Taesun **Whang**
NATURAL LANGUAGE PROCESSING & ARTIFICIAL INTELLIGENCE
(+82) 10-4905-1211 | taesunwhang@gmail.com | taesunwhang

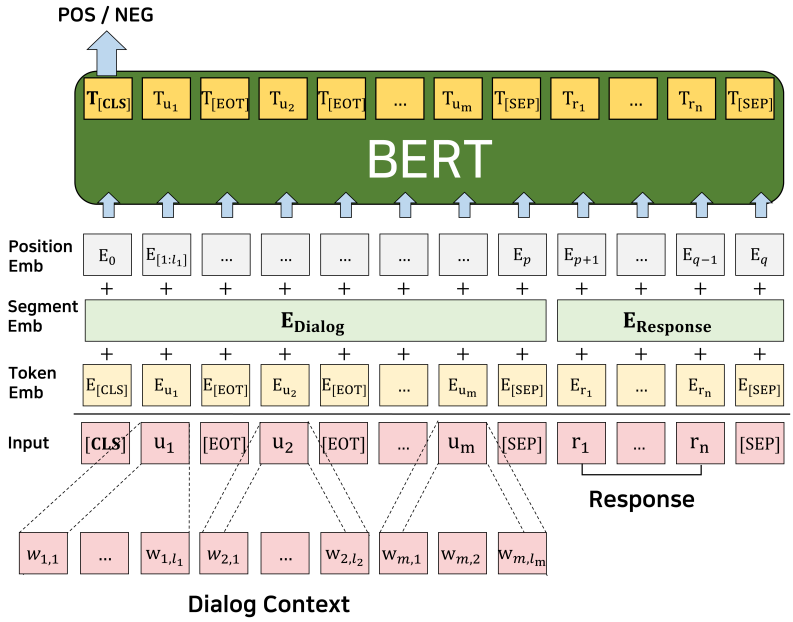
Multi-turn Response Selection

An Effective Domain Adaptive Post-Training Method for BERT in Response Selection

- **Whang, T.**, Lee, D., Lee, C., Yang, K., Oh, D., & Lim, H. (2020). An effective domain adaptive post-training method for BERT in response selection. In Proceedings of 22nd INTERSPEECH (pp. 1585-1589).
- Github: https://github.com/taesunwhang/BERT-ResSel

01. 연구 내용

- 대용량 코퍼스에 대해 언어 문맥을 고려하며 학습을 진행한 pre-trained language model BERT는 다양한 자연어처리 분야에서 state-Of-The-Art (SOTA)의 성능을 보여주었습니다. 뛰어난 성능을 보여준 비지도 학습의 목적 함수인 Masked Language Model (MLM)과 Next Sentence Prediction (NSP)에도 불구하고, 대화의 경우 일반적으로 BERT 모델이 학습했던 코퍼스 (e.g., English Wikipedia, Book Corpus)와는 다른 점이 존재합니다.
- 문어체의 문장 구조가 많고 일반적으로 많이 활용되는 단어 또는 문장들로 구성되어 있는 일반적인 코퍼스와 달리, 대화 코퍼스의 경우 문법 구조에 맞지 않는 구어체, 줄임말, 은어, 오타 등의 다양한 특징들이 존재합니다.
- 이러한 점을 보완하고자 본 연구에서는 대화 코퍼스를 위한 효율적인 BERT 추가 학습 (Post-Training) 방법을 제안하였으며, 이를 Response Selection task에 적용함으로써 대화를 위한 추가 학습이 뛰어난 성능 향상을 불러온다는 것을 입증하였습니다.
- 아래 그림은 본 연구에서 제안한 Response Selection을 위한 모델 구조도를 나타낸 것입니다. 모델의 입력으로는 대화와 응답 발화가 [SEP]을 통해 구분지어져 들어가게 되며, 대화의 각 발화 사이에는 [EOT]라는 새로운 토큰이 추가된 구조입니다. [EOT] 토큰을 추가함으로써 모델은 하나의 문장과 같이 들어간 구조에서 발화 간의 관계를 학습할 수 있게 되며, 이에 대한 효과는 뒤의 ablation 실험에서도 확인할 수 있었습니다.



02. 실험 및 결과

- 본 연구에서는 Response Selection에서 가장 많이 사용되는 Ubuntu Corpus V1(왼쪽)과 DSTC 7에서 공개한 Advising Corpus(오른쪽)에 대해 모델 성능 평가를 진행하였습니다. 기존 RNN 기반의 모델에서부터 Attention, Transformer 기반의 모델들까지 baseline으로 설정하고 성능을 비교하였습니다.

- BERT 관련 모델의 경우 Google에서 공개한 BERT$_{base}$, 본 연구에서 제안하는 방식으로 추가 학습을 진행한 BERT-DPT, Fine-tuning 단계에서 BERT의 상위 k개의 layer만 학습을 진행한 BERT-VFT, 마지막으로 Pointwise 방식으로 학습할 때 Positive 응답과 Negative 응답의 비율을 기존 1:1에서 1:4로 증가시킨 BERT-VFT(DA)에 대해서 실험을 진행하였습니다.
- 실험 결과, BERT-DPT가 BERT$_{base}$에 비해 두 데이터 모두에 대해서 Recall@1이 3.4% 향상된 것을 확인하였습니다. 뿐만 아니라, BERT-VFT, Data Augmentation도 BERT 모델의 효과적인 Fine-tuning을 가능하게 하였습니다.

Model	R_{10}@1	R_{10}@2	R_{10}@5
DualEncoder$_{rnn}$	0.403	0.547	0.819
DualEncoder$_{cnn}$	0.549	0.684	0.896
DualEncoder$_{lstm}$	0.638	0.784	0.949
DualEncoder$_{bilstm}$	0.630	0.780	0.944
MultiView	0.662	0.801	0.951
SMN	0.726	0.847	0.961
AK-DE-biGRU	0.747	0.868	0.972
DUA	0.752	0.868	0.962
DAM	0.767	0.874	0.969
MRFN	0.786	0.886	0.976
IoI	0.796	0.894	0.974
MSN	**0.800**	**0.899**	**0.978**
BERT$_{base}$	0.817	0.904	0.977
BERT-DPT	0.851	0.924	0.984
BERT-VFT	**0.855**	**0.928**	**0.985**
BERT-VFT(DA)	**0.858**	**0.931**	**0.985**

Model	R_{100}@1	R_{100}@10	R_{100}@50	MRR
Vig and Remma [9]	0.186	0.580	0.942	0.312
Chen et al. [18]	0.214	0.630	**0.948**	0.339
BERT$_{base}$	0.236	0.656	0.946	0.359
BERT-DPT	0.270	**0.668**	0.942	0.395
BERT-VFT	**0.274**	0.654	0.932	**0.400**
BERT-VFT(DA)	0.274	0.664	0.942	0.399

- 아래 표는 발화 간 관계를 학습하기 위해 추가한 [EOT] 토큰 및 BERT의 각 비지도 목적 함수에 대해 Ablation을 진행한 결과를 나타낸 것입니다.
- Ablation 결과 [EOT] 토큰의 추가와 관계 없이 MLM과 NSP를 함께 학습하였을 때, 가장 높은 성능을 보여주었습니다. 각 목적 함수를 비교하였을 때, 문장의 문맥 정보 학습에는 MLM이 많은 영향을 미치는 것을 본 실험을 통해 확인 하였으며, [EOT] 토큰의 추가가 대화 이해를 위한 BERT 추가 학습에 많은 영향을 준다는 것도 확인할 수 있었습니다.

Post-Training	Special Token	R_{10}@1	R_{10}@2	R_{10}@5	MRR
NSP	without EOT	0.816	0.903	0.977	0.884
MLM		0.834	0.916	0.981	0.896
MLM + NSP		**0.839**	**0.920**	**0.982**	**0.900**
NSP	with EOT	0.819	0.906	0.978	0.886
MLM		0.838	0.918	0.982	0.899
MLM + NSP		**0.851**	**0.924**	**0.984**	**0.907**