KOREA UNIVERSITY

DSBA CS224n 2021 Study

# [Lecture 17]
# Model Analysis and Explanation

DSBA
Data Science & Business Analytics

고려대학교 산업경영공학과

Data Science & Business Analytics Lab

발표자 : 이성계

# 00 발표 목차

Contents

1. Motivating model analysis and explanation
2. Out-of-domain evaluation sets
3. Influence studies and adversarial examples
4. Analyzing representations
5. Revisiting model ablations as analysis
6. Conclusion

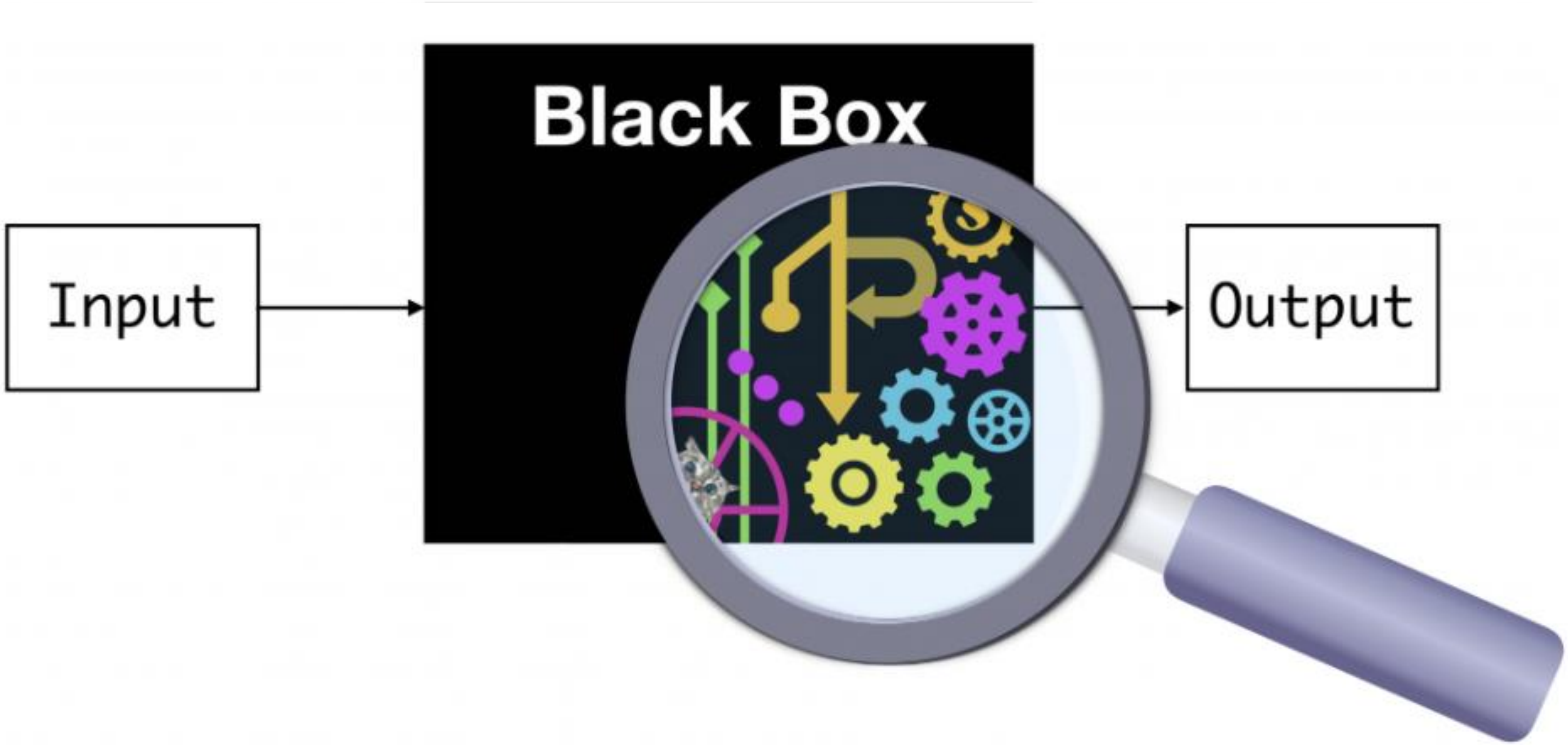
# 01 Contents

1. Motivating model analysis and explanation
2. Out-of-domain evaluation sets
3. Influence studies and adversarial examples
4. Analyzing representations
5. Revisiting model ablations as analysis
6. Conclusion

# 01 Motivating model analysis and explanation

Model은 어떻게 무엇을 학습하는가?

- Model의 예측 정확도는 **숫자**로 확인이 가능하지만, Model이 어떻게 동작하는지는 알기 어려움



01

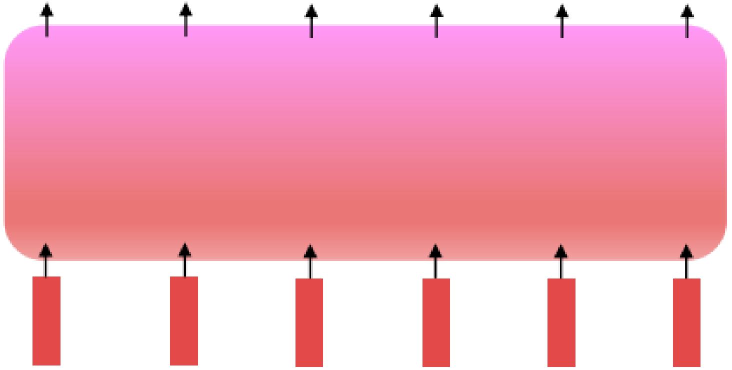
# Motivating model analysis and explanation

Tomorrow's Model

What can be learned via language model pretraining?

What will replace the Transformer?

What **can't** be learned via language model pretraining?



What does deep learning struggle to do?

How are our models affecting people, and transferring power?

What do neural models tell us about language?

➡ 적절한 Model 분석을 통해 Robust하고 예측 정확도가 높은 Model 생성

01

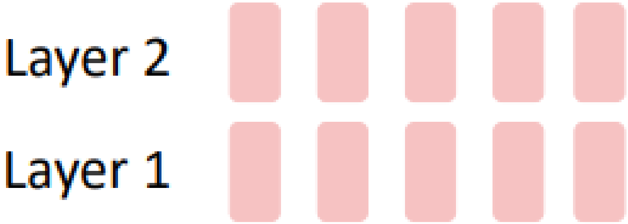
# Motivating model analysis and explanation

Model 분석의 level of abstraction

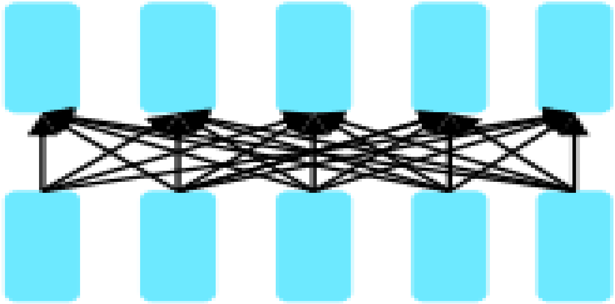
1. As a probability distribution and decision function

$$p_{\text{model}}(y|x)$$

2. As a sequence of vector representations in depth and time



3. Parameter weights, specific mechanisms like attention, dropout, +++

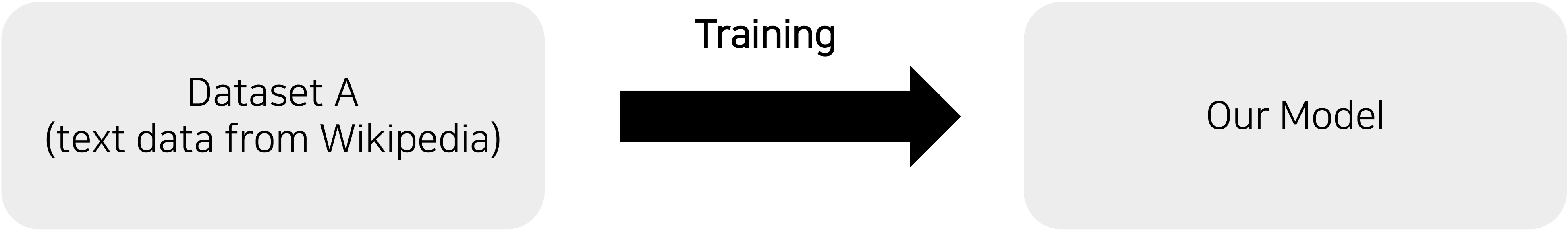


# 02 Contents

1. Motivating model analysis and explanation
2. **Out-of-domain evaluation sets**
3. Influence studies and adversarial examples
4. Analyzing representations
5. Revisiting model ablations as analysis
6. Conclusion

# 02 Out-of-domain evaluation sets

Out-of-domain 이란?



Dataset A
(text data from Wikipedia)

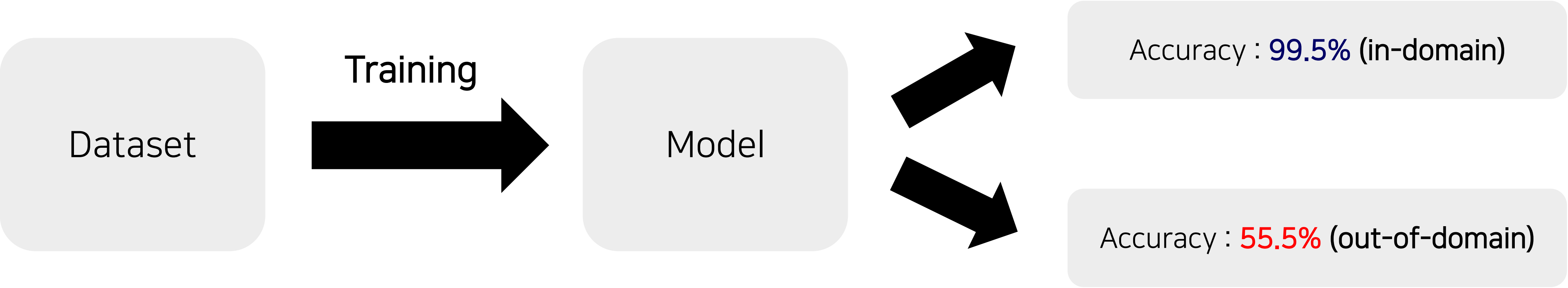
Training

Our Model

- In-Domain data : Text data from Wikipedia
- Out-of-Domain data : Except Wikipedia Data → Training 에 사용하지 않은 data

# 02 Out-of-domain evaluation sets

Fitting the dataset vs learning the task

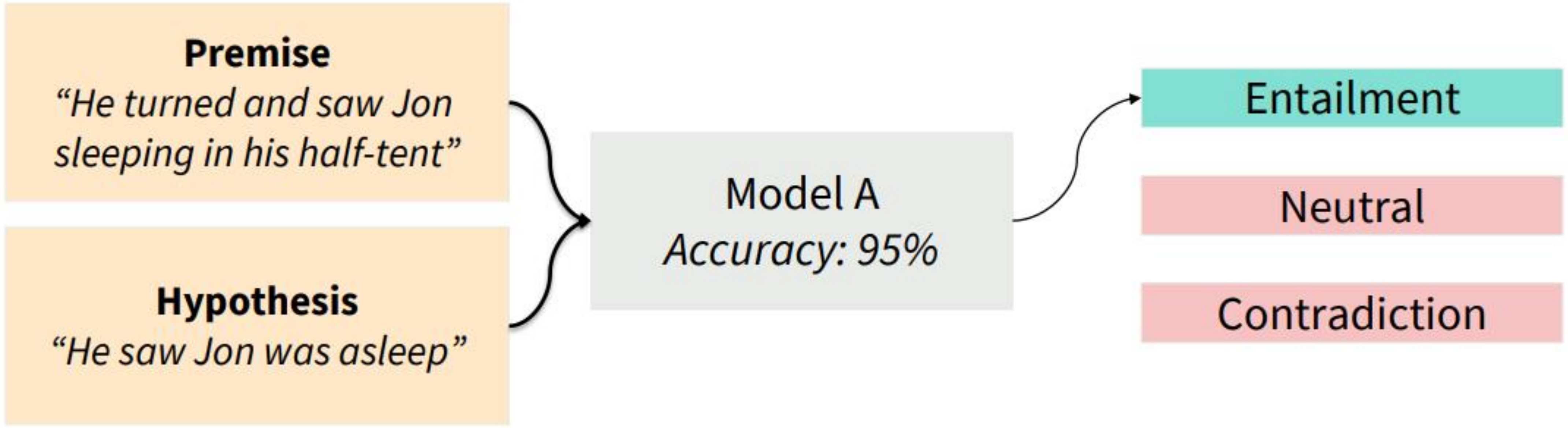

Dataset
Training
Model
Accuracy : 99.5% (in-domain)
Accuracy : 55.5% (out-of-domain)

- In-Domain data 에 대해서만 높은 accuracy 를 기록함
- 결국 Model 이 'task' 를 학습하는 것이 아닌 data를 단순히 fitting함을 예상 할 수 있음
- Out-of-domain data를 통해 Model을 분석할 수 있음을 시사

# 02 Out-of-domain evaluation sets

NLI(Natural Language Inference) 를 이용한 Model 평가

- NLI : 문장(Premise)이 주어졌을 때 다른 문장(Hypothesis)에 의미적으로 수반(Entail) 되는가?



**Premise**
*"He turned and saw Jon sleeping in his half-tent"*

**Hypothesis**
*"He saw Jon was asleep"*

Model A
*Accuracy: 95%*

Entailment

Neutral

Contradiction

※ Entailment : 같은 의미, Neutral : 무관함, Contradiction : 모순

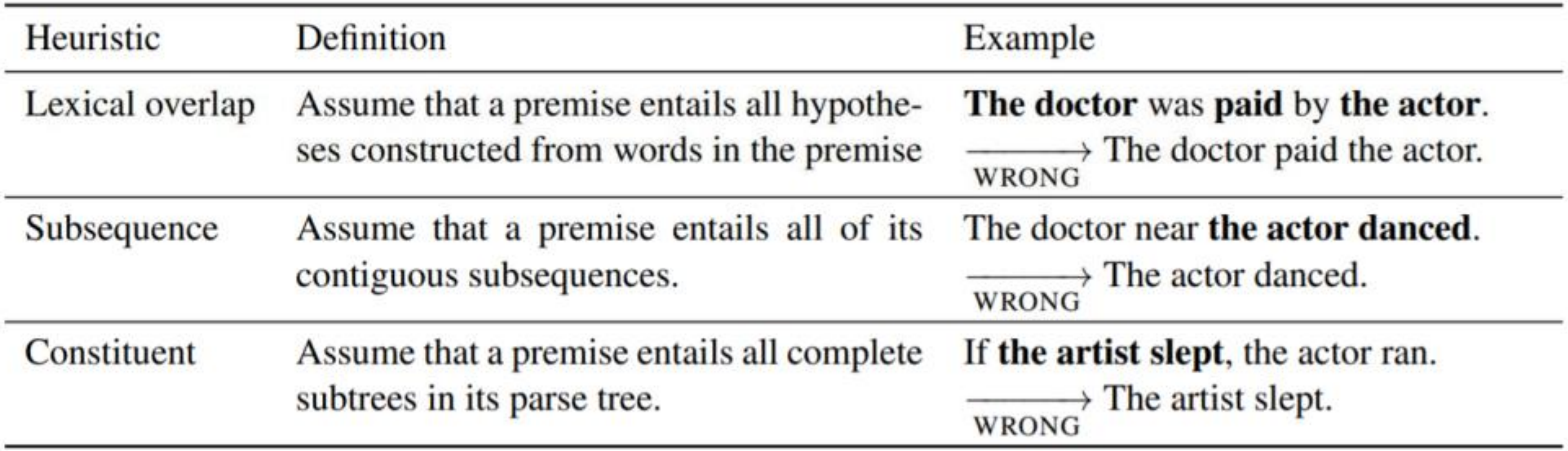
02

# Out-of-domain evaluation sets

HANS(Heuristic Analysis for NLI Systems) → diagnostic test set [McCoy et al., 2019]

- Model 이 Heuristic 한 방법을 통해 좋은 성능을 내는 것인지 확인하기 위한 용도의 Dataset

- Lexical overlap, Subsequence, Constituent 와 같은 syntactic heuristics 를 확인



| Heuristic | Definition | Example |
|---|---|---|
| Lexical overlap | Assume that a premise entails all hypotheses constructed from words in the premise | **The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor. |
| Subsequence | Assume that a premise entails all of its contiguous subsequences. | The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced. |
| Constituent | Assume that a premise entails all complete subtrees in its parse tree. | If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept. |

※ Constituent ⊂ Subsequence ⊂ Lexical overlap

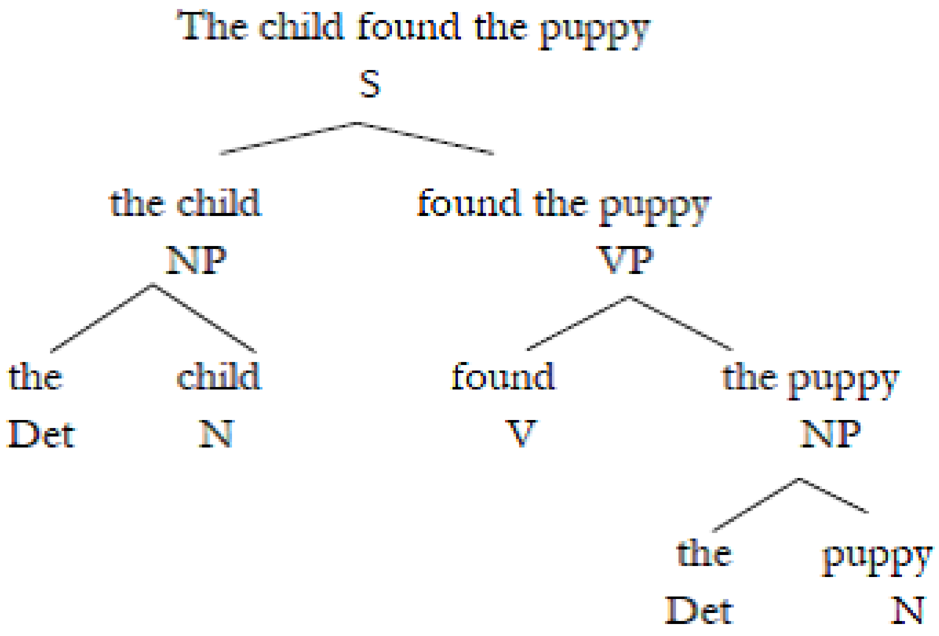
→ 본 논문에서는 하위 Heuristic 방식을 채택 (ex. Constituent 이면 Subsequence로는 사용 X)

02

# Out-of-domain evaluation sets

Constituent 란?

- **Constituent(구성 요소)**
- word or a group of words that **function as a single unit(subtree)** within a **hierarchical structure.**

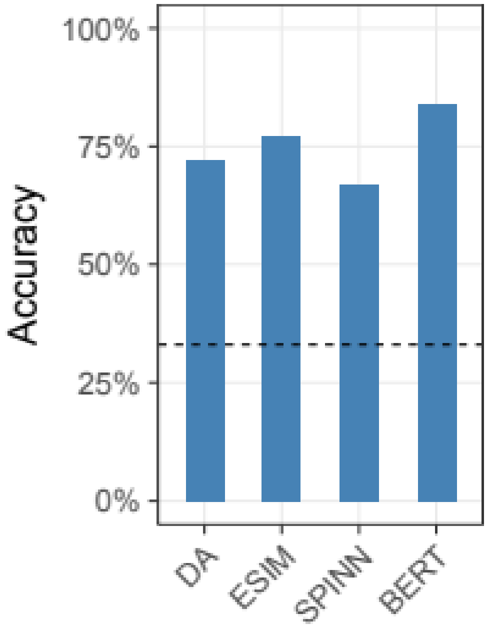


- The child, found the puppy → O
- The child found, found the .. → X

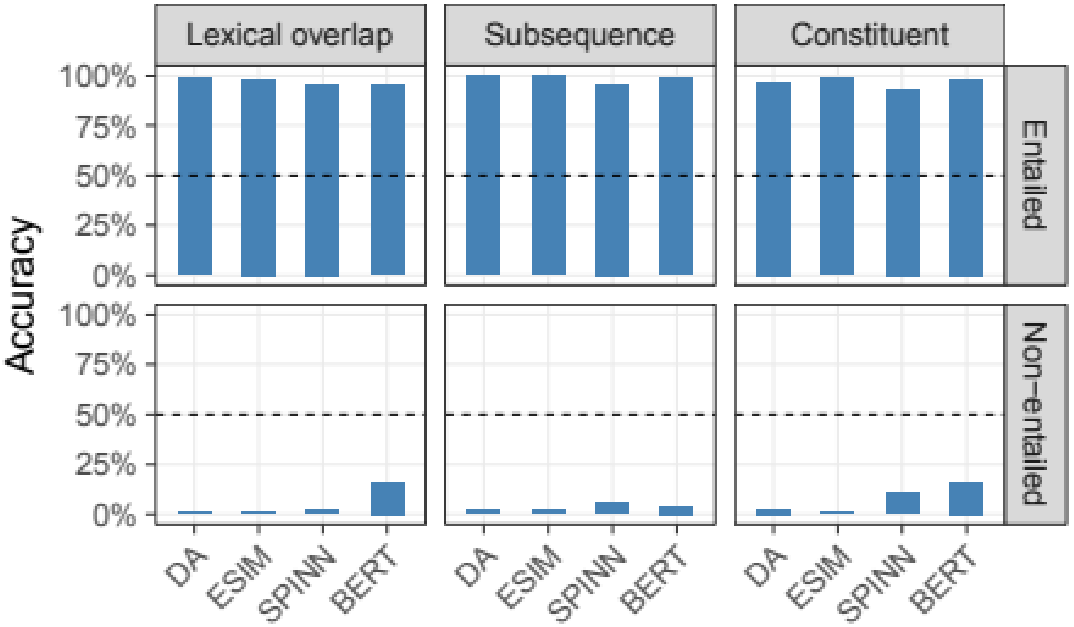
02

# Out-of-domain evaluation sets

HANS model analysis in natural language inference [McCoy et al., 2019]



a. Accuracy on the MNLI(In-domain)



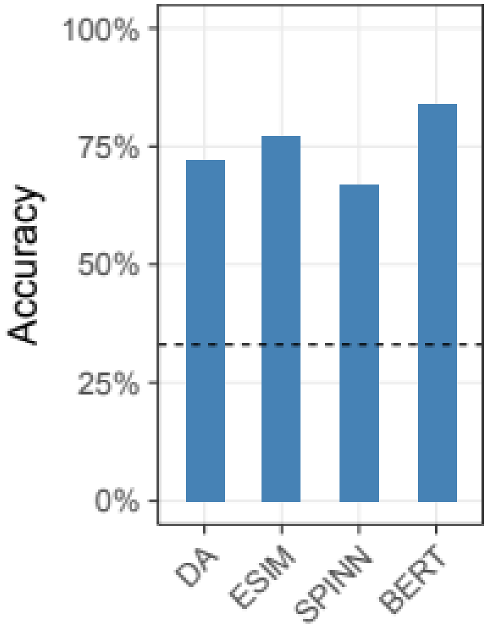
b. Accuracy on the HANS evaluation set

- MNLI (In-Domain) data에 대해서는 우수한 성능을 보임
- HANS dataset의 경우 syntactic heuristics 이 fail 하는 경우 accuracy가 매우 낮음

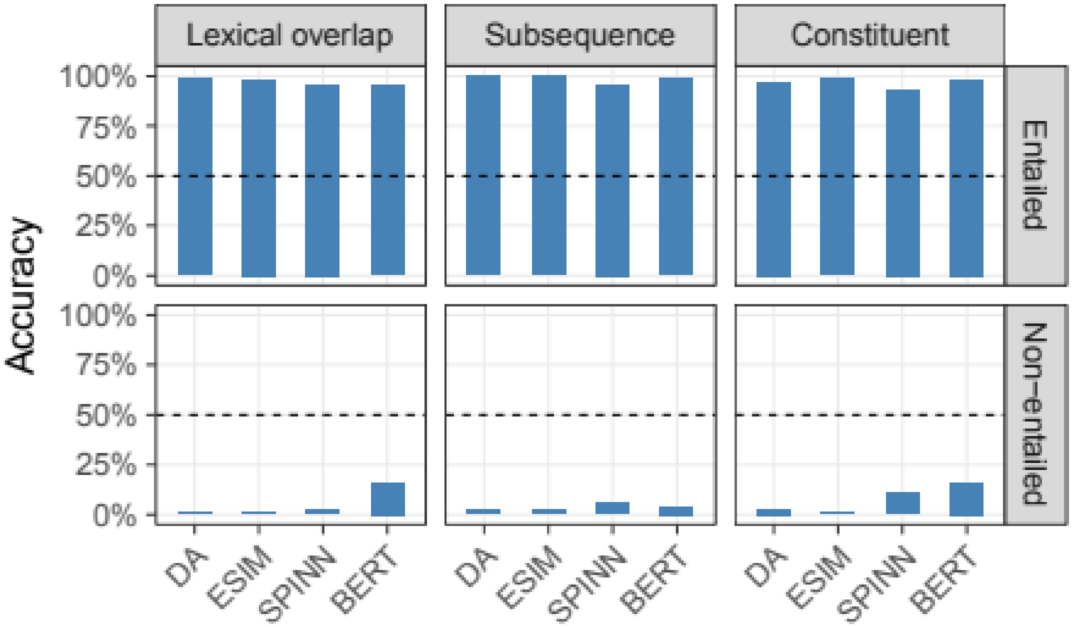
02

# Out-of-domain evaluation sets

Fitting the dataset vs Learning the task



a. Accuracy on the MNLI(In-domain)



b. Accuracy on the HANS evaluation set

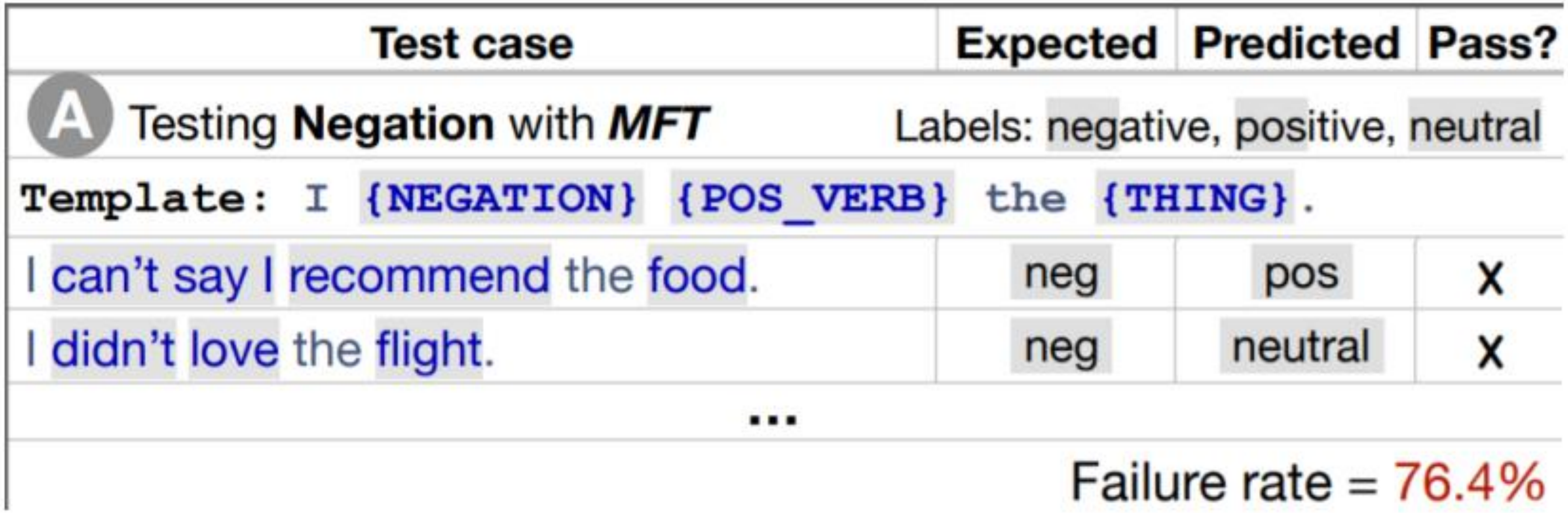
- 결국, Model이 NLI task를 학습하는 것이 아닌 Heuristic한 task를 사용한다고 추측할 수 있음

→ reasonable out-of-domain test set을 통해 Model을 test 하는 것은 유의미한 task가 될 수 있다.

02

# Out-of-domain evaluation sets

Careful test sets as unit test suites: CheckListing [Ribeiro et al., 2020]



| Test case | Expected | Predicted | Pass? |
|---|---|---|---|
| A Testing **Negation** with ***MFT*** | Labels: negative, positive, neutral | | |
| **Template:** I {NEGATION} {POS_VERB} the {THING}. | | | |
| I can't say I recommend the food. | neg | pos | X |
| I didn't love the flight. | neg | neutral | X |
| ... | | | |
| | | Failure rate = 76.4% | |

- 위와 같은 template을 이용하여 감성 분석 Model이 잘 작동하지 않는 test set을 만들 수 있다.

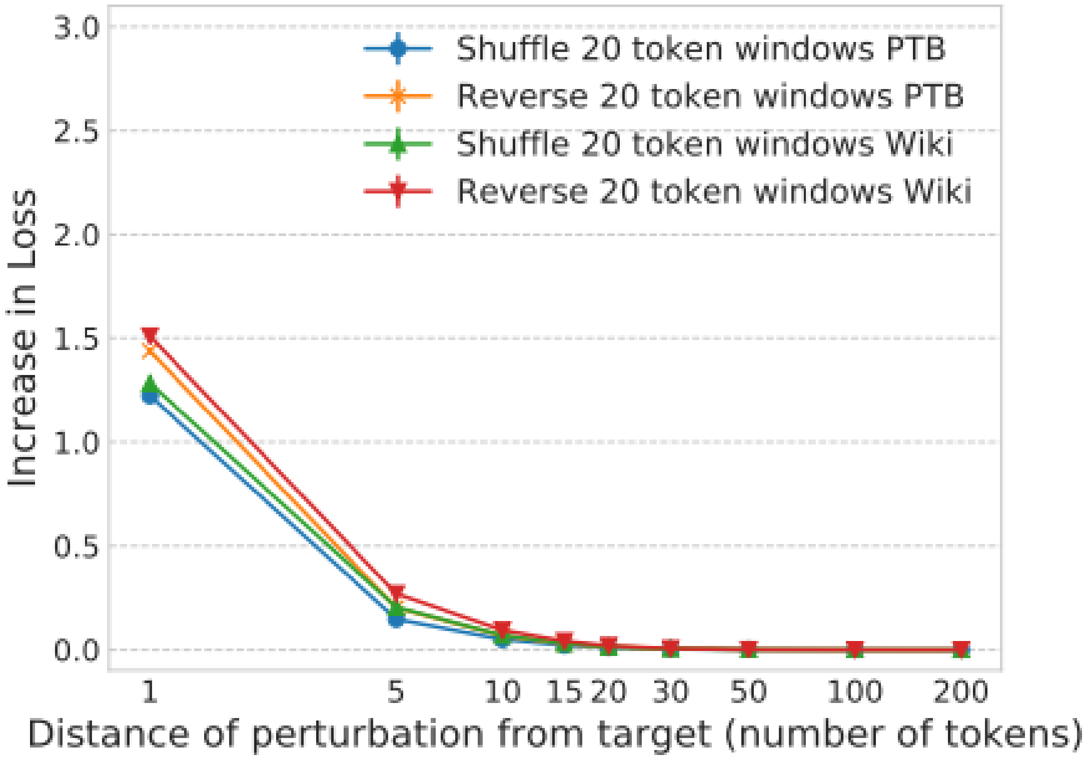
# 03 Contents

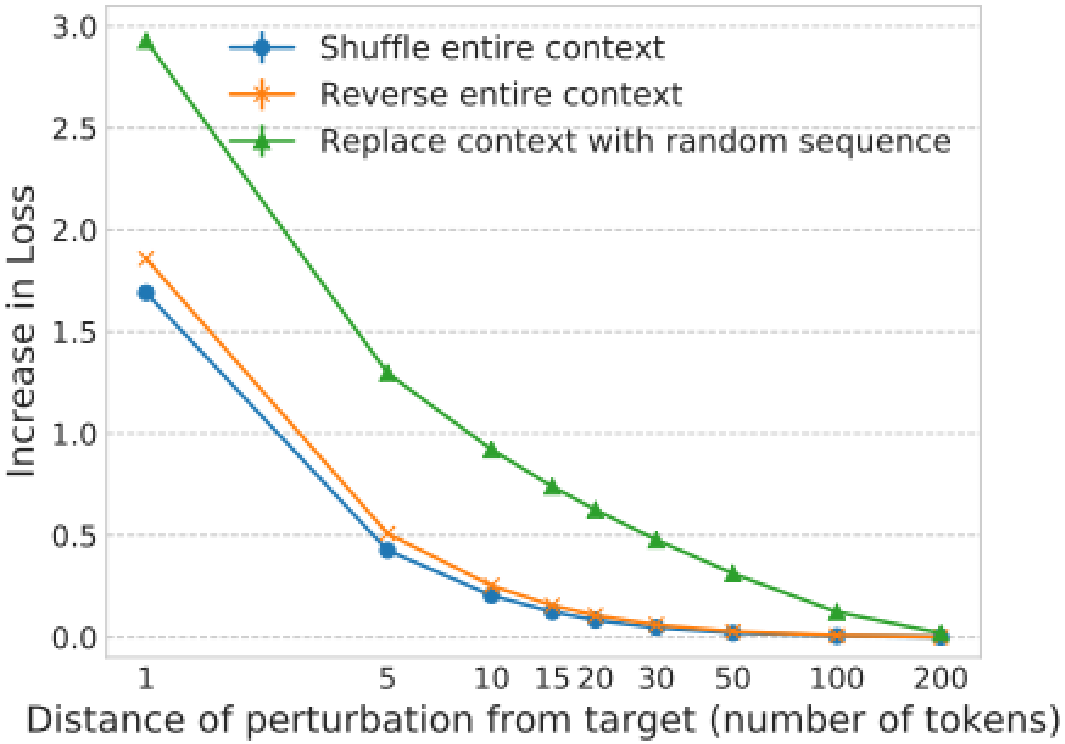
1. Motivating model analysis and explanation
2. Out-of-domain evaluation sets
3. **Influence studies and adversarial examples**
4. Analyzing representations
5. Revisiting model ablations as analysis
6. Conclusion

03

# Influence studies and adversarial examples

Model은 long-distance context를 정말 활용할까? [Khandelwal et al., 2018]





03

# Influence studies and adversarial examples

Model은 long-distance context를 정말 활용할까? [Khandelwal et al., 2018]

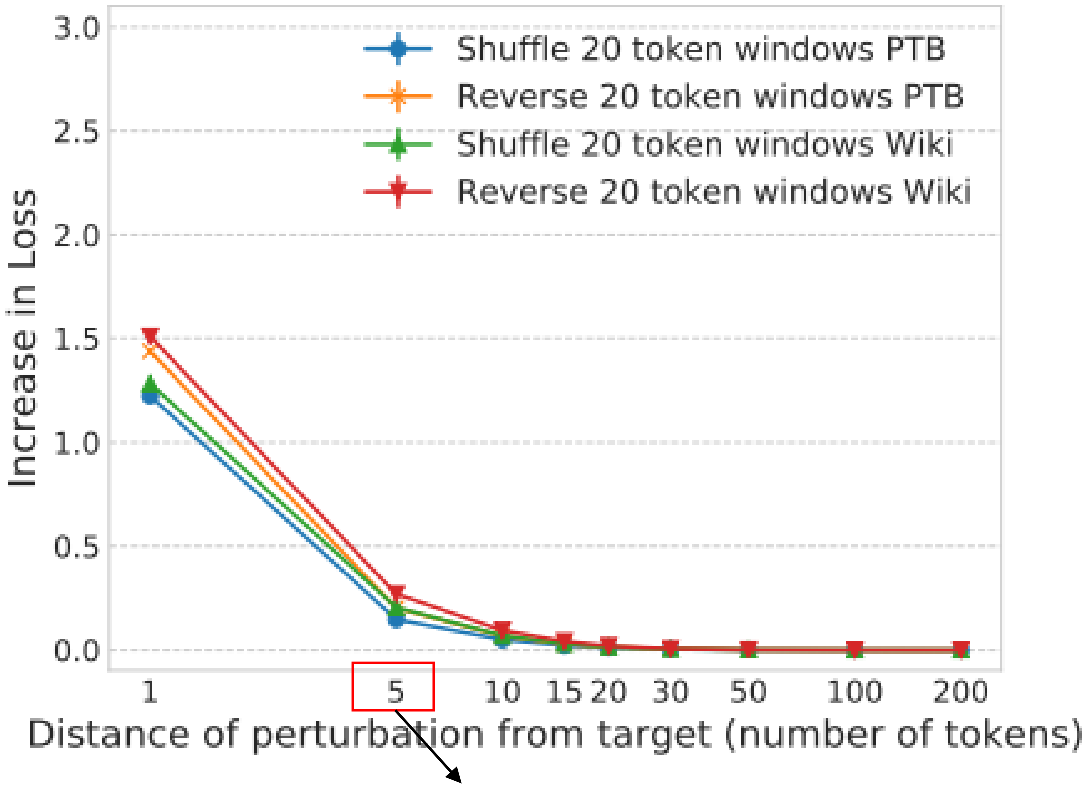
- LSTM model을 사용하여 실험
- Token의 permutation 전/후의 loss를 비교
- $\boldsymbol{\delta_{permute}(w_{t-1}, \cdots, w_{t-n}) = (w_{t-1}, \ldots \rho(w_{t-s_1-1}, \cdots, w_{t-s_2}), \ldots w_{t-n})}$
- $\boldsymbol{\rho} \in \{\boldsymbol{shuffle}, \boldsymbol{reverse}\},\ (\boldsymbol{s_1}, \boldsymbol{s_2}]: \boldsymbol{permutation}$ 범위 → permutable span
- Local word order (20 tokens), Global word order 에 대해 실험 진행

03

# Influence studies and adversarial examples

Model은 long-distance context를 정말 활용할까? [Khandelwal et al., 2018]

- Local word order (20 tokens)



- Target distance가 20을 넘어가면 Loss 차이 X
- Dataset, permutation 방식에 상관 없이 비슷한 Loss 변화 추이를 보임
- 20개의 token을 permutation 해서 20 이후로 Loss 차이가 없는 것 일까?

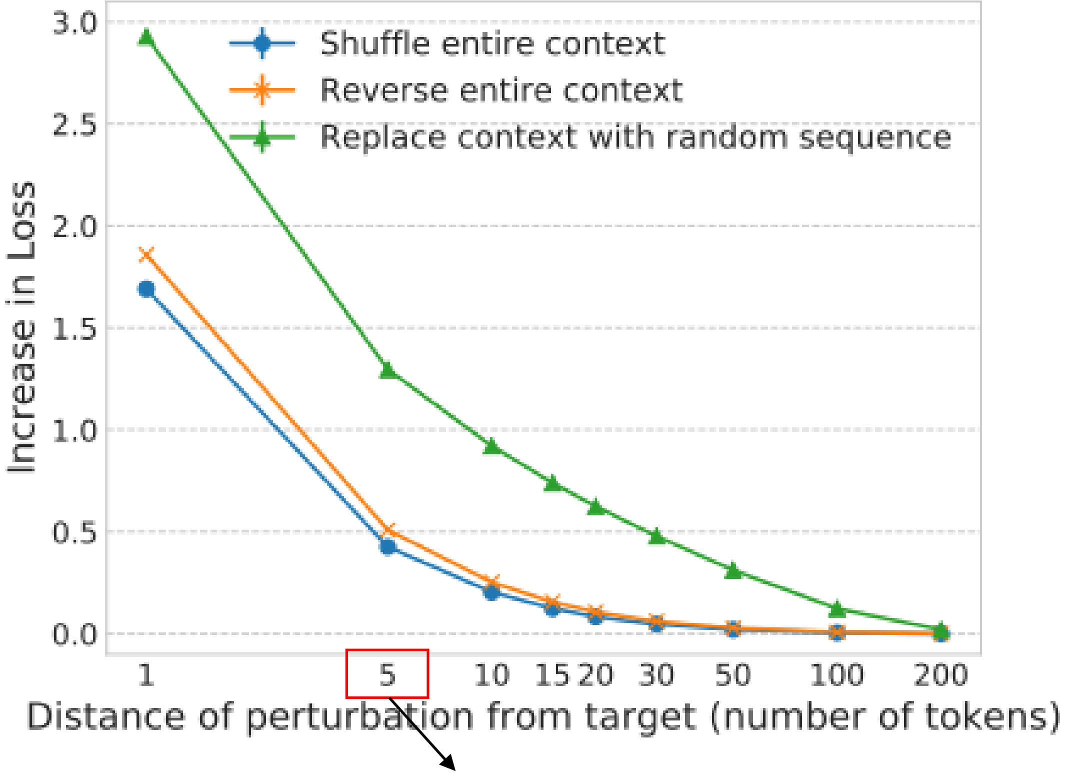
$s_1 = 5, s_2$ = 25, 20 token의 permutation만 진행하므로, $s_2 = s_1 + 20$

$\delta_{permute}(w_{t-1}, \cdots, w_{t-n}) = (w_{t-1}, \ldots \rho(w_{t-s_1-1}, \cdots, w_{t-s_2}), \ldots w_{t-n})$

# 03 Influence studies and adversarial examples

Model은 long-distance context를 정말 활용할까? [Khandelwal et al., 2018]

- Global word order



- Word order or Long-range context ?

  → Replace context with random sequence

- Shuffle, reverse 의 경우 50개 이후로 loss 변화 X

- 50개 이후 단어의 'identity' 는 중요

  → Replace case 의 경우 Loss가 증가

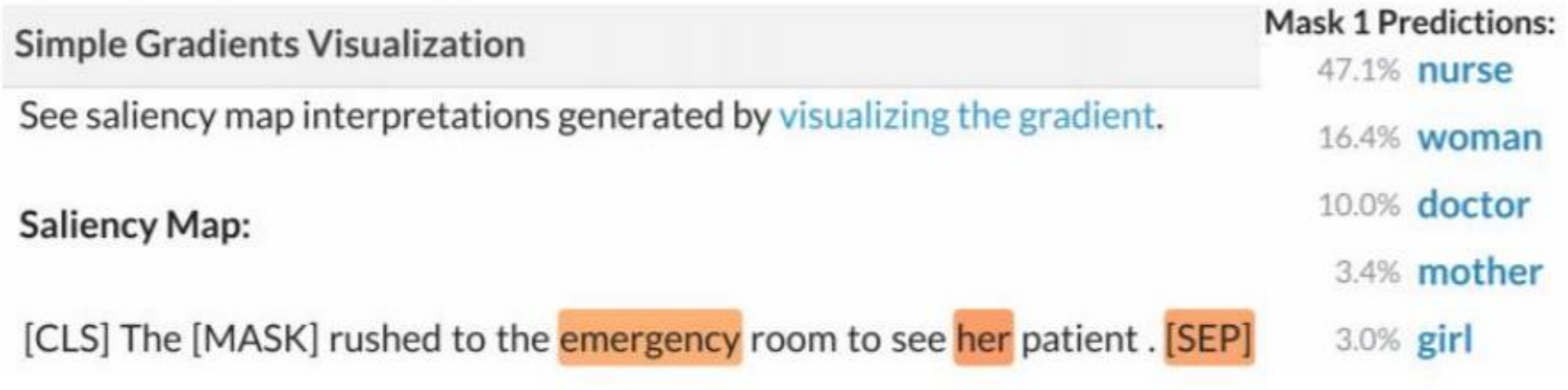
$s_1 = 5, s_2 = \mathbf{n}$, context 전체를 permutation 하므로, $s_2 = n$

$$\delta_{permute}(w_{t-1}, \cdots, w_{t-n}) = (w_{t-1}, \ldots\ \rho(w_{t-s_1-1}, \cdots, w_{t-s_2}), \ldots\ w_{t-n})$$

# 03 Influence studies and adversarial examples

Prediction explanations: input과 output의 연관관계 (saliency map)

- Saliency maps : model 의 prediction 에 input이 얼마나 영향을 미쳤는지 scoring



Simple Gradients Visualization

See saliency map interpretations generated by visualizing the gradient.

Saliency Map:

[CLS] The [MASK] rushed to the emergency room to see her patient . [SEP]

Mask 1 Predictions:

| | |
|---|---|
| 47.1% | nurse |
| 16.4% | woman |
| 10.0% | doctor |
| 3.4% | mother |
| 3.0% | girl |

- BERT 의 MLM task 진행한 예시 (오른쪽 비율이 Masking word 추론 비율)

03

# Influence studies and adversarial examples

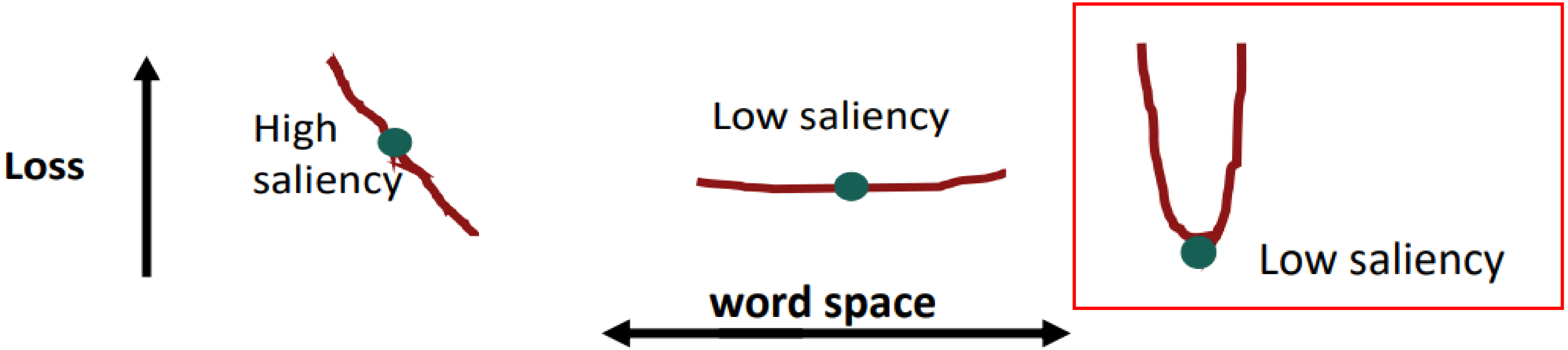
Prediction explanations: input과 output의 연관관계 (saliency map)

- Idea of saliency map → Simple gradient method

Input words $x_1, x_2, .., x_n$ 이고, model의 scoring 함수가 $S_c(x_1, x_2, \dots, x_n)$ 일 때 salience 는 다음과 같다.

$$salience(x_i) = \left\| \nabla_{x_i} S_c(x_1, x_2, \dots, x_n) \right\|$$

결국, Gradient가 크다는 것은 score에 영향을 많이 준다는 것을 의미하므로 위와 같이 정의 가능



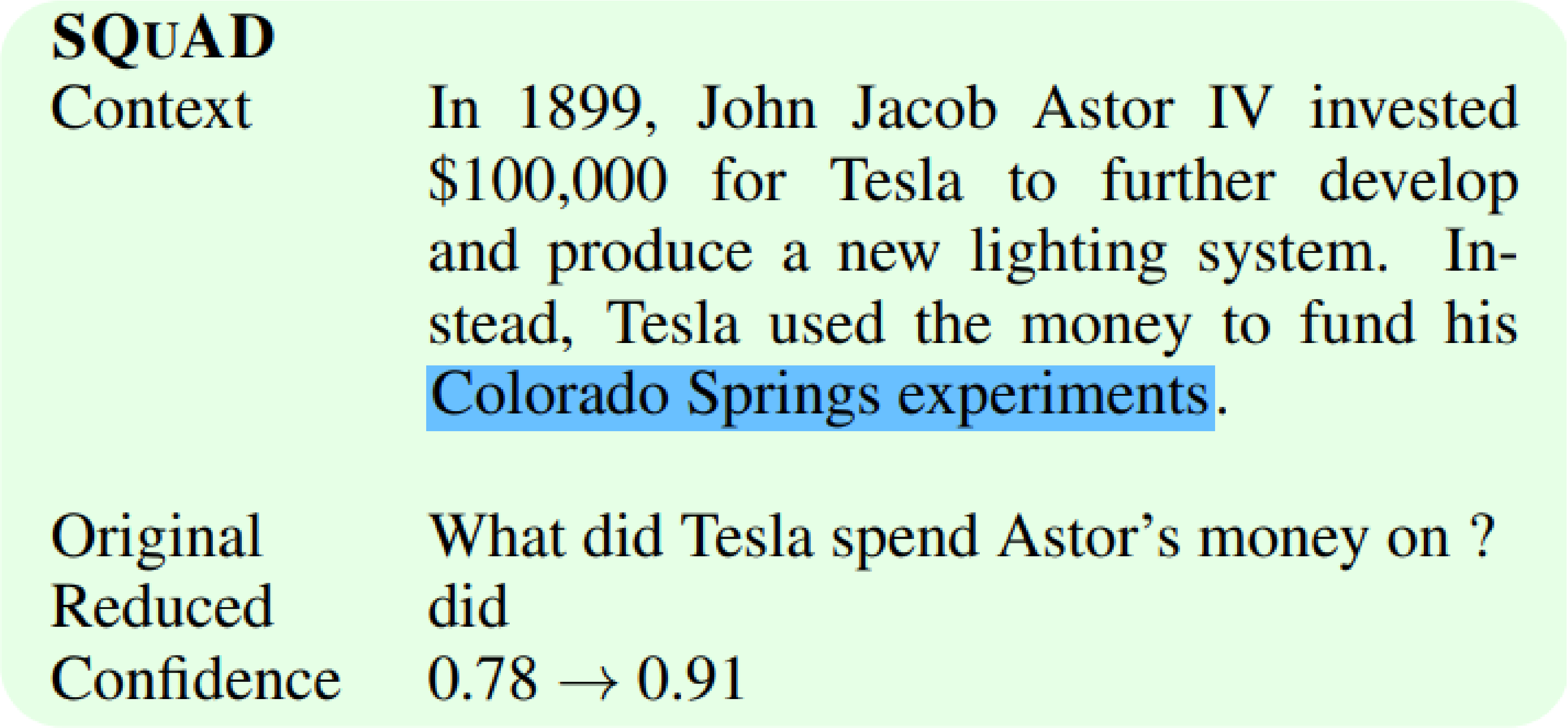
But, local, global optimal 과 같이 gradient가 0인 경우에는 중요도를 잘 반영하지 못한다는 단점이 있음

03

# Influence studies and adversarial examples

Explanation by input reduction [Feng et al., 2018]

- 같은 prediction을 갖는 가장 작은 input은? (rubbish example)
  - rubbish example : 사람은 대답하기 힘들지만, model의 confidence는 높은 예제



SQuAD

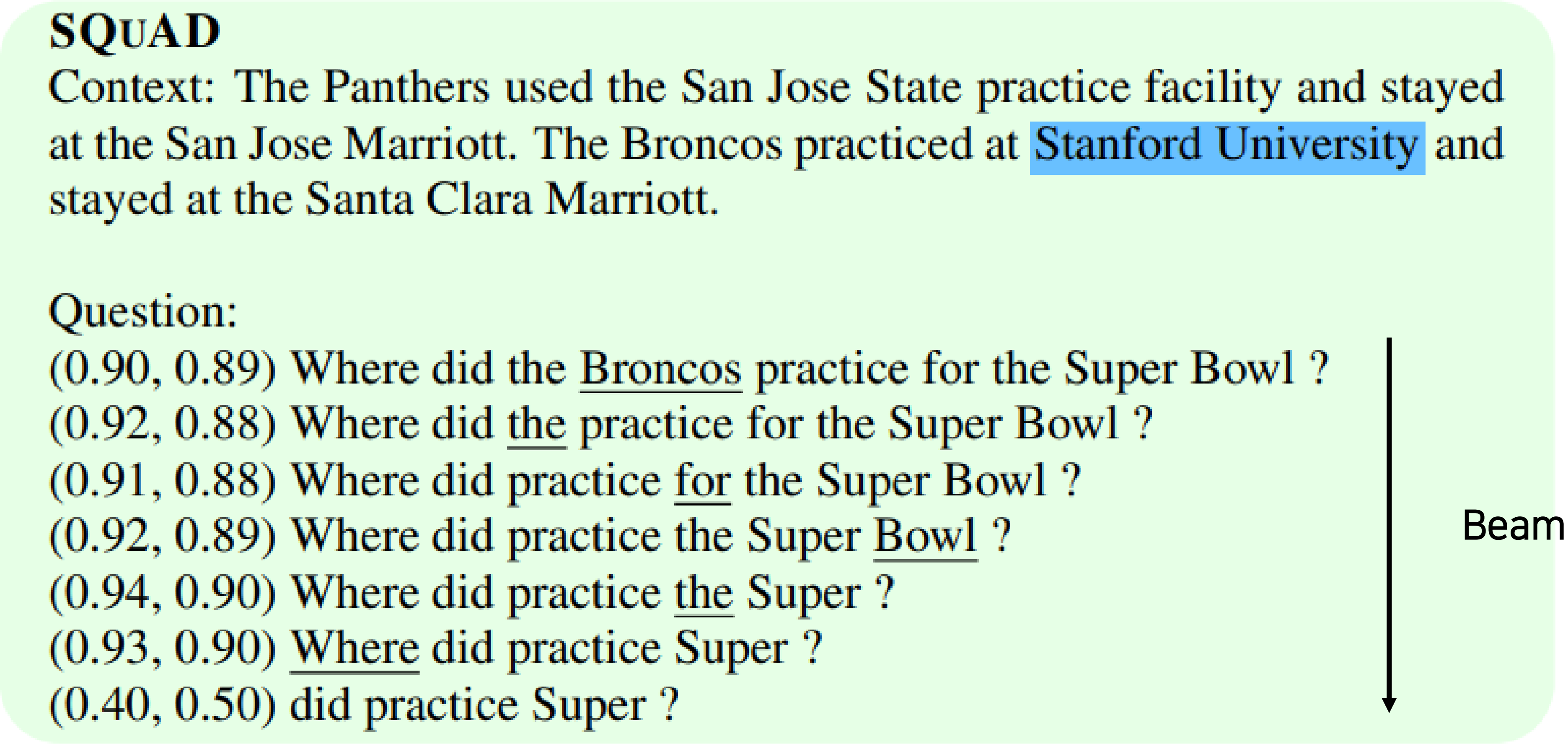
| | |
|---|---|
| Context | In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments. |
| Original | What did Tesla spend Astor's money on ? |
| Reduced | did |
| Confidence | 0.78 → 0.91 |

03

# Influence studies and adversarial examples

Explanation by input reduction [Feng et al., 2018]

- Idea : saliency map을 반복적으로 적용하여 가장 중요도가 낮은 word를 삭제



**SQuAD**

Context: The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott. The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott.

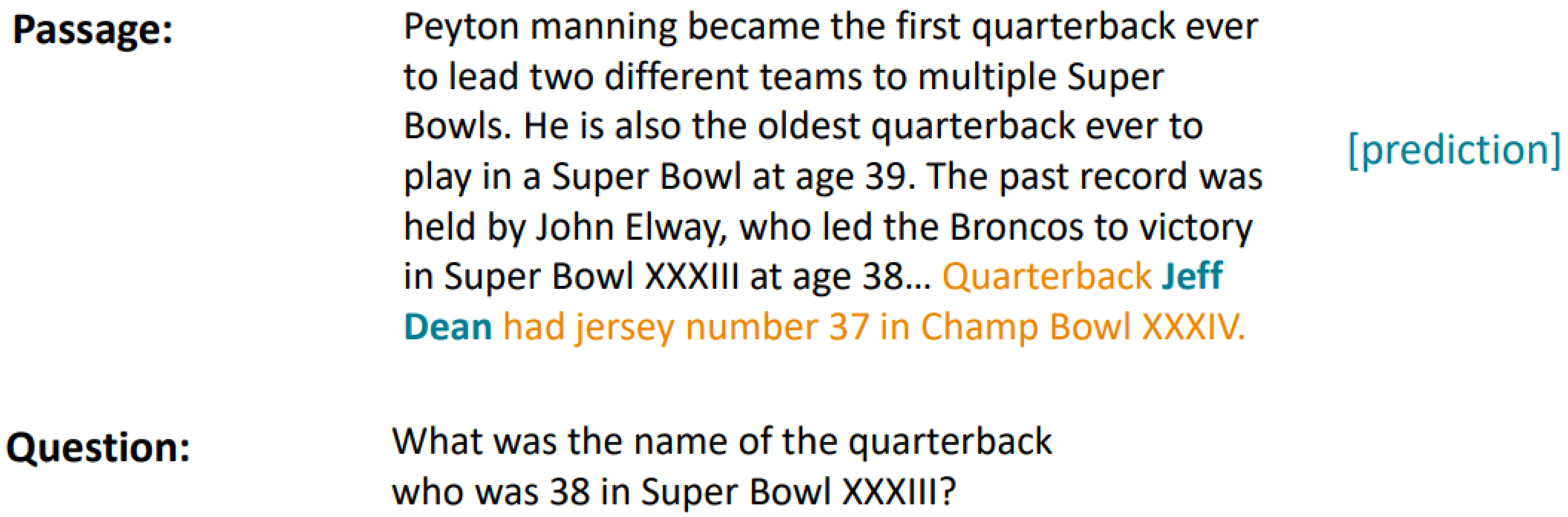
Question:
(0.90, 0.89) Where did the Broncos practice for the Super Bowl ?
(0.92, 0.88) Where did the practice for the Super Bowl ?
(0.91, 0.88) Where did practice for the Super Bowl ?
(0.92, 0.89) Where did practice the Super Bowl ?
(0.94, 0.90) Where did practice the Super ?
(0.93, 0.90) Where did practice Super ?
(0.40, 0.50) did practice Super ?

Beam search

# 03 Influence studies and adversarial examples

Analyzing models by breaking them [Jia et al., 2017]

- Idea : Paragraph 에 문장을 추가하여 model을 부셔보자!



**Passage:** Peyton manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38... Quarterback **Jeff Dean** had jersey number 37 in Champ Bowl XXXIV.

[prediction]

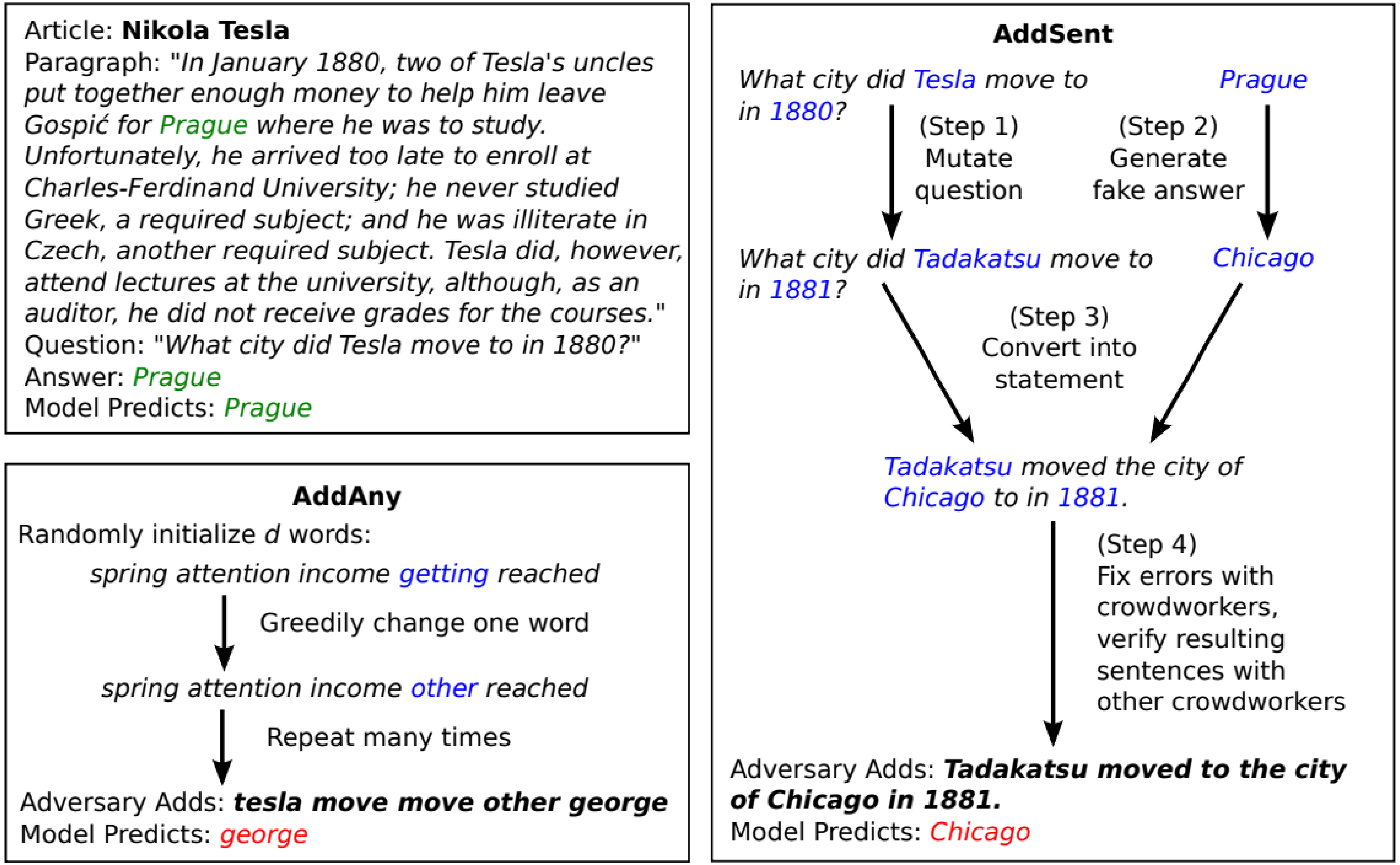
**Question:** What was the name of the quarterback who was 38 in Super Bowl XXXIII?

추가된 문장이 Q의 답을 바꾸지는 않지만 예측은 바뀜 → model 이 우리의 생각대로 동작하지 않음을 시사

03

# Influence studies and adversarial examples

Method to make adversarial example [Jia et al., 2017]



Article: **Nikola Tesla**
Paragraph: "*In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*"
Question: "*What city did Tesla move to in 1880?*"
Answer: *Prague*
Model Predicts: *Prague*

**AddAny**

Randomly initialize *d* words:

*spring attention income getting reached*

↓ Greedily change one word

*spring attention income other reached*

↓ Repeat many times

Adversary Adds: ***tesla move move other george***
Model Predicts: *george*

**AddSent**

*What city did Tesla move to in 1880?* — *Prague*

(Step 1) Mutate question

(Step 2) Generate fake answer

*What city did Tadakatsu move to in 1881?* — *Chicago*

(Step 3) Convert into statement

*Tadakatsu moved the city of Chicago to in 1881.*

(Step 4) Fix errors with crowdworkers, verify resulting sentences with other crowdworkers

Adversary Adds: ***Tadakatsu moved to the city of Chicago in 1881.***
Model Predicts: *Chicago*

# 03 Influence studies and adversarial examples

ADDSENT [Jia et al., 2017]

- **Idea : Question 과 비슷하지만 정답에는 영향을 주지 않는 문장을 만들자**

Step1 : Question에 semantics-altering perturbations 을 적용하여 새로운 question 생성

- 명사, 형용사를 antonym 으로 대체 (WordNet)
- Named entities, number 를 GloVe 에서 가장 가까운 word로 변경
- 만약 해당 사항이 없으면 pass

What **ABC** division handles **domestic** TV distribution? → What NBC division handles foreign TV distribution?

Step2 : 기존 정답과 **같은 type(NER, POS 등의 기준으로 분류)의** fake answer 생성

Step3 : Step1의 **question**과 **Fake answer**를 결합하여 sentence 생성

Q: What NBC division handles foreign TV distribution? + Fake answer : Central Park

→The NBC division of Central Park handles foreign TV distribution

Step4 : Step3에서 생성된 sentence의 문법적 오류를 crowdsourcing을 통해 수정

# 03 Influence studies and adversarial examples

ADDANY [Jia et al., 2017]

- Idea : d개의 단어로 이루어진 random sequence를 생성하여 추가하자 (논문의 경우 d=10)

Step1 : common English word 10개 random하게 생성(Brown corpus)

Step2 : Candidate words W 생성 → common English word 20개 words + Question에 있는 모든 단어

Step3 : 6epochs 만큼 각 단어들에 대해 local search 진행(F1-Score를 최소화 하는 방향으로)

Step4 : 만약 3epoch까지 학습해도 F1-score가 0이 안 되면, random하게 4개의 단어를 초기화

# 04 Contents

1. Motivating model analysis and explanation
2. Out-of-domain evaluation sets
3. Influence studies and adversarial examples
4. **Analyzing representations**
5. Revisiting model ablations as analysis
6. Conclusion

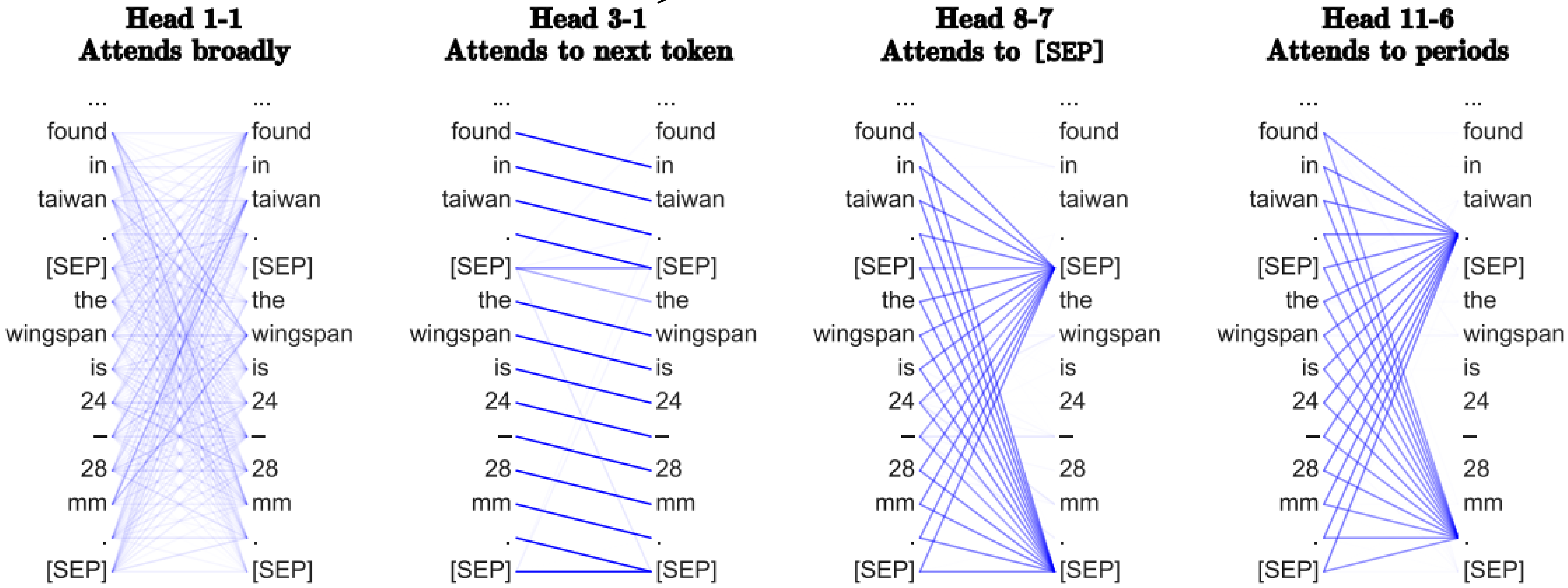
04

# Analyzing representations

Analysis of "interpretable" architecture components [Clark et al., 2018]

- Idea : Attention map을 이용하여 각 layer가 어떤 역할을 하는지 추측하자.

<layer>-<head_number>



# 05 Contents

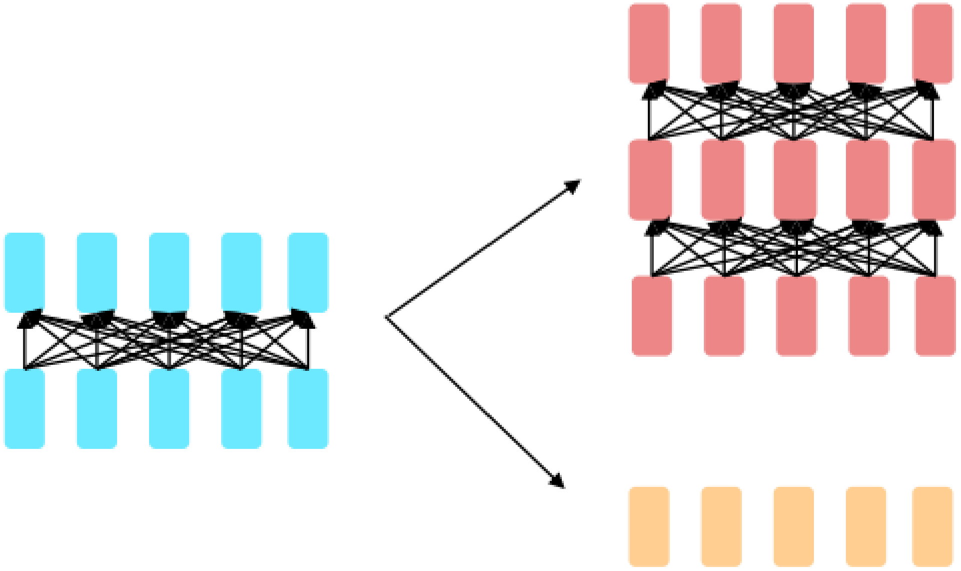
1. Motivating model analysis and explanation
2. Out-of-domain evaluation sets
3. Influence studies and adversarial examples
4. Analyzing representations
5. **Revisiting model ablations as analysis**
6. Conclusion

05

# Revisiting model ablations as analysis

Recasting model tweaks and ablations as analysis

- Ablation을 analysis 관점으로 접근



- Layer를 깊게 vs 얇게 ?

  Model의 accuracy 상승을 위해 **Model tuning**을 진행

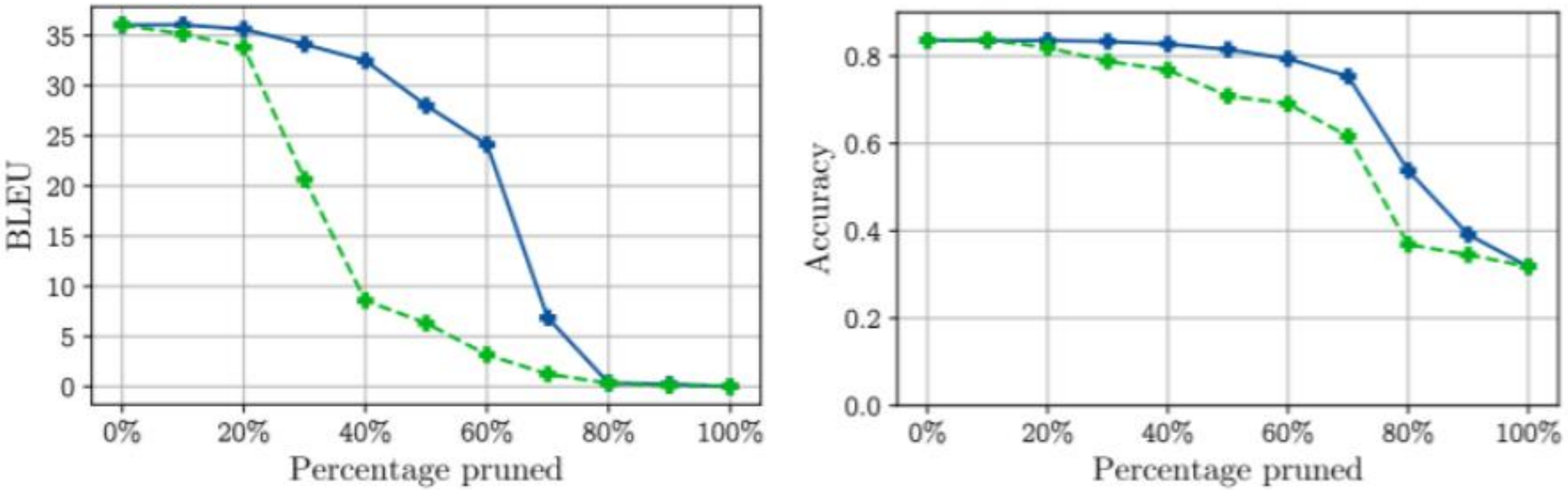
  → 일련의 tuning 과정을 analysis의 관점으로도 해석 가능

# 05 Revisiting model ablations as analysis

Ablation analysis: do we need all these attention heads? [Michel et al., 2019]

- Attention heads 수에 따른 예측 정확도

몇몇 Attention head는 제거되어도 Model의 accuracy에 영향을 주지 않음을 시사



(a) Evolution of BLEU score on newstest2013 when heads are pruned from WMT.

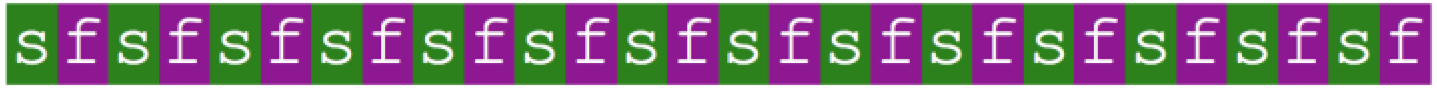
(b) Evolution of accuracy on the MultiNLI-matched validation set when heads are pruned from BERT.

※ 파란색, 초록색 선은 각각 다른 pruning method를 사용 (파랑 : proxy score, 초록: masking attention head task result)

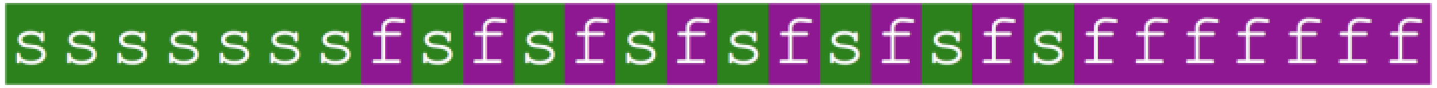
# 05 Revisiting model ablations as analysis

Ablation analysis: What’s the right layer order for a Transformer? [Press et al., 2019]

- Transformer model layer의 순서
  - Transformer의 경우 **Self-attention → Feed-forward** 의 순서를 반복하는 구조 (Layer norm, Residual 제외)
  - **Self-attention → Feed-forward** 를 반복하는 구조가 최적의 구조일까?



Achieves 18.40 perplexity on a language modeling benchmark



Achieves 17.96 perplexity on a language modeling benchmark

# 06 Contents

1. Motivating model analysis and explanation
2. Out-of-domain evaluation sets
3. Influence studies and adversarial examples
4. Analyzing representations
5. Revisiting model ablations as analysis
6. Conclusion

# 06 Conclusion

## Summary

- Neural model은 복잡하고 특징을 잡아내는 것은 어렵다.
- Model의 직관적인 설명은 어렵지만, 특정 부분에 집중하여 분석을 하는 다양한 방법이 존재한다.
- Model을 분석하고 설명하는 것이 주된 목표가 아니더라도 Model의 구축 과정에 접목할 수 있다.
- 좋은 Model을 구축하여 SOTA를 달성하는 것도 좋은 논문 주제이지만, 기존 Model을 분석, 추론하는 것 역시 좋은 주제가 될 수 있을 것이다.

# 감사합니다

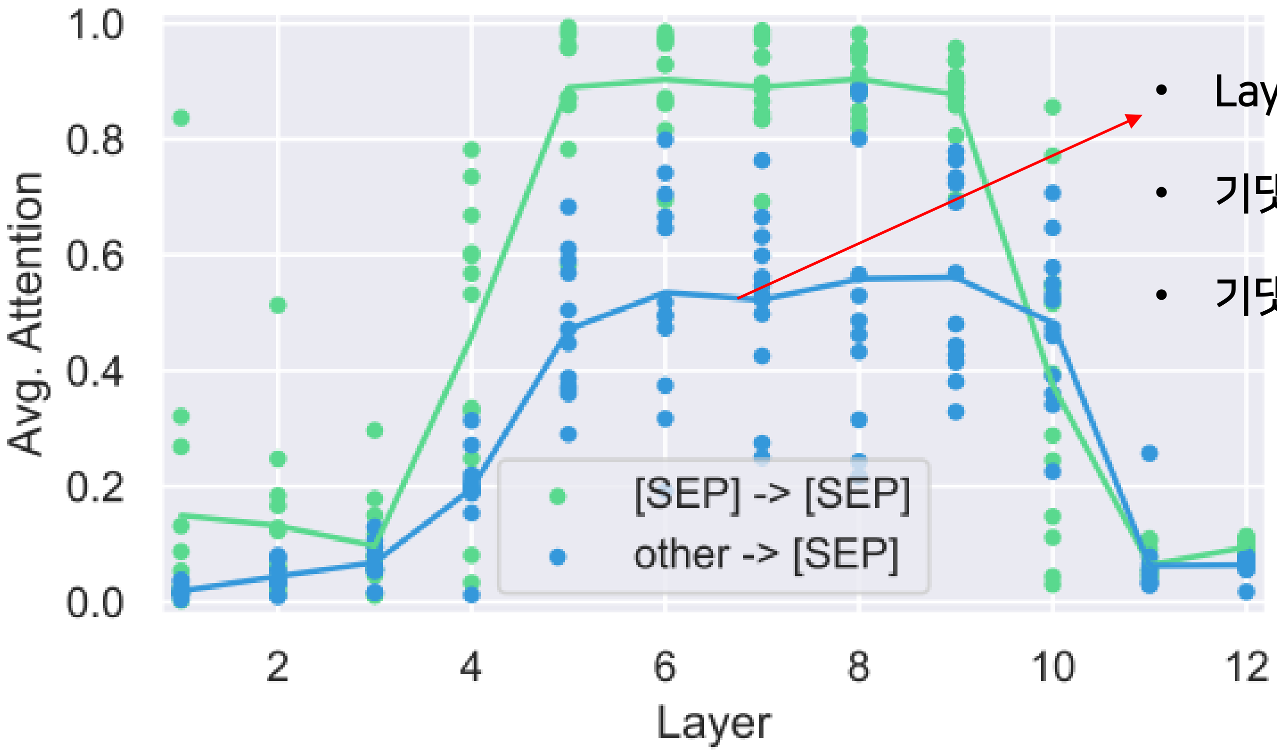
04

# Analyzing representations

Analysis of “interpretable” architecture components [Clark et al., 2018]

- Case 분석 : [SEP] token

가설1 :[SEP] TOKEN 이 정보를 모아주는 역할을 할 것 이다.



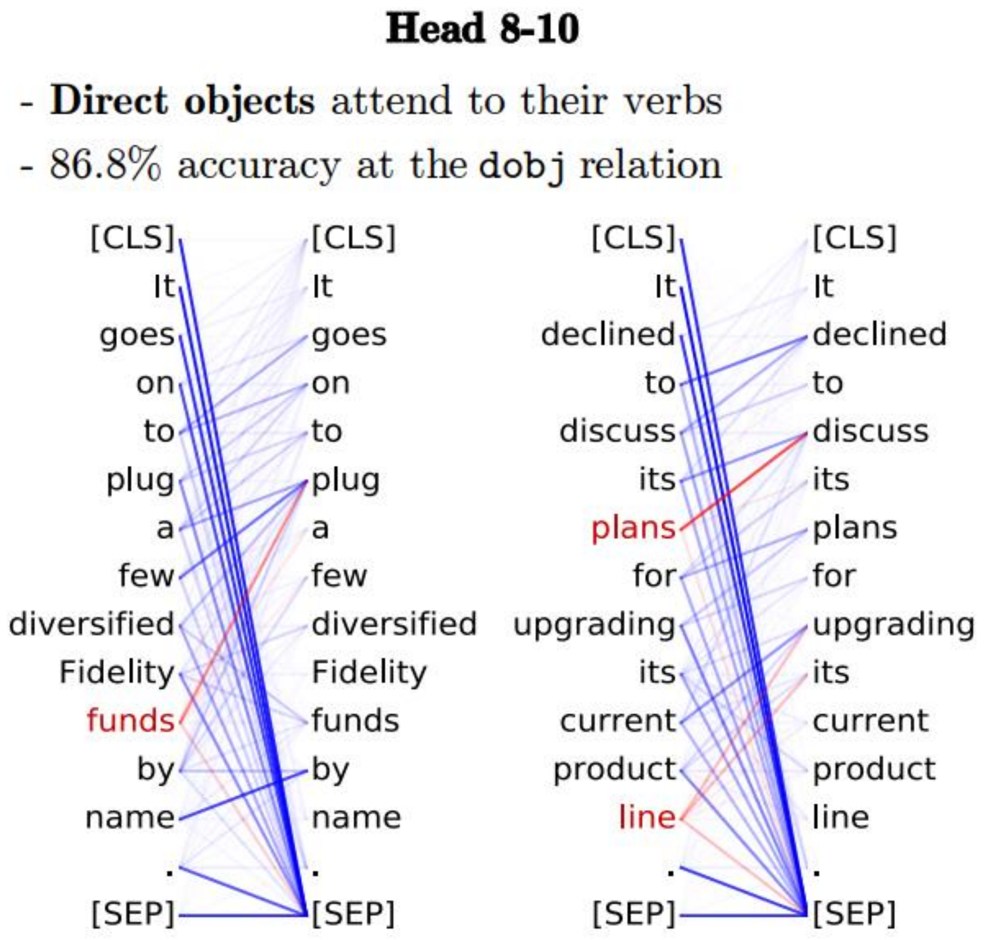
- Layer 6 ~ 10 의 경우 50% 이상의 attention 이 [SEP]
- 기댓값은 1/64 (Segment 128 tokens, [SEP] 2 개 포함)
- 기댓값 대비 상당히 유의미하게 많은 부분을 차지

만약 위 가설이 맞으면 Attention head가 [SEP]을 통해 전체적인 Segment 처리를 해야한다.

# Analyzing representations

Analysis of "interpretable" architecture components [Clark et al., 2018]

- Case 분석 : [SEP] token



- Object의 경우 각 object의 verb와 강한 연결
- Non-noun 의 경우 [SEP] 과 강한 연결

→ [SEP] TOKEN 이 no-op 의 역할을 한다!

(Attention이 기능을 제대로 하지 못할 때)

00

# Appendix

000