SpanBERT : Improving Pre-training by Representing and Predicting Spans

2019. 11. 16

발표자 : 김정미

Contents

1 . Introduction

- 많은 NLP 과제들이 두 개 이상의 span사이의 관계에 대한 추론을 포함
  - 예) “Which NFL team won Super Bowl 50? “라는 질문이 주어졌을 때, “Broncos”라는 단어를 알고 “Denver”를 유추하는 것보다 “Denver Broncos”를 유추하는 것이 훨씬 어려움.

- 이 논문은 BERT에 영감을 받아 span level의 pre-training 접근 방식을 소개

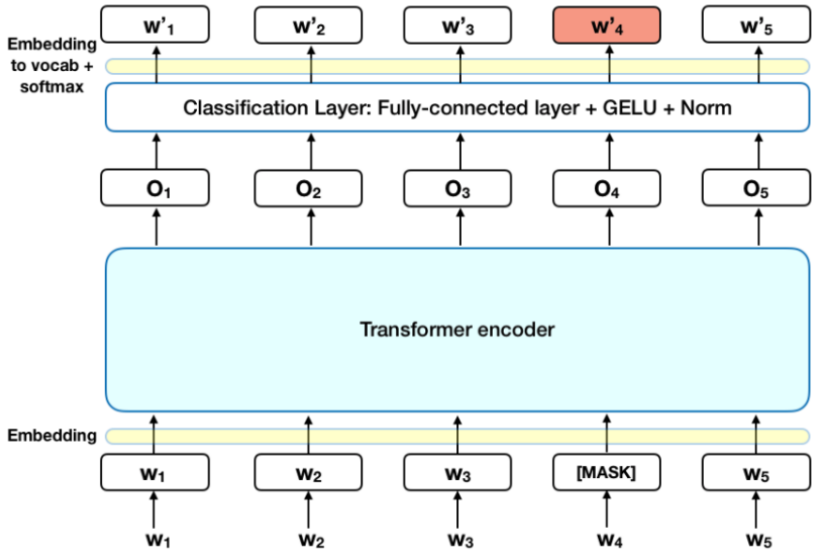
- BERT와의 차이점
  - Masking random contiguous span
  - span-boundary objective (SBO)
    - 마스킹 된 span의 경계로 부터 span을 유추

- SpanBERT를 구현하면서, Single segment를 pre-training하고 NSP를 제외하는 것이 성능에 좋은 영향을 준 것을 발견

2 . Background : BERT

- **Masked Language Model(MLM)**

  - 그림 1처럼, 일단 단어 중의 일부를 [MASK] token 으로 바꾸어 줌.
  - 이를 통해 LM의 left-to-right (혹은 r2l)을 통하여 문장 전체를 predict하는 방법론과는 달리, [MASK] token 만을 predict하는 pre-training task를 수행.
  - 해당 token을 맞추어 내는 task를 수행하면서, BERT는 문맥을 파악하는 능력을 길러내게 됨.
  - [MASK] token으로 바꾸어준 비율은 15%
  - 80% [MASK] , 10% 전혀 다른 set의 토큰 , 10%는 원본 토큰 그대로



[그림 1] BERT Masked Language Model

- **Next Sentence Prediction(NSP)**

  - 2개의 시퀀스 X_A, X_B가 input으로 들어 올 때, X_B가 X_A의 직접적인 연속 인지 확인
  - 50% : sentence A, B가 실제 next sentence
  - 50% : sentence A, B가 corpus에서 random으로 뽑힌(관계가 없는) 두 문장
  - 예를 들어

```
Input = [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP] LABEL =
IsNext

Input = [CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]
Label = NotNext
```

3 . Model

3.1 Span Masking

- 일련의 토큰들 $X = (x_1, \ldots, x_n)$ 이 주어지면, X의 15%가 masking될 때까지, 텍스트의 span을 반복적으로 샘플링 하여 토큰 $Y \subseteq X$ 의 하위 집합을 선택.
  - token sequence의 15%가 masking되어있도록 만듦.

- span의 길이는 geometric distribution $\ell \sim Geo(p)$ 에 의해 선택, 그 후 span의 시작점을 임의로 (균일한 확률로) 선택
  - 저자는, p = 0.2 , ℓ_{max} = 10 으로 설정, ℓ의 평균은 3.8
  - span의 길이 측정은 완전한 단어 단위가 되게끔 함

- BERT 역시 선택된 15%의 masking 된 토큰 사용
  - 그 중 80%는 [mask] , 10% 원단어, 10%는 랜덤한 token
  - 하지만, 저자는 각각의 token 단위가 아닌 span 레벨로 대체

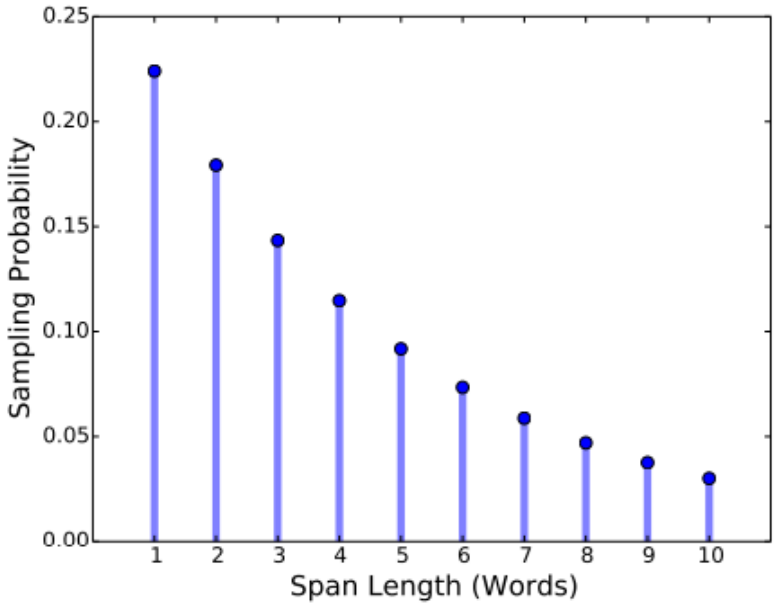


Figure 2: We sample random span lengths from a geometric distribution $\ell \sim Geo(p = 0.2)$ clipped at $\ell_{max} = 10$.

3 . Model

3.2 Span Boundary Objective (SBO)

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\mathbf{x}_7) + \mathcal{L}_{\text{SBO}}(\mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_7)$$

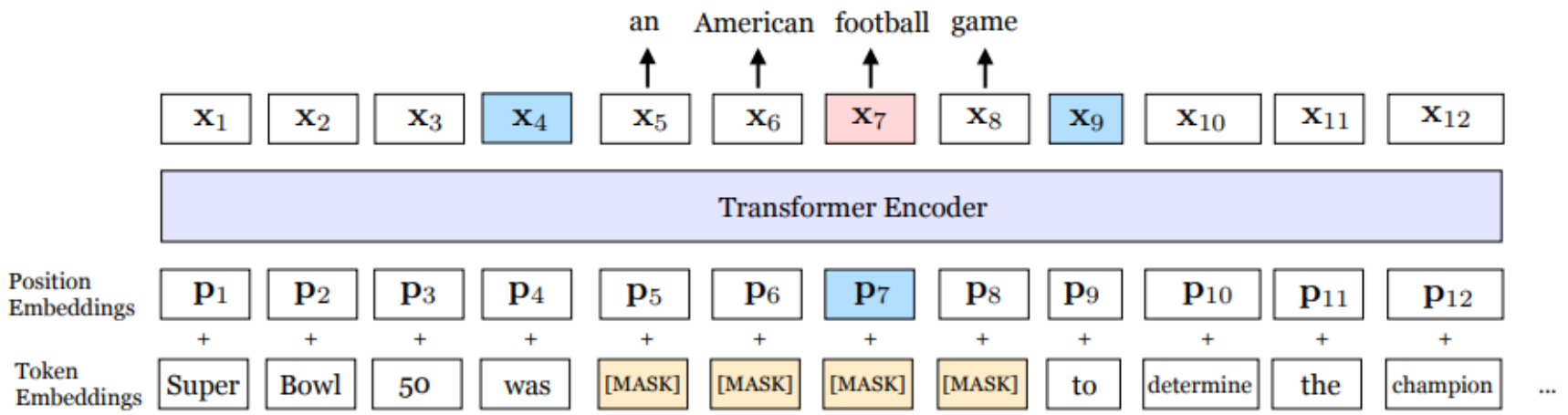


Figure 1: An illustration of SpanBERT training. In this example, the span *an American football game* is masked. The span boundary objective then uses the boundary tokens *was* and *to* to predict each token in the masked span.

- SBO : 경계에서 관찰 된 token의 표현만 사용하여, 마스크 된 span의 각 token을 예측.
  - Span selection models은 일반적으로 span의 경계 토큰을 사용하여 , fixed-length representation을 만듦.
  - Span의 끝에 최대한 많은 내부의 span 내용을 요약하는 것이 이상적.

3 . Model

3.2 Span Boundary Objective (SBO)

[구현한 방법]

- $y_i = f\left(X_{s-1}, X_{e+1}, P_i\right)$
  - P_i : 타겟 토큰의 positinoal embedding값
  - X_{s-1}, X_{e+1}: 경계 token의 인코딩 값

- $f(.)$를 구현하기 위해
- GeLU activations과 layer normalization 기능이 있는 2-layer feed-forward network 사용
  - $\mathbf{h} = LayerNorm(GeLU(W_1 \cdot [X_{s-1}; X_{e+1}; P_i]))$
  - $f(\cdot) = LayerNorm(GeLU(W_2 \cdot \mathbf{h}))$

- MLM 처럼 cross-entropy loss 사용하는데, 최종적으로 SBO loss와 MLM loss가 함께 합해져 학습하게 됨.

3 . Model

3.3 Single-Sequence Training

- BERT의 경우 sequence를 두 개 뽑아 (X_A, X_B) 모델이 NSP를 예측하지만, 이러한 셋팅이 성능을 더 안 좋은 결과를 보임을 관측

- 오히려, single sequence를 사용하고 NSP를 제외하는 것이 결과가 더 좋음 (4.3 자세히 설명)
  - 모델이 더 긴 context를 보는 것이 이득
  - NSP를 위해 다른 document에서 시퀀스를 뽑을 경우 오히려 노이즈 추가 될 수 있음.

- 저자는 최대 512개의 토큰으로 이루어진 단일 연속 세그먼트(a single contiguous segment)를 간단히 샘플링 함.

4 . Experimental Setup

4.1 Tasks

- Extractive Question Answering
  - 짧은 텍스트 구절과 질문이 주어지면, 구절과 관련도가 높은 문장을 답변으로 선택하는 것
  - 사용된 dataset 종류
    - SQuAD 1.1 and 2.0
    - Five more datasets from the MRQA
      - NewsQA (Trischler et al., 2017), SearchQA (Dunn et al., 2017), TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), NaturalQA (Kwiatkowski et al., 2019)
  - Data set은 도메인 및 수집 방법론에 따라 다르므로, 사전 훈련된 모델이 여러 데이터 분포에서 잘 일반화 될 수 있는 지 여부를 판단 하기 위해

- Coreference Resolution(상호 참조 해결)
  - 동일한 실제 엔티티를 나타내는 mention끼리 clusting 하는 것
  - 사용된 dataset 종류 : CoNLL-2012
    - document-level의 coreference resolution을 위한 data set
  - Lee et al.'s (2018)
    - 원래 LSTM 기반의 encoder를 BERT의 pre-trained transformer encoders로 대체하여 비교 실험 진행

4 . Experimental Setup

4.1 Tasks

- Relation Extraction
  - 사용된 dataset 종류 : TACRED
    - 관계 추출을 위한 data set
    - no_relation을 포함하여, 42개의 사전 정의된 관계 타입을 이용해서 span사이의 관계를 예측
  - Soares et al., 2019
    - pre-trains relation representations using distant supervision from entity-linked text

- GLUE
  - consists of 9 sentence-level classification tasks
    - 2 single-sentence tasks: CoLA , SST-2
    - 3 sentence similarity tasks: MRPC , STS-B , QQP3
    - 4 natural language inference tasks: MNLI , QNLI , RTE, WNLI

4 . Experimental Setup

4.2 Implementation

- BERT-large 모델 사용, 동일 corpus를 사용하여 모든 model을 훈련
  - 대소문자 구분 토큰을 사용하는 BooksCorpus와 English Wikipedia
- 주요 차이점
  - 저자
    - 각 epoch 마다 다른 mask를 사용
    - 문서 경계에 도달할 때까지 항상 최대 512개의 토큰 시퀀스를 가져옴
  - BERT
    - 데이터 처리 중에 각 sequence에 대해 10개의 다른 mask를 샘플링
    - 작은 확률(0.1)로 짧은 시퀀스를 샘플링
- Optimization은 기존 2.4M steps를 수행하는 것을 벗어나, Adam에 대해 1e-8의 epsilon 사용
- Pre-training은 32 Volta V100 GPU에서 15일 동안
- Fine-tuning은 HuggingFace의 코드 기반으로

4.3 Baselines

- Google BERT
- Our BERT
- Our BERT-1seq

5 . Results

- Extractive Question Answering

	SQuAD 1.1		SQuAD 2.0	
	EM	F1	EM	F1
Human Perf.	82.3	91.2	86.8	89.4
Google BERT	84.3	91.3	80.0	83.3
Our BERT	86.5	92.6	82.8	85.9
Our BERT-1seq	87.5	93.3	83.8	86.6
SpanBERT	**88.8**	**94.6**	**85.7**	**88.7**

2%

2.8%

Table 1: Test results on SQuAD 1.1 and SQuAD 2.0.

	NewsQA	TriviaQA	SearchQA	HotpotQA	NaturalQA	(Avg)
Google BERT	68.8	77.5	81.7	78.3	79.9	77.3
Our BERT	71.0	79.0	81.8	80.5	80.5	78.6
Our BERT-1seq	71.9	80.4	84.0	80.3	81.8	79.7
SpanBERT	**73.6**	**83.6**	**84.8**	**83.0**	**82.5**	**81.5**

Table 2: Performance (F1) on the five MRQA extractive question answering tasks.

5 . Results

- Coreference Resolution

	MUC			B^3			$CEAF_{\phi_4}$			
	P	R	F1	P	R	F1	P	R	F1	Avg. F1
Prev. SotA: (Lee et al., 2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
Google BERT	84.9	82.5	83.7	76.7	74.2	75.4	74.6	70.1	72.3	77.1
Our BERT	85.1	83.5	84.3	77.3	75.5	76.4	75.0	71.9	73.9	78.3
Our BERT-1seq	85.5	84.1	84.8	77.8	76.7	77.2	75.3	73.5	74.4	78.8
SpanBERT	**85.8**	**84.8**	**85.3**	**78.3**	**77.9**	**78.1**	**76.4**	**74.2**	**75.3**	**79.6**

Table 3: Performance on the OntoNotes coreference resolution benchmark. The main evaluation is the average F1 of three metrics – MUC, B^3, and $CEAF_{\phi_4}$ on the test set.

5．Results

- Relation Extraction

	P	R	F1
Curr. SotA: (Soares et al., 2019)	-	-	71.5
Google BERT	69.1	63.9	66.4
Our BERT	67.8	67.2	67.5
Our BERT-1seq	**72.4**	67.9	70.1
SpanBERT	70.8	**70.9**	**70.8**

Table 5: Test set performance on the TACRED relation extraction benchmark.

5．Results

- GLUE

	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	(Avg)
Google BERT	59.3	**95.2**	88.5/84.3	86.4/88.0	71.2/89.0	86.1/85.7	93.0	71.1	80.4
Our BERT	58.6	93.9	90.1/86.6	88.4/89.1	71.8/89.3	87.2/86.6	93.0	74.7	81.1
Our BERT-1seq	63.5	94.8	**91.2**/87.8	89.0/88.4	**72.1/89.5**	88.0/87.4	93.0	72.1	81.7
SpanBERT	**64.3**	94.8	90.9/**87.9**	**89.9/89.1**	71.9/**89.5**	**88.1/87.7**	**94.3**	**79.0**	**82.8**

Table 4: Test set performance metrics on GLUE tasks. MRPC: F1/accuracy, STS-B: Pearson/Spearmanr correlation, QQP: F1/accuracy, MNLI: matched/mistached accuracies. WNLI (not shown) is always set to majority class (65.1% accuracy) and included in the average.

5 . Results

5.2 Overall Trends

- 17개의 benchmarks에서 3개의 BERT baselines 중에서도 SpanBERT가 BERT보다 거의 모든 작업에서 우수
- SpanBERT는 특히 extractive question answering에 특히 더 나은 성능을 보임
- single-sequence 훈련이 next sentence prediction(NSP)를 사용한 bi-sequence 훈련보다 훨씬 효과적

6 . Ablation Studies

6.1 Masking Schemes

- 논문에서 사용된 Random Spans 방식과 언어학적으로 알려진 masking Schemes를 비교 실험 진행

	SQuAD 2.0	NewsQA	TriviaQA	Coreference	MNLI-m	QNLI
Subword Tokens	83.8	72.0	76.3	**77.7**	86.7	92.5
Whole Words	84.3	72.8	77.1	76.6	86.3	92.8
Named Entities	84.8	72.7	78.7	75.6	86.0	93.1
Noun Phrases	85.0	**73.0**	77.7	76.7	86.5	93.2
Random Spans	**85.4**	**73.0**	**78.8**	76.4	**87.0**	**93.3**

Table 6: The effect of replacing BERT's original masking scheme (Subword Tokens) with different masking schemes. Results are F1 scores for QA tasks and accuracy for MNLI and QNLI on the development sets. All the models are based on bi-sequence training with NSP.

6 . Ablation Studies

6.2 Auxiliary Objectives

	SQuAD 2.0	NewsQA	TriviaQA	Coreference	MNLI-m	QNLI
Span Masking (2seq) + NSP	85.4	73.0	78.8	76.4	87.0	93.3
Span Masking (1seq)	86.7	73.4	80.0	76.3	87.3	93.8
Span Masking (1seq) + SBO	**86.8**	**74.1**	**80.3**	**79.0**	**87.6**	**93.9**

Table 7: The effects of different auxiliary objectives, given MLM over random spans as the primary objective.