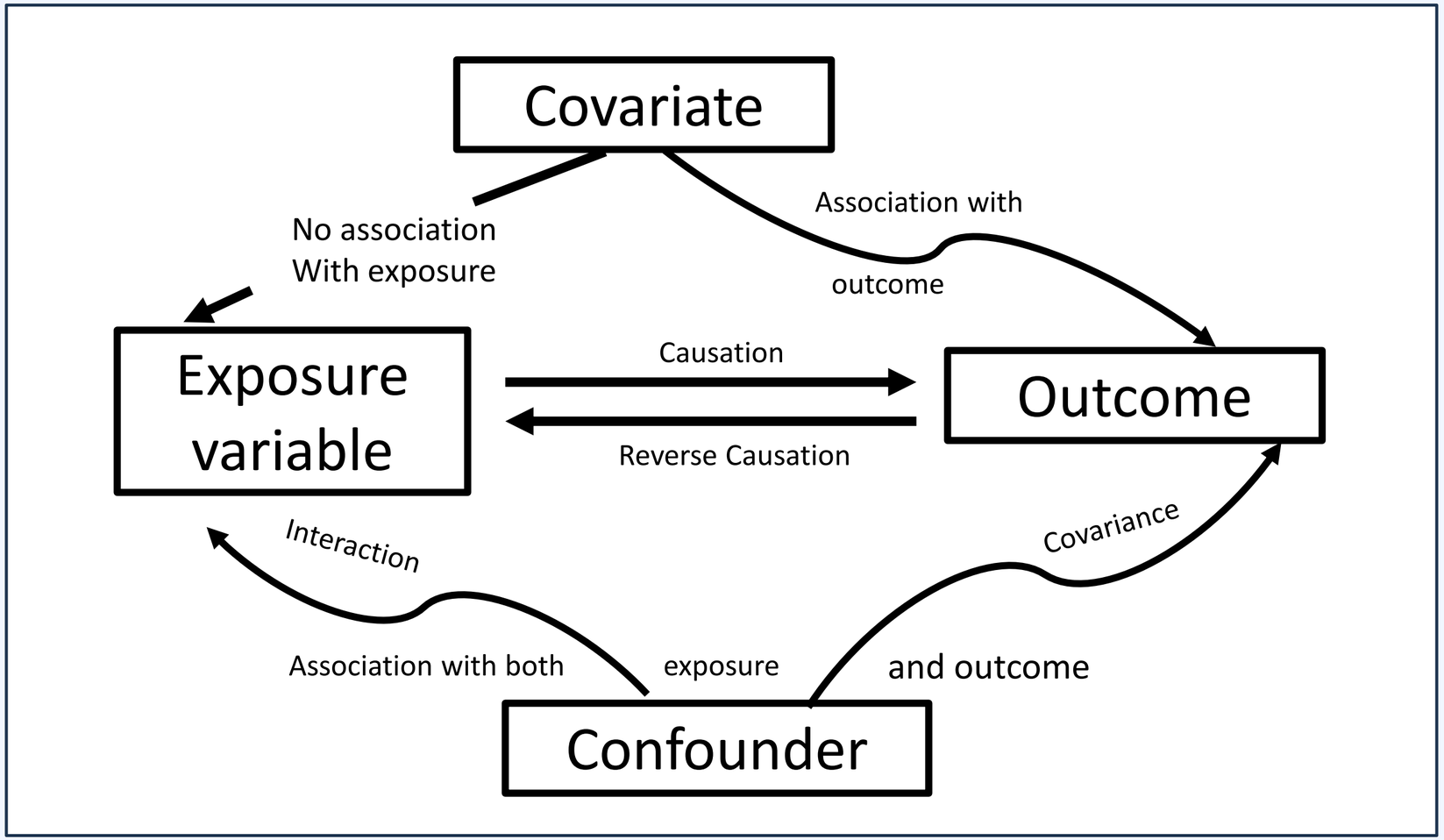
KUBIG CONTEST

NLP 분반 CausaLM

15기 박지우, 18기 백성은

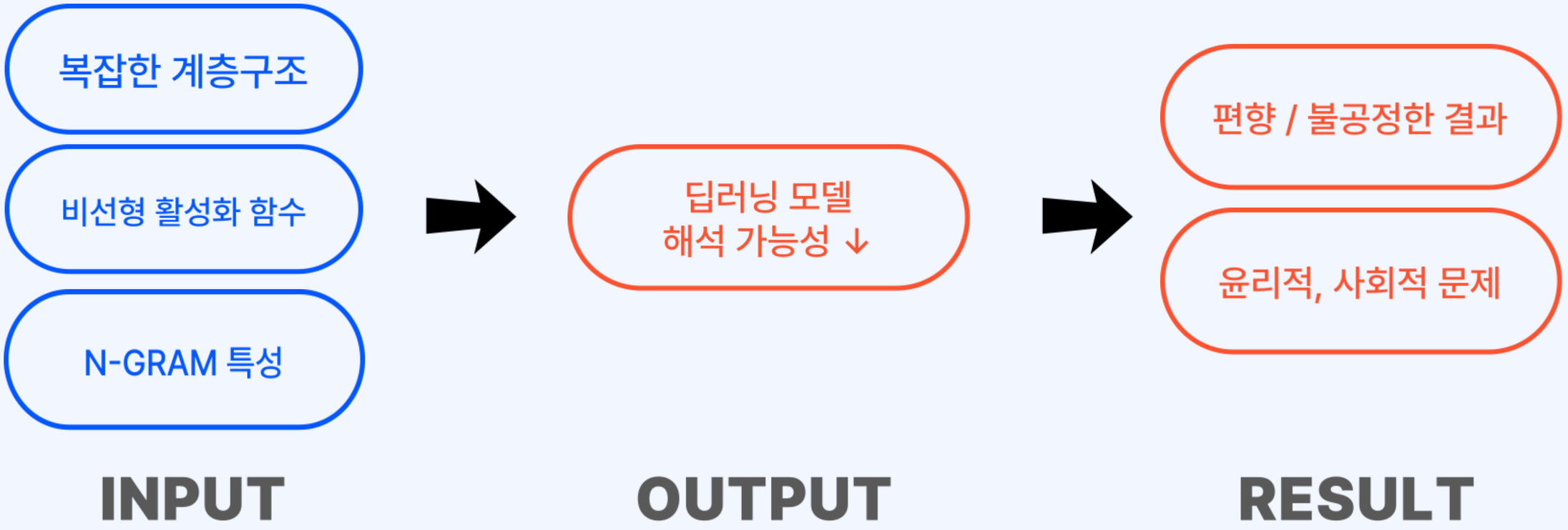
Content

1.Preliminary



Causality

2. Motivation



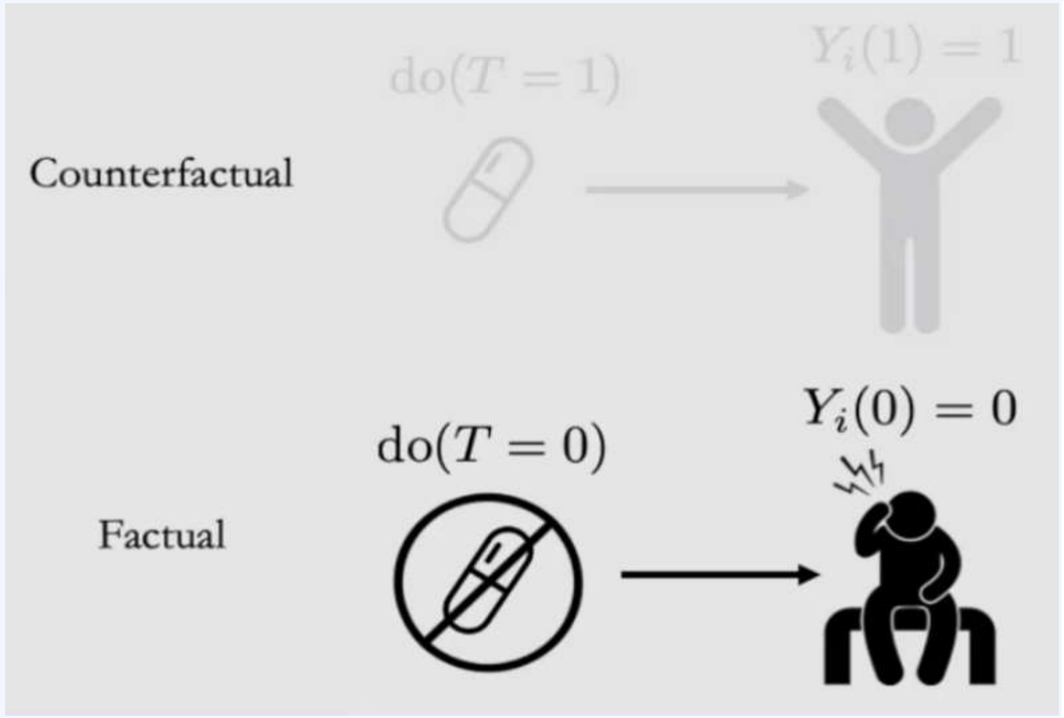
3. CausaLM

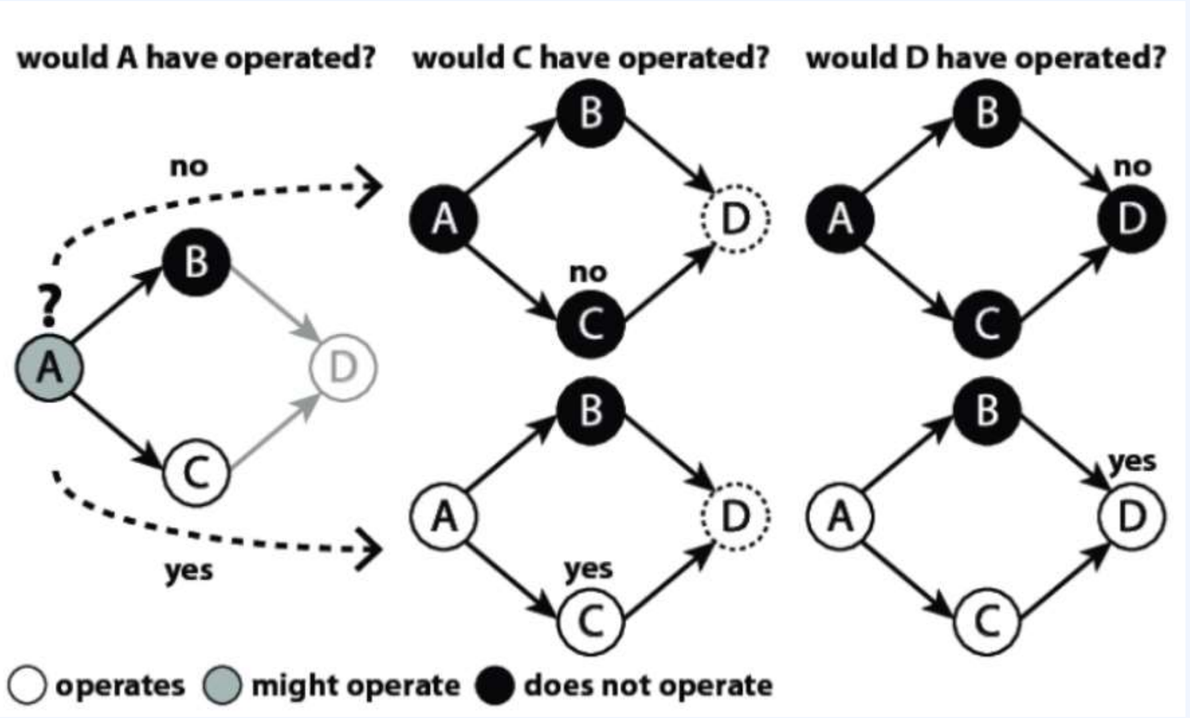
President **Trump** did his best imitation of **Ronald Reagan** at the State of the Union address, falling just short of declaring it Morning in America, the **iconic** imagery and message of a campaign ad that **Reagan** rode to re-election in 1984. **Trump** talked of Americans as pioneers and explorers; he lavished praise on members of the military, several of whom he recognized from the podium; he **optimistically** declared that the best is yet to come. It was a **masterful** performance – but behind the **sunny** smile was the same old **Trump**: **petty**, **angry**, **vindictive** and **deceptive**. He refused to shake the hand of House Speaker **Nancy Pelosi**, a snub she returned in kind by ostentatiously ripping up her copy of the President's speech at the conclusion of the address, in full view of the cameras.

- Adjective (형용사) → 모델 성능에 영향
- 'Trump'와 같은 논란이 될만한 내용들은 어떻게 ?

causal concept-based explanation

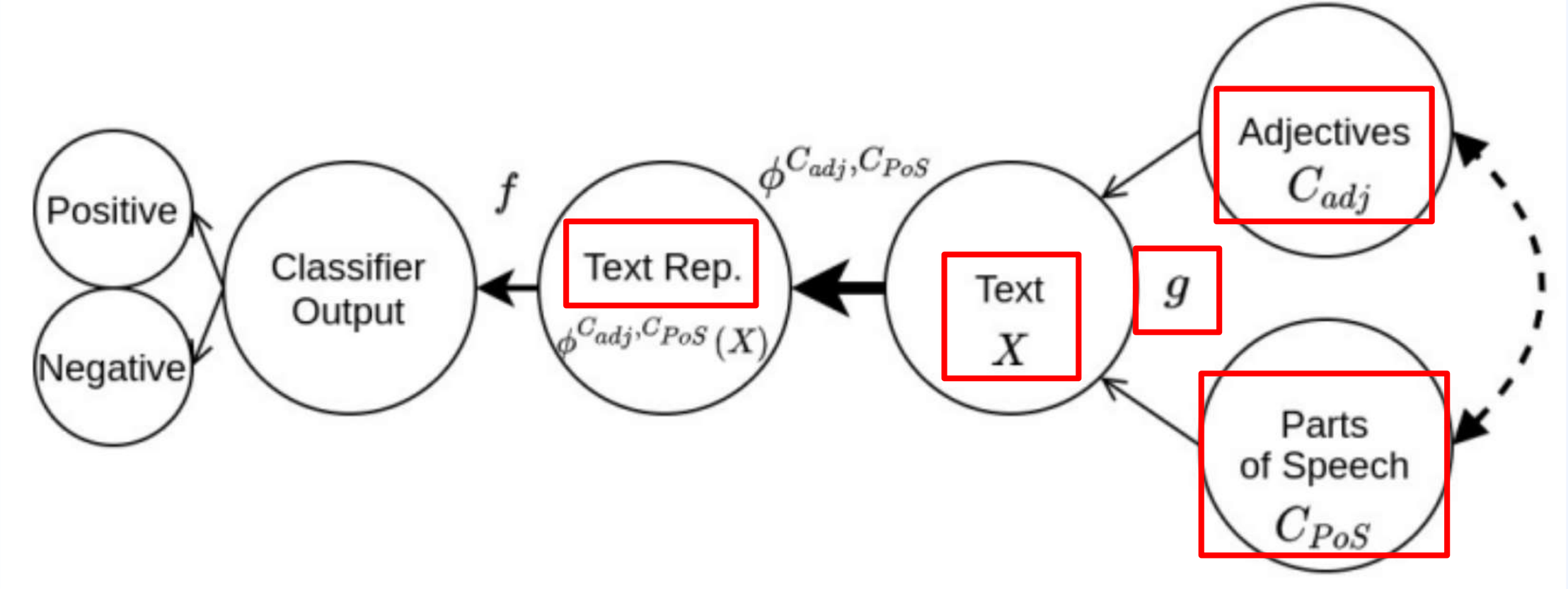
3. CausaLM





Counterfactual Example vs Factual Example

3. CausaLM (Causal graph)



3. CausaLM (Generating concept)

Treated concept

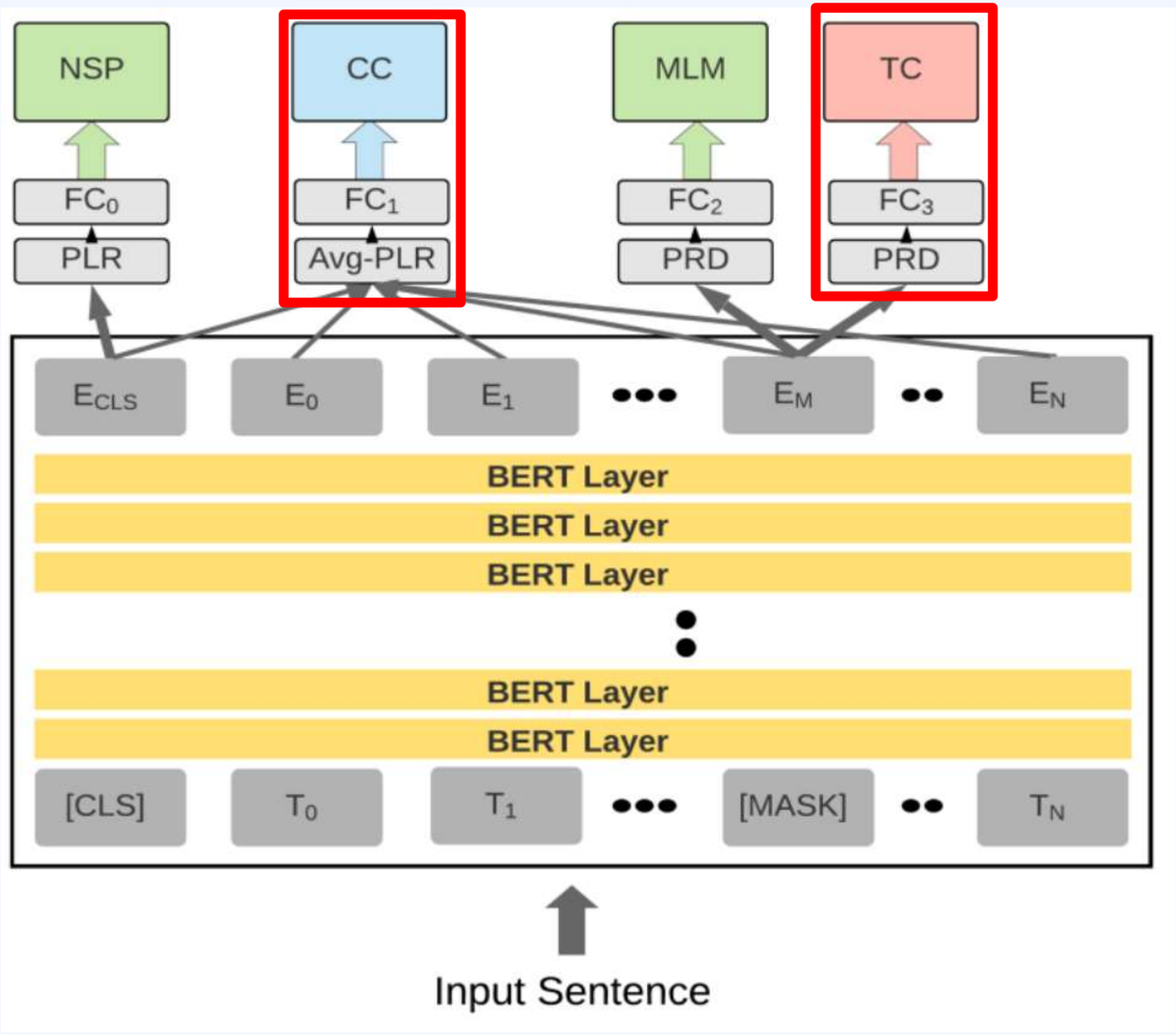
- 실험군
- 실험을 위한 Treatment를 적용하여 텍스트를 생성하는 Concept
- 최종 결과에서 treated concept의 effect가 'Forget' 되었는지 확인
- Adjective / Not_Adjective

vs

Control concept

- 대조군
- Treated concept을 제외한 나머지 concept
- 최종 결과에서 control concept이 계속 'Remember' 되어 있는지 확인
- ex) '미국' / 'Topic'

3. CausaLM (BERT Model)



3. CausaLM (Loss function & TReATE)

- **Loss Function**

$$\mathcal{L}(\theta_{bert}, \theta_{mlm}, \theta_{nsp}, \theta_{cc}, \theta_{tc}) = \frac{1}{n}\Big(\sum_{i=1}^{n} \mathcal{L}^{i}_{mlm}(\theta_{bert}, \theta_{mlm}) + \sum_{i=1}^{n} \mathcal{L}^{i}_{nsp}(\theta_{bert}, \theta_{nsp}) + \sum_{i=1}^{n} \mathcal{L}^{i}_{cc}(\theta_{bert}, \theta_{cc}) - \lambda \sum_{i=1}^{n} \mathcal{L}^{i}_{tc}(\theta_{bert}, \theta_{tc}) \Big)$$

- **TReATE**

$$\mathrm{ATE}_T = \mathbb{E}\left[Y | do(T=1)\right] - \mathbb{E}\left[Y | do(T=0)\right]$$

4. Experiment

REPRODUCING(MLM)

Balanced

Positvie & Negative

Gentle

Postive & Negative(-50%)

Aggressive

Positive(-50%) & Negative (-50%)

	Balanced	Gentle	Aggressive
ATE	0.397	0.376	0.634
TREATE	0.202	0.177	0.313

4. Experiment

REPRODUCING(CF)

Balanced

Positvie & Negative

Gentle

Postive & Negative(-50%)

Aggressive

Positive(-50%) & Negative (-50%)

	Balanced	Gentle	Aggressive
ATE	0.397	0.376	0.634
TREATE	0.273	0.361	0.475

4. Experiment

PRODUCING TEST(MLM)

	Bert-cased in paper	Bert-cased in test	Bert-large-uncased
Aggressive → balance	0.744		0.947

4. Experiment

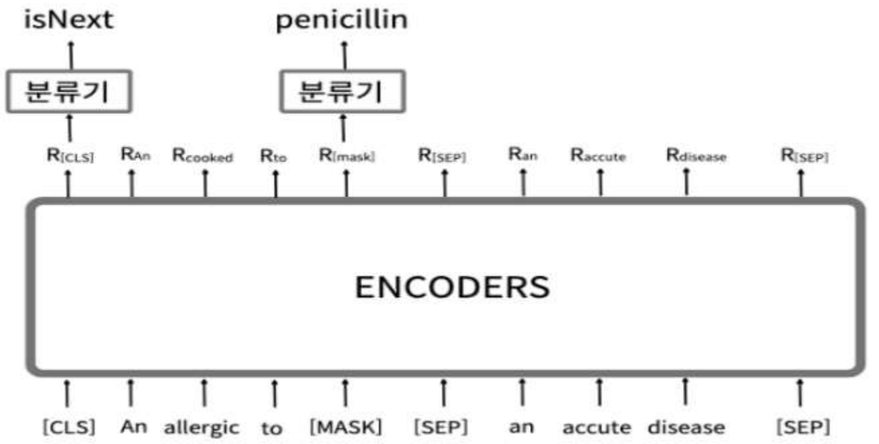
PRODUCING TEST(CF)

	Bert-cased in paper	Bert-cased in test	Bert-large-uncased
Aggressive → balance	0.793		0.512

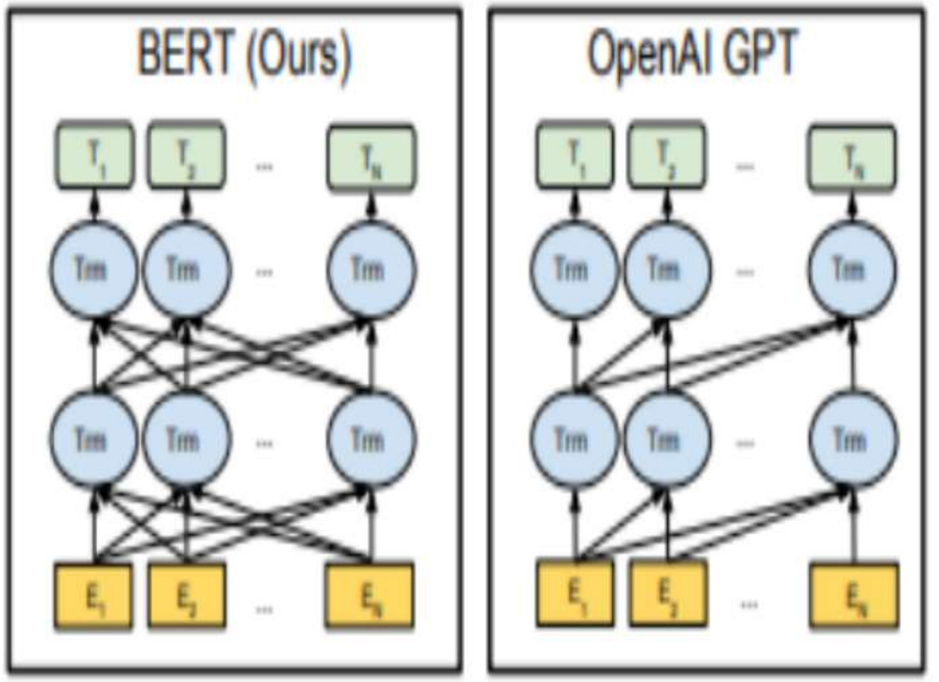
5. Further study

Clinical BERT

- ClinicalBERT는 대규모 임상 말뭉치에서 사전 학습된 임상 domain-BERT 모델이다.
- Domain Bert : 특정 도메인에 대한 데이터셋 학습한 BERT
  - 임상노트(clinical note), 진행 노트(progress note)는 환자에 대한 매우 유용한 정보를 포함한다



BERT → GPT



6. Error

```
elif n_gpu > 1:
    model = torch.nn.DataParallel(model)
```