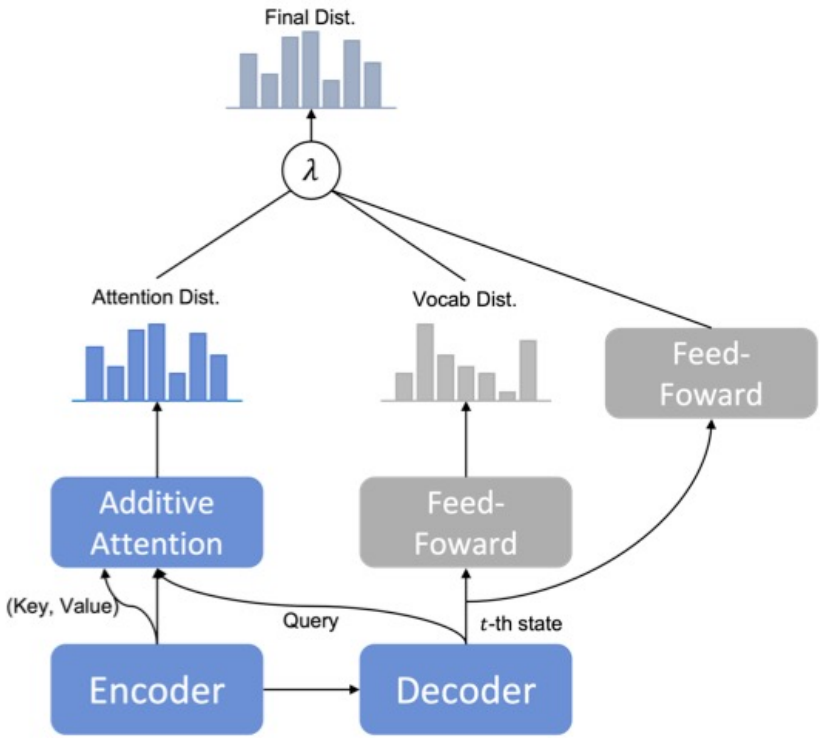
- 연구 개요
  - 최근 다양한 연구에서 [5, 6, 7, 8, 9, 10, 11]와 같은 언어모델(PLM, Pretrained Language Model)로 미세 조정 실험을 함.
    - 생성 관련 태스크에서는 T5, MASS, BART, GPT, PALM 등이 사용됨.
  - 본 연구에서는 PALM 언어 모델을 사전 학습하고 T5, BART 대비 어느정도 성능 향상되었는지 비교함.

- PALM



  - 기존 T5, BART 에서 마스킹한 X'를 복구하는 Objective 를 이용하여 사전 학습됨.
  - PALM은 문단의 80%를 X′(BERT Masking Scheme과 동일하게 마스킹) 인코더의 입력으로, 나머지 20% Y를 디코더의 입력으로 넣고 Y를 복구하는 Objective를 이용하여 사전 학습함. (데이터 집합은 한국어 위키피디아를 사용)
    - $L_{AR} = \sum_{(x,y)\in(X,Y)} \log \prod_{t=1}^{n} P(y_t | y_{<t}, x; \theta)$
    - $L(\theta) = L_{MLM} + L_{AR}$

- PALM



– 사전 학습 단계에서 [4]에서 제안되었던 Pointer-Generator Network 가 사용됨.

• 실험 결과

표 1. 한국어 기계독해 실험 결과: KorQuAD 1.0 (Dev Set)

Model	*EM*	*F*1
BERT-ETRI[15]	84.82	92.27
RoBERTa[13]	85.03	93.37
SpanBERT[16]	85.53	93.42
PALM-Base	78.48	88.27
RoPALM-Base	84.46	92.81

표 3. 평가 데이터 집합 Rouge-1, 2, L 및 EMFS 평가 결과

Model	Rouge-1	Rouge-2	Rouge-L
T5-Base	39.45	23.40	35.45
BART-Base	50.00	33.63	41.91
RoPALM-Base	54.64	34.89	45.89

- KorQuAD[14](질의 응답)와 AI-HUB의 신문기사 데이터 집합 (생성 요약)으로 실험을 진행
- KorQuAD
  - 기존 사전 학습된 RoBERTa[13]로 인코더를 초기화한 RoPALM-Base가 그렇지 않은 PALM-Base 보다 더 좋은 성능을 보여줌.
  - 비록 인코더만 초기화하더라도 디코더가 제대로 사전 학습됨을 보여줌.
- 신문기사 데이터 집합
  - {T5, BART}-Base 보다 RoPALM-Base 가 Rouge-{1, 2, L}에서 좋은 성능을 보여줌.

- 결론
  - T5, BART보다 RoPALM이 더 좋은 생성 요약 성능을 보여줌.
  - 본 연구의 성능을 추후 연구를 위한 베이스라인으로 삼을 수 있음.
  - 추후 연구:
    - 질문 생성(Question Generation), 검색기(Retriever) 등의 사실 검증 모듈을 이용한 정교하고 개선된 문서 생성 요약 연구 진행

감사합니다