SpanBERT: Improving Pre-training by Representing and Predicting Spans

이봉석

- **Author**
  - Mandar Joshi
  - Danqi Chen
  - Yinhan Liu
  - Daniel S. Weld
  - Luke Zettlemoyer
  - Omer Levy

- **Title of Conference(Journal)**
  - TACL 2020

01. Introduction

많은 NLP task는 두개 이상의 span 사이의 관계에 대한 추론을 포함한다.

예를 들어 question answering에서 “Denver Broncos”가 “NFL 팀”인지 결정하는 것은 “어떤 NFL팀이 Super Bowl 50에서 우승했습니까?”라는 질문에 대답하는데 중요하다.

=> 그래서 text의 span을 보다 잘 예측 표현하는 pre-training method를 디자인하자

01. Introduction

SpanBERT는 BERT로 부터 영감을 받은 모델인데 3가지 차이점이 있다.

1. **Span Masking**
2. **Span Boundary Objective**
3. **Single-Sequence Training**

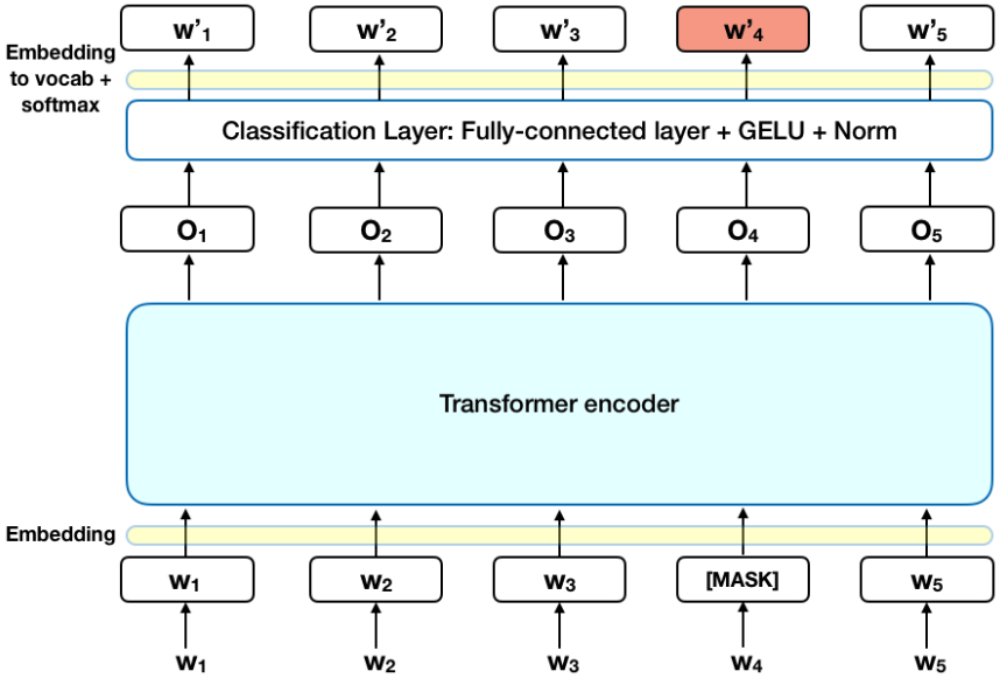
02. Background: BERT

- **BERT는 특정 downstream task에 맞게 fine-tuning하기전에 deep transformer encoder를 pre-training하는 self-supervised 방식이다.**

- **BERT는 MLM(Masked Language Modeling)과 NSP(Next Sentence Prediction)라는 두 가지 training objective를 optimization한다.**

02. Background: BERT

- **MLM(Masked Language Modeling)**



- **MLM은 sequence에서 mask된 token을 예측하는 task.**
- **BERT 구현에서 mask는 token중 15%를 차지한다. 그중 80%는 [MASK], 10%는 Random Token, 10%는 변경하지 않는다.**
- **BERT에서 mask되는 tokens은 각 individial token을 무작위로 선정하지만 SpanBERT는 contiguous span을 무작위로 선택한다.**

02. Background: BERT

- **NSP(Next Sentence Prediction)**

```
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
```

- **NSP task는 XA, XB 두 sequence를 입력으로 사용하고 XB가 XA와 직접적인 연속인지 여부를 예측.**
- **두 sequence 는 [SEP] token으로 구분.**
- **[CLS]token이 XA, XB에 추가되어 모델 입력을 형성.**
- **여기서 [CLS]를 통해 XB가 XA 뒤에 오는지 안오는지를 예측.(binary classification)**
- **SpanBERT에서는 NSP objective를 제거하고 single full-length sequence를 사용.**

03. Model

SpanBERT는 BERT로 부터 영감을 받은 모델인데 3가지 차이점이 있다.

1. **Span Masking(individual token이 아닌 다른 방식으로 token span을 마스킹)**
2. **Span Boundary Objective(span boundary에서 token representation만 사용하여 전체 mask된 span을 예측하고자 하는 새로운 auxiliary objective인 span boundary objective)**
3. **Single-Sequence Training(하나의 contiguous한 text segment를 샘플링)**

03. Model(Single-Sequence Training)

- **BERT는 두 개의 text sequence (XA, XB)와 연결여부(NSP)를 예측하기 위해 모델을 학습시키는 objective가 포함되어 있다.**
- **NSP objective없이 single sequence를 사용하는 것이 더 좋은 성능을 보인다는 것을 발견**
- **더 긴 길이의 context를 보는 것이 이득이고, NSP를 위해 다른 documnet에서 시퀀스를 뽑을 경우 오히려 노이즈가 추가될 수 있다고 추측했다.**

03. Model(Span Masking)

- token으로 구성된 sequence X가 주어지면 masking budget(e.g., 15% of X)이 사용될 때까지 text span을 반복적으로 샘플링한다.
- 각 반복마다 geometric distribution를 통해 span 길이를 샘플링한다.
- 이 분포는 더 짧은 스팬으로 편향된다. (실험을 통해 확률이 0.2일때가 가장 좋고 이 때 길이의 평균이 3.8이다.)
- 이후 span의 시작점을 random하게(균일하게) 선택한다.
- 또한 subword token이 아닌 complete word로 span 길이를 측정한다.
- BERT와 마찬가지로 15%를 마스킹 - 80% [MASK], 10% Random Token, 10% Original Token으로 한다.

03. Model(Span Masking)

확률변수 X가 기하확률변수 일 때, 확률질량함수 $f(x)$는

$$f(x)=pq^{x-1}$$

단, $x=1, 2, \ldots$ 이고, $p+q=1$ 이다.

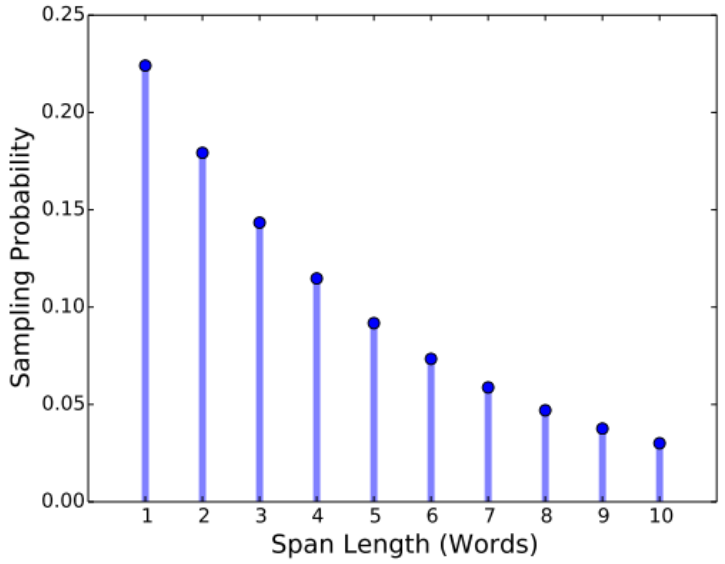


Figure 2: We sample random span lengths from a geometric distribution $\ell \sim Geo(p = 0.2)$ clipped at $\ell_{max} = 10$.

03. Model(Span Boundary Objective)

- boundary에서 관찰된 token의 representation만 사용하여 mask된 span의 각 token을 예측하는 SBO를 도입하여 이를 수행한다.
- Masked span (x_s, … , x_e) ∈ Y, 여기서 (s, e)는 시작 및 끝 위치를 나타낸다.
- external boundary token x_{s-1}과 x_{e+1}의 encoding target token P_i의 position embedding을 사용하여 span의 각 token x_i를 나타낸다.

$$\mathbf{y}_i = f(\mathbf{x}_{s-1}, \mathbf{x}_{e+1}, \mathbf{p}_{i-s+1})$$

03. Model(Span Boundary Objective)

- 본 논문에서는 GeLU activation function 및 Layer Normalization을 사용하여 representation function f를 2-layer Feed-Forward Network로 구현한다.

$$\begin{aligned} \mathbf{h}_0 &= [\mathbf{x}_{s-1}; \mathbf{x}_{e+1}; \mathbf{p}_{i-s+1}] \\ \mathbf{h}_1 &= \text{LayerNorm}(\text{GeLU}(\mathbf{W}_1 \mathbf{h}_0)) \\ \mathbf{y}_i &= \text{LayerNorm}(\text{GeLU}(\mathbf{W}_2 \mathbf{h}_1)) \end{aligned}$$

- 이후 vector representation y_i를 사용하여 x_i를 예측하고 정확히 MLM objective와 같은 cross-entropy loss를 계산한다.
- SpanBERT는 span boundary 및 mask된 span의 각 token에 대한 regular masked language modeling objective의 loss를 합산한다.

$$\begin{aligned} \mathcal{L}(x_i) &= \mathcal{L}_{\text{MLM}}(x_i) + \mathcal{L}_{\text{SBO}}(x_i) \\ &= -\log P(x_i \mid \mathbf{x}_i) - \log P(x_i \mid \mathbf{y}_i) \end{aligned}$$

03. Model(Span Boundary Objective)

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$

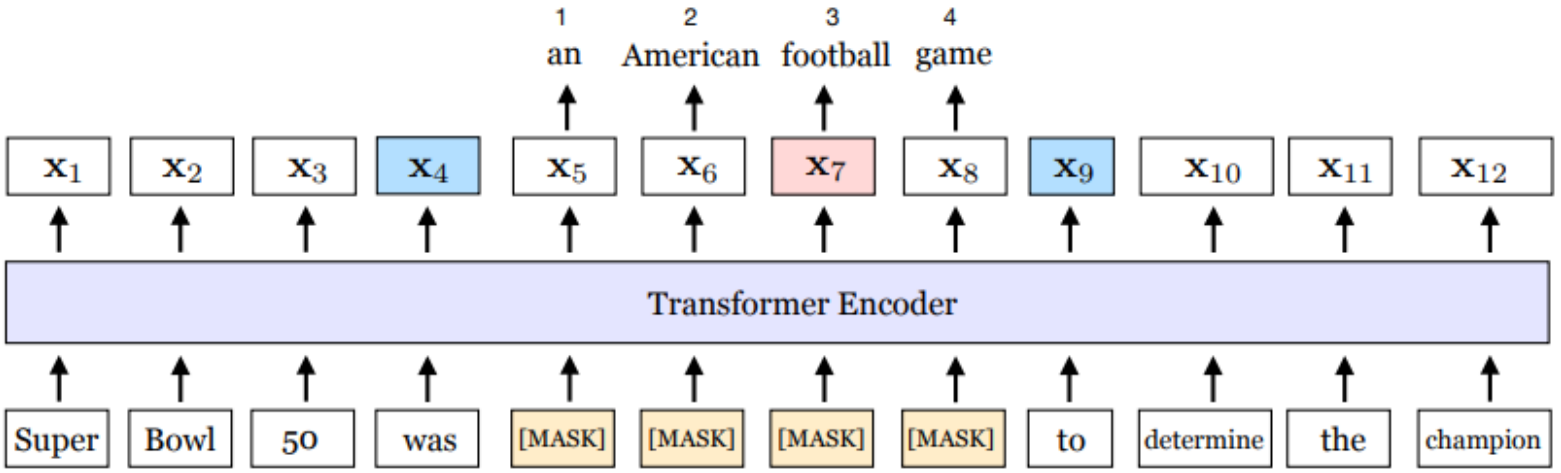


Figure 1: An illustration of SpanBERT training. The span *an American football game* is masked. The span boundary objective (SBO) uses the output representations of the boundary tokens, $\mathbf{x}_4$ and $\mathbf{x}_9$ (in blue), to predict each token in the masked span. The equation shows the MLM and SBO loss terms for predicting the token, *football* (in pink), which as marked by the position embedding $\mathbf{p}_3$, is the *third* token from x_4.

04. Experiment

- **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**

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

	NewsQA	TriviaQA	SearchQA	HotpotQA	Natural Questions	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.

04. Experiment

➢ 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.

04. Experiment

- **Relation Extraction**

	p	R	F1
$BERT_{EM}$(Soares et al., 2019)	–	–	70.1
$BERT_{EM}$+MTB*	–	–	**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 4: Test performance on the TACRED relation extraction benchmark. $BERT_{large}$ and $BERT_{EM}$+MTB from Soares et al. (2019) are the current state-of-the-art. *: $BERT_{EM}$+MTB incorporated an intermediate ''matching the blanks'' pre-training on the entity-linked text based on English Wikipedia, which is not a direct comparison to ours trained only from raw text.

04. Experiment

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

감사합니다

Appendix

➢ GLUE

Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**	Accuracy / F1
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**	Pearson / Spearman
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**	Accuracy
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**	Accuracy