

# 함께 자라기

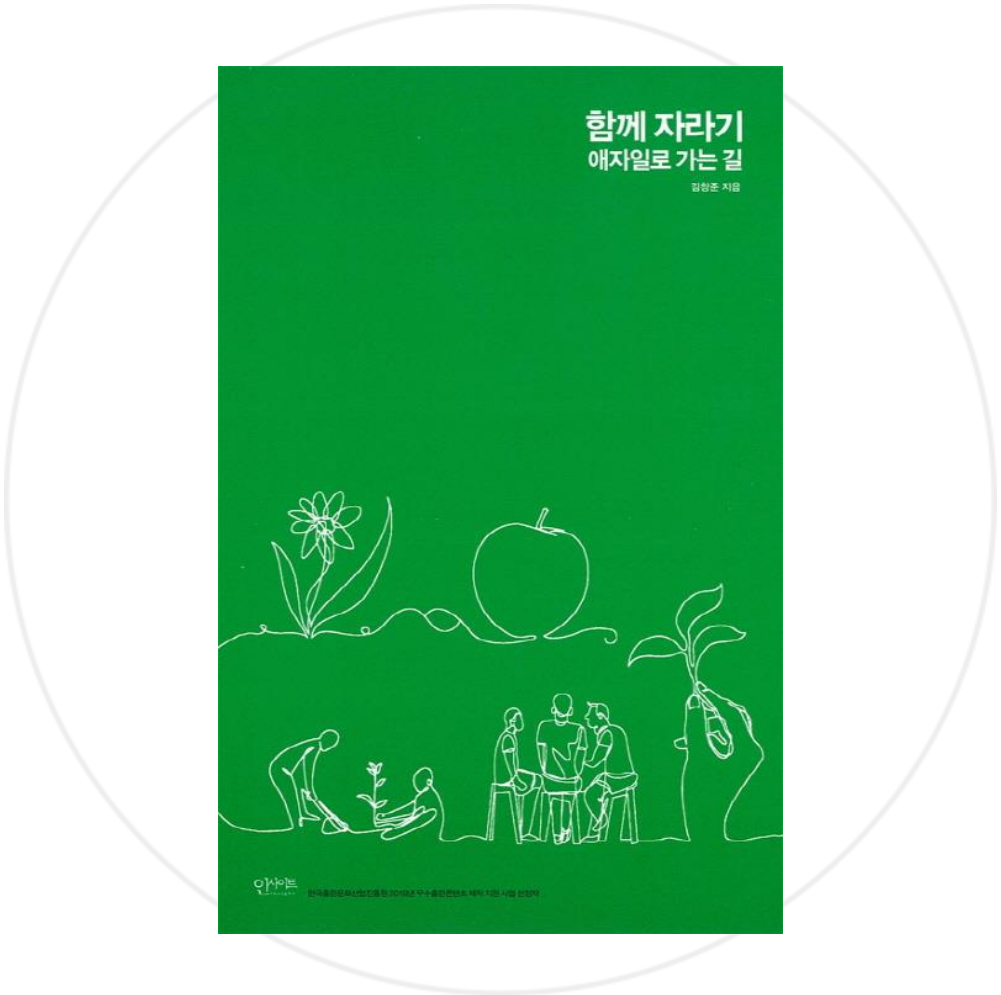
KLUE : 문장 내 개체간 관계 추출 대회  
SOLUTION 발표

7조 김준재 김현아 배현진 이강민 최성원

# INTRODUCTION



sesame street  
열려라 참깨



함께 자라기

안녕하세요! NLP 7조 참깨자라기 입니다!

# 협업 + @

## 이정표: 이번 대회에서 얻어가고 싶은 것

- 텍스트 데이터 Augmentation에 대한 충분한 경험
- 깃헙으로 협업하는 방법 익히기
- huggingface 자유롭게 사용하기

## ✪그라운드 룰(P 스테이지)


- 대회 제출 룰: 매일 밤 9시 슬랙 봇이 올리는 질문에 답변(오늘 제출 했는지/할 예정인지, 제출 안 할 것인지 → 이를 통해 여러분의 제출횟수가 나오고, 이걸 쓴 사람은 스레드로 보고하기)
- 보류: Validation 데이터 고정하기(모두의 실험 조건을 동일하게 만듭니다.)
  - metric:
    - 1) no\_relation class를 제외한 micro F1 score
    - 2) 모든 class에 대한 area under the precision-recall curve (AUPRC)
      - 2가지 metric으로 평가하며, micro F1 score가 우선시 됩니다.
- 팀 대회 리더보드 실험 로그엔 본인이 제출한 모델의 정보만 올릴 것 → 각자의 자잘한 실험 로그는 개인 실험 로그에 따로 기록
- WandB를 통해 실험 내용 공유하기

## GIT


- 브랜치 규칙

Conventional Commits

A specification for adding human and machine readable meaning to commit messages

 <https://www.conventionalcommits.org/en/v1.0.0/>

- master - 모두가 사용하게 될 브랜치
  - feature - 추가, 수정하게 될 코드를 실험을 끝낸 상태에서, 타 팀원에게 리뷰를 요청하는 브랜치
    - test - 개인이 추가, 수정하고 싶은 코드를 바로 적용하여 셀프 디버깅 하는 브랜치
- <https://gmlwjd9405.github.io/2017/10/27/how-to-collaborate-on-GitHub-1.html>
- Pull request 양식
  - <https://velog.io/@always0ne/Commit-Pull-Request-issue-템플릿-사용하기>

 Slack Notification

오후 4:10

HyunAh-Kim-Clou

Ref  
refs/heads/master

Actions URL  
[Slack Notification](#)

Message  
Merge pull request #17 from boostcampaitech3/feature/hidden\_emb

Feature/hidden\_emb

Event  
push

Commit  
400b74

Powered By rtCamp's GitHub Actions Library

 Daily Submission Poll

워크플로

오후 9:00

3월 21일 월요일 오후 9:00:16 오늘의 제출 현황 조사 @channel

오늘 중으로 더 제출할 게 있는 분들은 👍


아닌 분들은 😊 눌러주세요!

👍 2

😊 3

😊

현진



3개의 답글 18일 전 마지막 답글

Projects

Create new project

Hidden Emb test  
growing\_sesame

32 runs  
Last ran 19 hours ago

[junejae]eval\_aug\_test  
growing\_sesame

70 runs  
Last ran 19 hours ago

FAST\_TEST  
growing\_sesame

66 runs  
Last ran 19 hours ago

baseline  
growing\_sesame

94 runs  
Last ran 19 hours ago

KFOLD\_TEST  
growing\_sesame

10 runs  
Last ran 23 hours ago

Model\_Test  
growing\_sesame

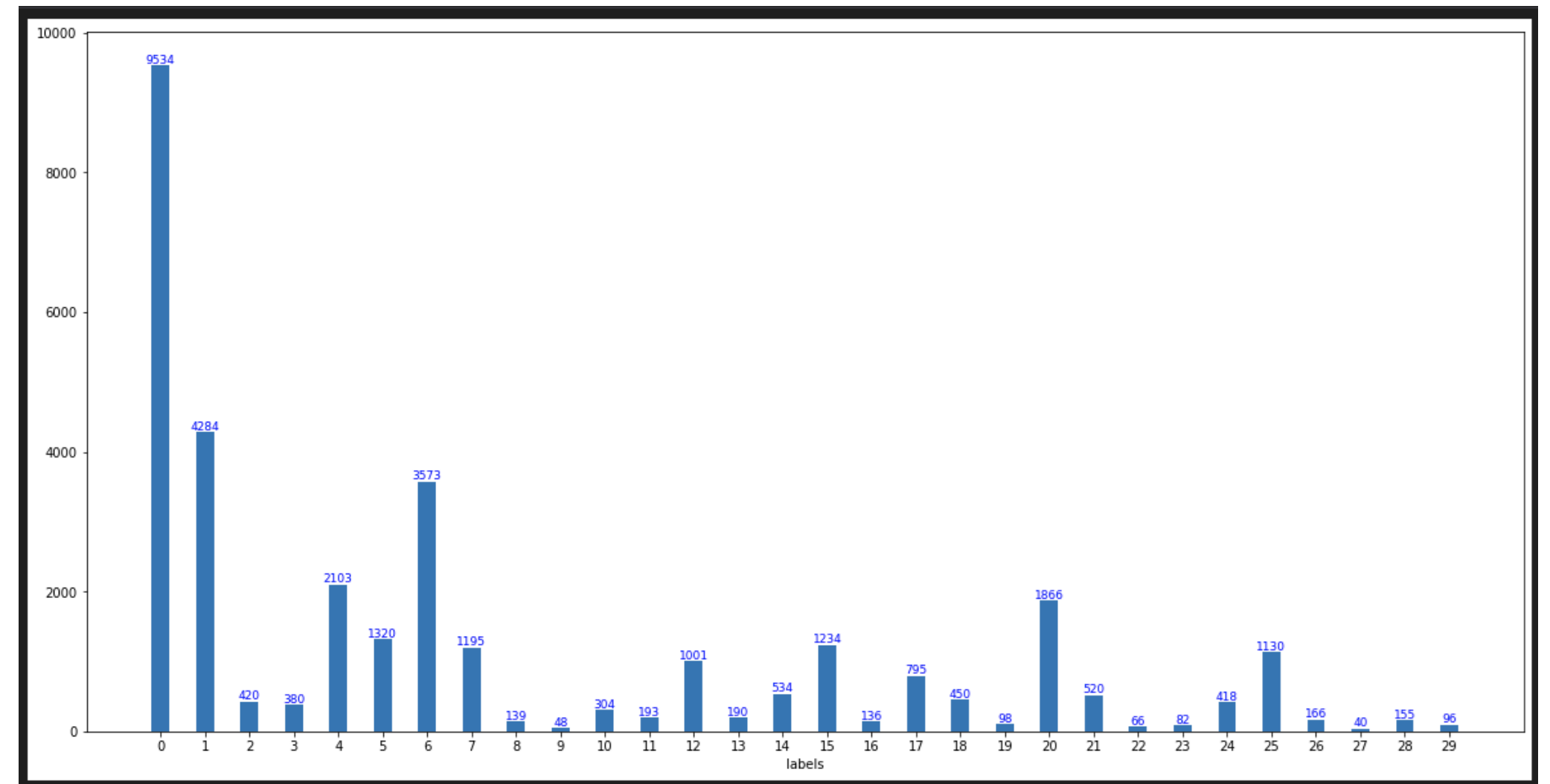
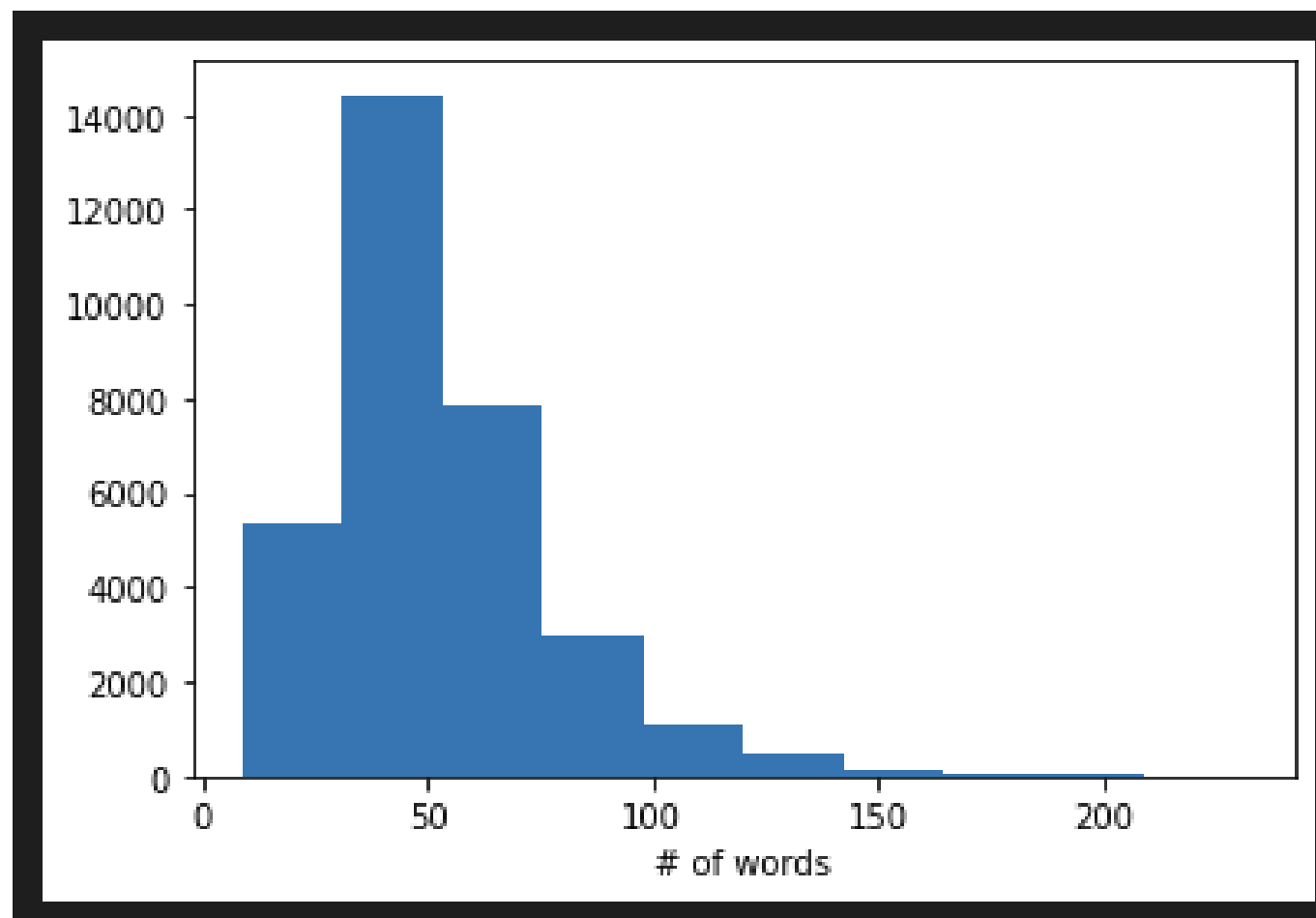
1 run  
Last ran 2 days ago

See all →

<input type="checkbox"/>	0 Open ✓ 20 Closed
<input type="checkbox"/>	<a href="#">add weighted voting ensemble</a> ✓ #20 by HyunAh-Kim-Clou was merged 2 days ago 3 tasks
<input type="checkbox"/>	<a href="#">added question sentence</a> ✓ #19 by hyunbool was merged 2 days ago 5 tasks
<input type="checkbox"/>	<a href="#">add ensemble_probs fn</a> ✓ #18 by HyunAh-Kim-Clou was merged 3 days ago 3 tasks
<input type="checkbox"/>	<a href="#">Feature/hidden_emb</a> ✓ #17 by HyunAh-Kim-Clou was merged 3 days ago 5 tasks
<input type="checkbox"/>	<a href="#">Feature/sota</a> ✓ #16 by hyunbool was merged 4 days ago 5 tasks
<input type="checkbox"/>	<a href="#">Feature/sota</a> ✓ #15 by hyunbool was closed 4 days ago 4 tasks
<input type="checkbox"/>	<a href="#">fixed code for load data</a> ✓ #14 by hyunbool was closed 6 days ago 5 tasks
<input type="checkbox"/>	<a href="#">Feature/hidden_emb</a> ✓ #13 by HyunAh-Kim-Clou was merged 6 days ago 3 tasks
<input type="checkbox"/>	<a href="#">updated modified_load_data.py</a> ✓ #12 by hyunbool was merged 8 days ago 4 tasks
<input type="checkbox"/>	<a href="#">Feature/tapt</a> ✓ #11 by Lkangmin was merged 9 days ago 1 of 6 tasks
<input type="checkbox"/>	<a href="#">Feature/text aug</a> ✓ #10 by Junejae was merged 10 days ago 2 of 4 tasks
<input type="checkbox"/>	<a href="#">update for better shell scripting</a> ✓ #9 by Junejae was merged 11 days ago 2 of 4 tasks

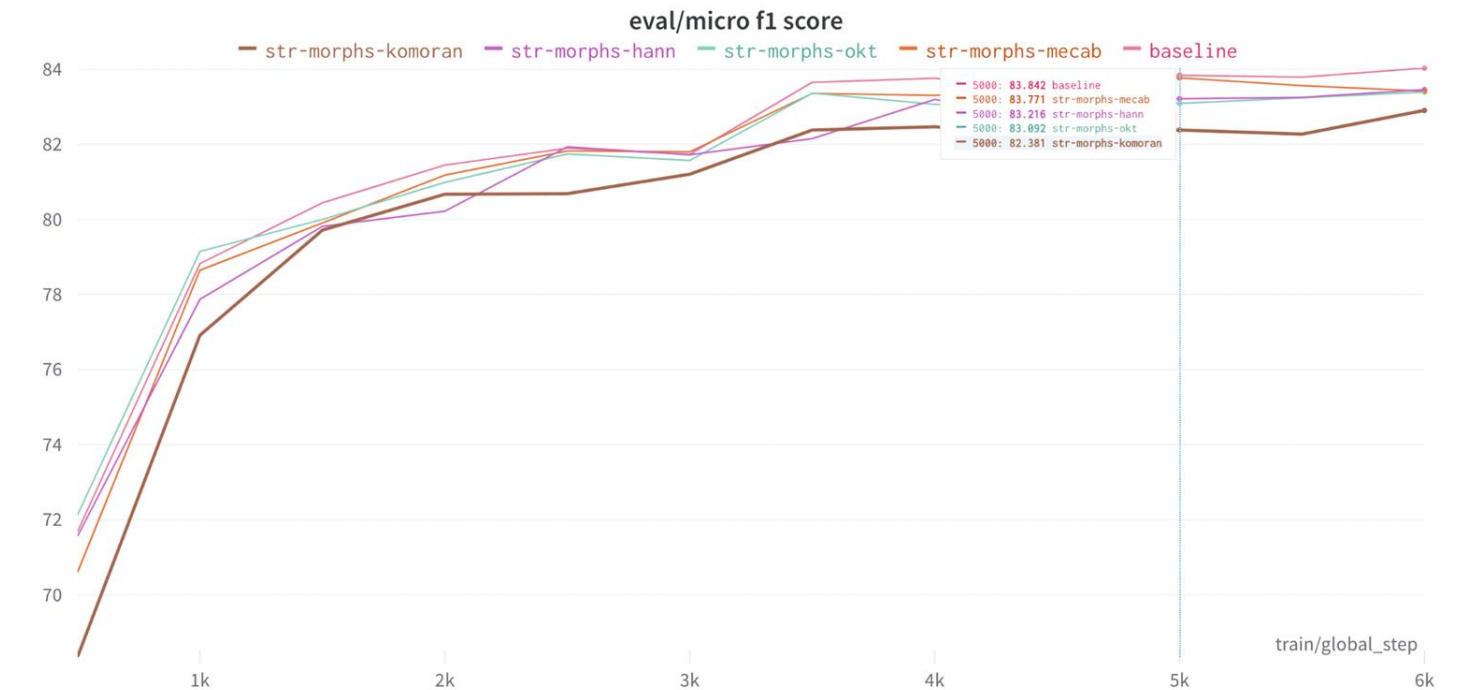
# EDA

- 간단한 EDA
  - 문장 길이 분포
  - 데이터 불균형



# Preprocessing

- 형태소 분석기를 사용하여 형태소 단위로 입력하면, 모델이 더 잘 인식할 수 있지 않을까
  - 명사, 형태소, 조사를 뺀 입력 형태에 대해 실험
- Text Data Augmentation: ktextaug 라이브러리를 이용해 단어 철자를 바꾸는 등의 변형 시도
- .csv 파일의 entity(subject, object) 정보들이 python dict형으로 저장되어 있음을 확인.
- eval() 함수를 통해 string을 dict로 코드화 하여 entity type 정보 추출
- 오피스 아워에서 힌트를 얻어 entity special token를 실험
- Entity Marker(punct), Typed entity marker 실험 시도
- 미스라벨링, 완전 중복 패턴 제거
- etc...



Method	Input Example	BERT <sub>BASE</sub>	BERT <sub>LARGE</sub>	RoBERTa <sub>LARGE</sub>
Entity mask	[SUBJ-PERSON] was born in [OBJ-CITY].	69.6	70.6	60.9
Entity marker	[E1] Bill [/E1] was born in [E2] Seattle [/E2].	68.4	69.7	70.7
Entity marker (punct)	@ Bill @ was born in # Seattle #.	68.7	69.8	71.4
Typed entity marker	<S:PERSON> Bill </S:PERSON> was born in <O:CITY> Seattle </O:CITY>.	71.5	72.9	71.0
Typed entity marker (punct)	@ * person * Bill @ was born in # ^ city ^ Seattle #.	70.9	72.7	74.6

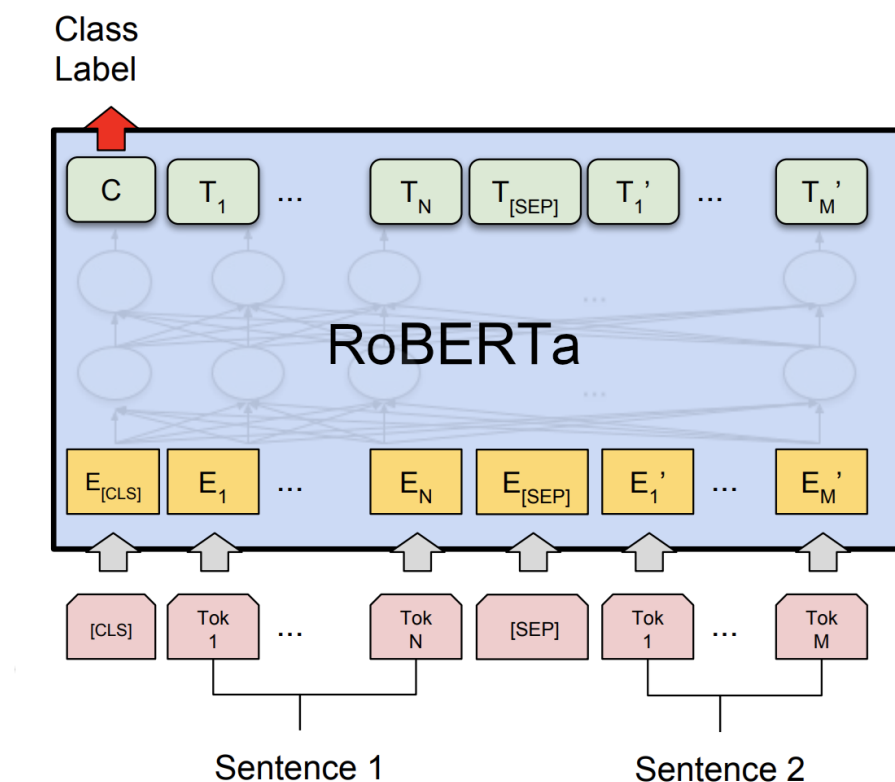
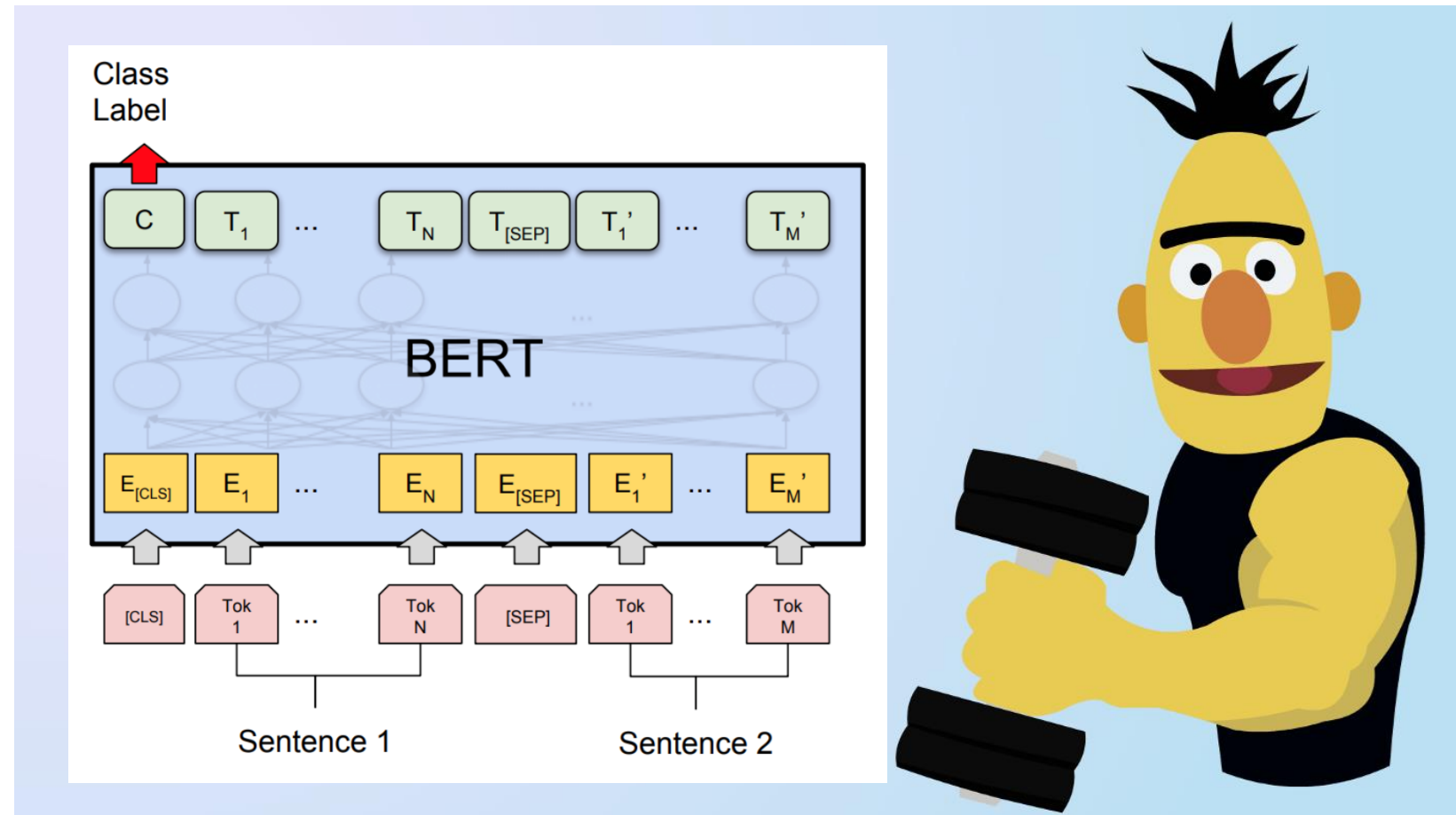
Table 3: Test  $F_1$  (in %) of different entity representation techniques on TACRED. For each technique, we also provide the processed input of an example text “*Bill was born in Seattle*”. Typed entity markers (original and punct) significantly outperforms others.

```
dict_e01 = eval(e01) # str을 코드화
dict_e02 = eval(e02)

e01_word = dict_e01['word']
e02_word = dict_e02['word']
```



# Model



o/e








```
# How to use: type 'sh train.sh' on your CLI
python train.py \
--load_data_filename load_data_junejae \
--load_data_func_load load_data \
--load_data_func_tokenized tokenized_dataset \
--load_data_func_tokenized_train tokenized_dataset \
--load_data_class RE_Dataset \
--metric_for_best_model 'eval_loss' \
--gradient_accumulation_steps 1 \
--use_augmentation True \
--aug_data ../dataset/train/augmented_phonologicalProcess.csv \
--seed 42 \
--model klue/roberta-large \
--train_data ../dataset/train/train_finalCorrection.csv \
--num_labels 30 \
--output_dir ./results \
--save_total_limit 10 \
--save_steps 500 \
--num_train_epochs 2 \
--learning_rate 1e-5 \
--per_device_train_batch_size 34 \
--per_device_eval_batch_size 64 \
--warmup_steps 500 \
--weight_decay 0.0 \
```

실패 (그렇게 2주가 흘러갔습니다.)

---



11	NLP_07조	    	69.9725	71.6246
----	---------	---	---------	---------

---

## 멘토님의 조언

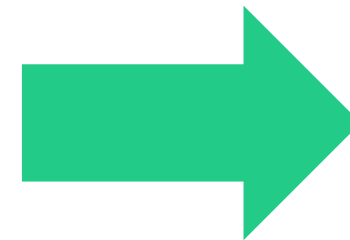


지금까지의 실험은 그저 성능을  
올리기 위한 비교 실험 같다

논문을 제대로 읽어보지 않은 느낌



모델과 데이터의 특성,  
task에 대한 이해 부족



작성





## 원인 분석

---

Metric에 대한  
이해 부족

오로지 F1만 보고 달려옴

논문에 대한  
이해 부족

좋은 성능을 낸 논문을 생각 없이  
따라서 구현만 함

Github으로  
협업하기  
<절망편>

(2주간의 실험은... 🤖)

Special Token 실험 코드가 이전 버전으로  
덮어 씌어진 것을 발견

미리 겁먹고  
시도해보지도 않기

Backtranslation의 entity 손상 가능성이  
두려워 시도조차 해보지 않음

하이퍼파라미터  
고려 부족

이미지 대회에서의 회고

모델에 대한  
고려 부족

빠른 실험을 위해 bert-base만 사용

---

## 원인 분석

- bert-base 모델만 활용하여 실험 (빠르게 실험 하려는 목적)  
⇒ roberta-large 모델에서도 실험
- python lightning AMP를 활용하신게 기억나서  
huggingface에도 비슷한 기능이 없나 찾아보게 되었다.  
⇒ fp16을 알게 되었다. (공식 문서의 중요성을 한번더 느낌)
- KLUE 논문에선 RE task를 할 때 max\_length = 128로 주는 것 발견
- roberta-large + fp16 + max\_length = 128 + 3 epoch  
⇒ 실험 시간 18분!



---

# 5일간의 버닝

---

# 5일간의 버닝

## 하이퍼파라미터 튜닝 (+ focal loss)

### 5.2 Fine-Tuning Configurations

For all the experiments, we use Huggingface Transformers [139] and PyTorch-Lightning.<sup>62</sup> We use AdamW optimizer [83] with the learning rate selected from  $\{10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$ , the warm-up ratio from  $\{0., 0.1, 0.2, 0.6\}$  and the weight decay coefficient from  $\{0.0, 0.01\}$ . We choose the batch size from  $\{8, 16, 32\}$  and the number of epochs from  $\{3, 4, 5, 10\}$ . We use the maximum sequence length of 512 for KLUE-MRC and WoS, and 128 for all the other tasks. We report the score obtained from the best hyperparameter configuration based on the dev set performance.



F1: 69.5652  
Auprc: 70.9330

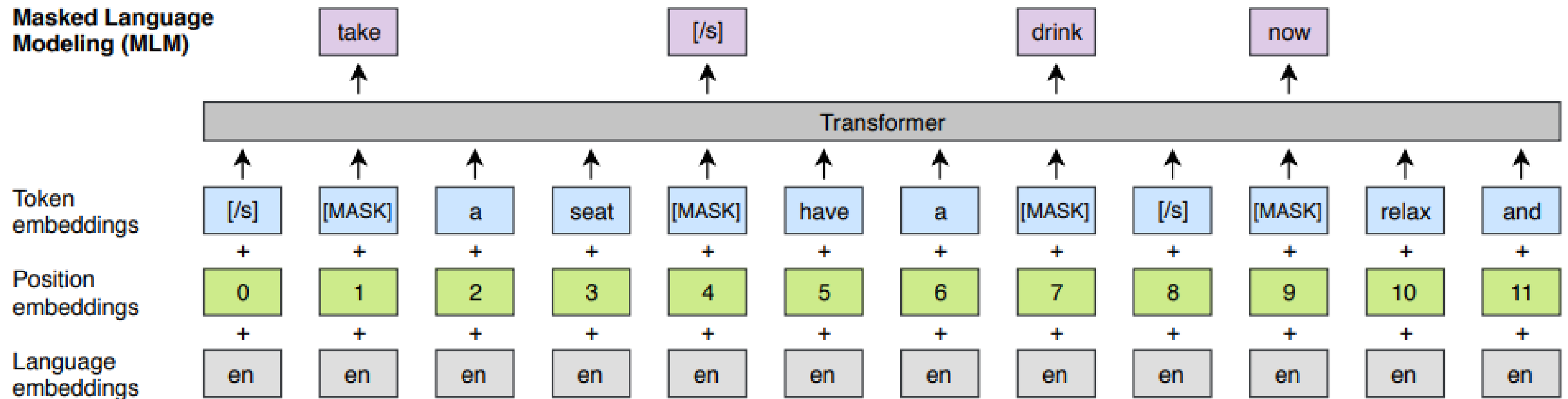


F1: 73.5558  
Auprc: 75.9027

1) KLUE: Korean Language Understanding Evaluation

## 5일간의 버닝

### Multilingual Model 사용 (XLM RoBERTa large)





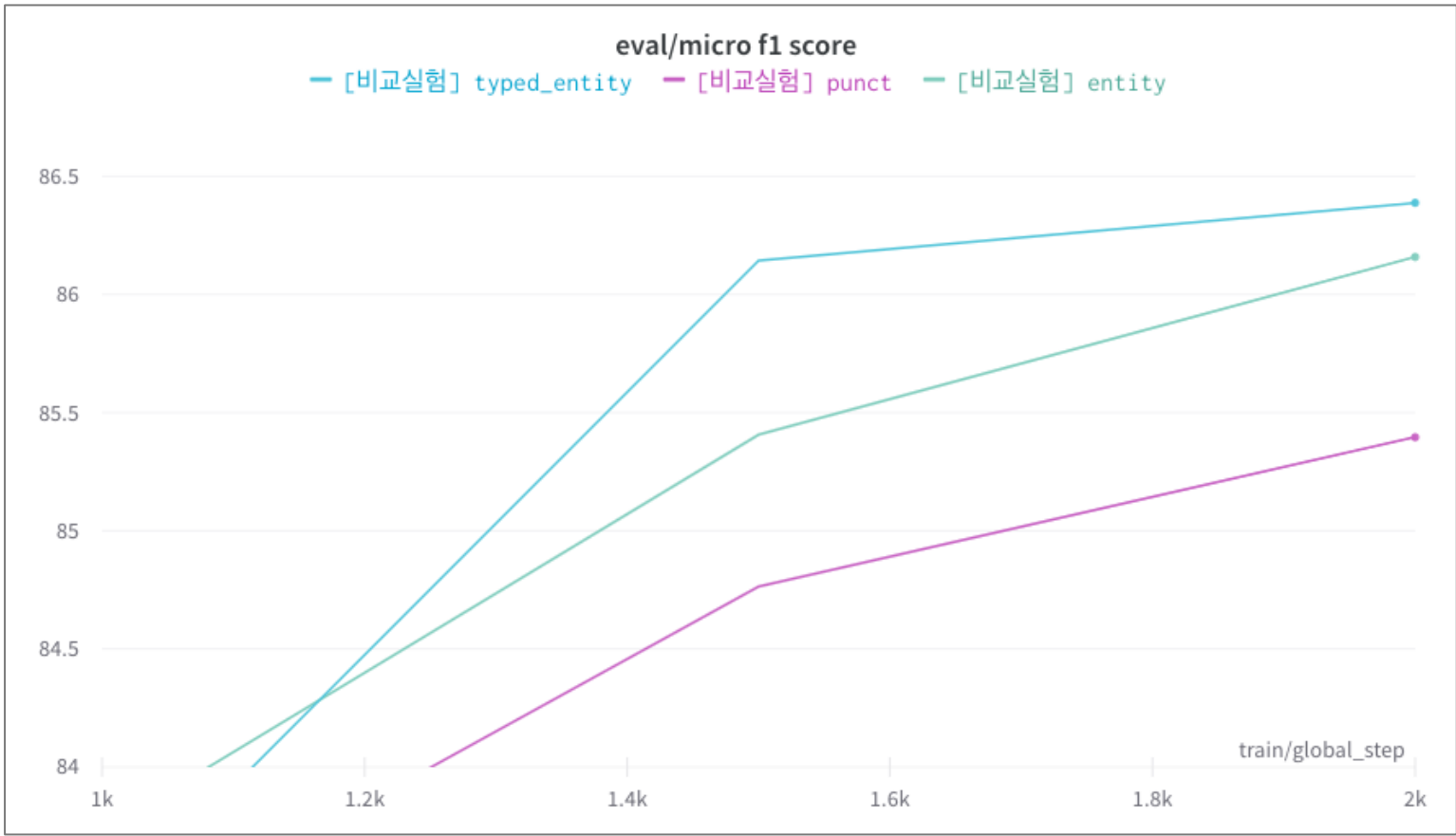
# 5일간의 버닝

## Special Token 적극 도입

Method	Input Example	BERT <sub>BASE</sub>	BERT <sub>LARGE</sub>	RoBERTa <sub>LARGE</sub>
Entity mask	[SUBJ-PERSON] was born in [OBJ-CITY].	69.6	70.6	60.9
Entity marker	[E1] Bill [/E1] was born in [E2] Seattle [/E2].	68.4	69.7	70.7
Entity marker (punct)	@ Bill @ was born in # Seattle #.	68.7	69.8	71.4
Typed entity marker	<S:PERSON> Bill </S:PERSON> was born in <O:CITY> Seattle </O:CITY>.	71.5	72.9	71.0
Typed entity marker (punct)	@ * person * Bill @ was born in # ^ city ^ Seattle #.	70.9	72.7	74.6

Table 3: Test  $F_1$  (in %) of different entity representation techniques on TACRED. For each technique, we also provide the processed input of an example text “Bill was born in Seattle”. Typed entity markers (original and punct) significantly outperforms others.

- 실험(Roberta-Large)
  - punctuation: [CLS] sub [SEP] obj [SEP] a b c @ sub @ d e # obj # f g [SEP]
  - entity token: [CLS] sub [SEP] obj [SEP] a b c <S:NER\_TYPE> sub </S:NER\_TYPE> d e <O:POS\_TYPE> obj </O:POS\_TYPE> f g [SEP]
  - typed entity marker(punct): [CLS] sub [SEP] obj [SEP] a b c @ \* SUB\_NER \* sub @ d e # ^ OBJ\_NER ^obj # f g [SEP]



### ● 분석

- Entity를 함께 주었을 때 성능이 좋았음
  - Entity Type이 관계를 이해하는데 도움을 준다.
- Entity와 Type을 한꺼번에 주는게 아니라 각각 위치를 따로 표시해주는 것이 성능 향상 면에 있어서 더 좋았다.
- 영어로 된 entity type이 더 성능이 좋음

# 5일간의 버닝

## Entity Marker Embedding

```
def get_entity_position_embedding(tokenizer, input_ids):
    special_token2id = {k:v for k,v in zip(tokenizer.all_special_tokens, tokenizer.all_special_ids)}

    sub_token_id = special_token2id['@'] # 36
    obj_token_id = special_token2id['#'] # 7

    pos_embeddings = []

    for y in input_ids:
        ss_embedding = []
        os_embedding = []
        for j in range(0, len(y)):
            if len(ss_embedding) + len(os_embedding) == 4:
                break
            if y[j] == sub_token_id:
                ss_embedding.append(j)
            if y[j] == obj_token_id:
                os_embedding.append(j)

        pos = ss_embedding + os_embedding

        pos_embeddings.append(pos)

    return torch.tensor(pos_embeddings, dtype=torch.int)
```

```
def forward(
    self,
    input_ids=None,
    attention_mask=None,
    token_type_ids=None,
    entity_position_embedding = None,
    position_ids=None,
    head_mask=None,
    inputs_embeds=None,
    labels=None,
    output_attentions=None,
    output_hidden_states=None,
    return_dict=None,
):
    return_dict = return_dict if return_dict is not None else self.config.use_return_dict

    outputs = self.model(
        input_ids,
        attention_mask=attention_mask,
        token_type_ids=token_type_ids,
        position_ids=position_ids,
    )
    pooled_output = outputs[0]

    idx = torch.arange(input_ids.size(0)).to(input_ids.device)
    entity_position_embedding = entity_position_embedding.T
    ss_emb = pooled_output[idx, entity_position_embedding[0].tolist()]
    se_emb = pooled_output[idx, entity_position_embedding[1].tolist()]
    os_emb = pooled_output[idx, entity_position_embedding[2].tolist()]
    oe_emb = pooled_output[idx, entity_position_embedding[3].tolist()]

    h = torch.cat((
        ss_emb,
        se_emb,
        os_emb,
        oe_emb
    ), dim=-1).to(input_ids.device)

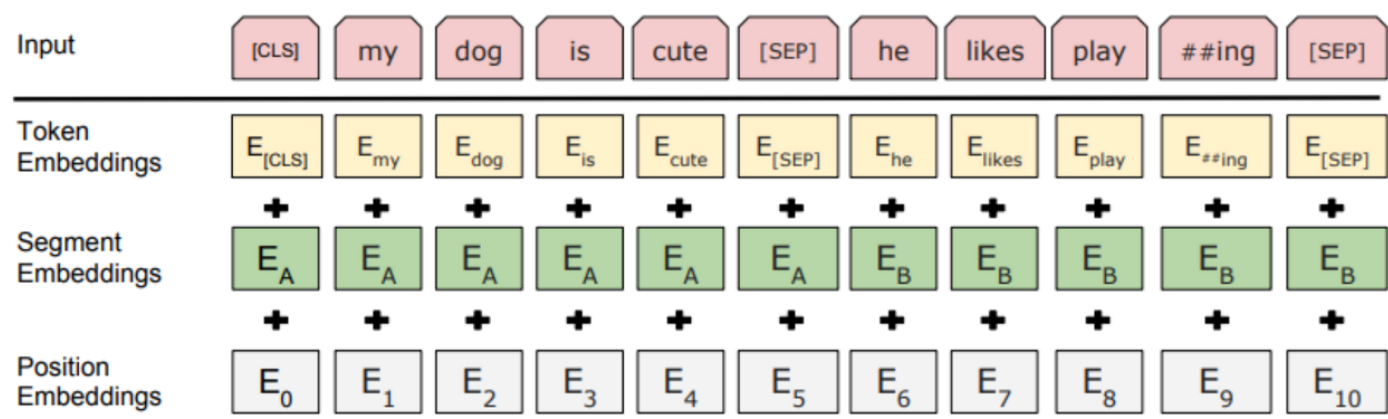
    logits = self.classifier(h)
    outputs = (logits,)
    if labels is not None:
        loss = self.loss_fn(logits.float(), labels)
        outputs = (loss, ) + outputs
    return outputs
```

- 1) [https://github.com/huggingface/transformers/blob/v4.17.0/src/transformers/models/bert/modeling\\_bert.py](https://github.com/huggingface/transformers/blob/v4.17.0/src/transformers/models/bert/modeling_bert.py)
- 2) [https://github.com/wzhouad/RE\\_improved\\_baseline/blob/main/model.py](https://github.com/wzhouad/RE_improved_baseline/blob/main/model.py)

# 5일간의 버닝

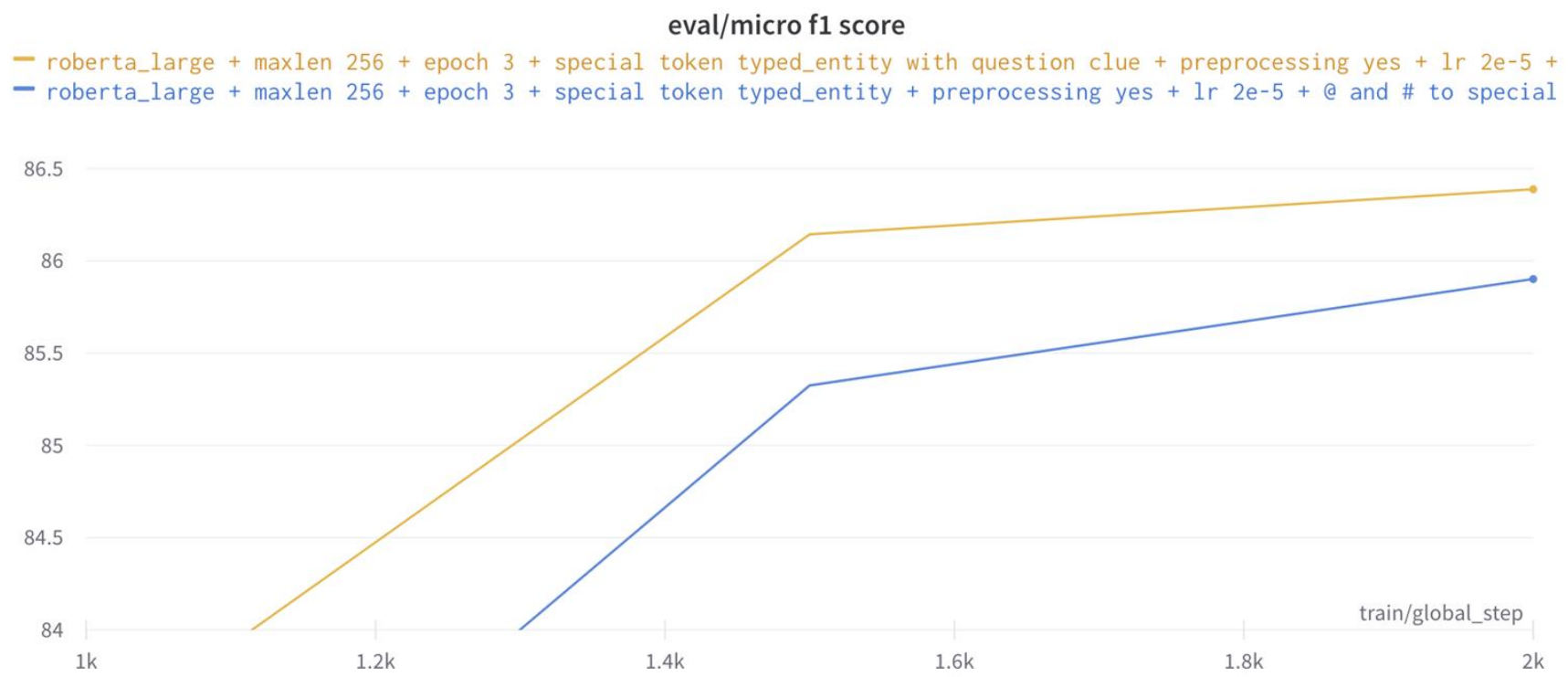
## 질문형 문장

- 첫주차 오피스 아워에서 힌트를 얻어 실험(Single vs Multi)
  - 두가지 문장을 함께 입력 받는 BERT의 구조를 그대로 가져갈 수 있도록 raw sentence 뒤에 문장을 하나 더 붙여주면 어떨까?



[CLS] Typed Entity Marker(punct)가 부착된 Raw Sentence [SEP] 질문형 문장 [SEP]

넣어주는 경우 F1 score 상승



# 5일간의 버닝

---

## 질문형 문장

- 그렇다면 질문으로 들어가는 내용도 중요하게 작용할까?

- 한글 질문
  - 이 문장에서 sub와(과) obj 은(는) 어떤 관계일까?
  - 에서 sub 와(과) obj 은(는)?
  - 에서 sub 와(과) obj의 관계는?
- 영어 질문
  - In this sentence, what is relationship between sub and obj?
  - In this sentence, sub and obj are?

- 결론

- 문장(sentence), 관계(relationship)과 같이 의미가 있는 단어가 함께 들어가는 경우 일반화 부분에서 성능 하락
  - i. 즉, raw sentence에 영향을 주지 않은 불용어 위주로 구성된 문장일수록 일반화가 더 잘 됨
- 한글 데이터셋으로 학습된 모델이다 보니 영어보다는 한글 질문의 성능이 더 좋음.

# 5일간의 버닝

---

## Data Augmentation

- googletrans 를 이용
- 한글 => 영어 => 한글 번역
- Ver 1: Entity를 번역 전에 특수문자로 치환, 최종 번역물 단계에서 복원
- Ver 2: 원본 문장 그대로 이중 번역, 이후 Entity가 온전히 살아남은 문장만 픽업
- 결과물을 원본 데이터와 비교하여 변경점이 없는 경우 제외

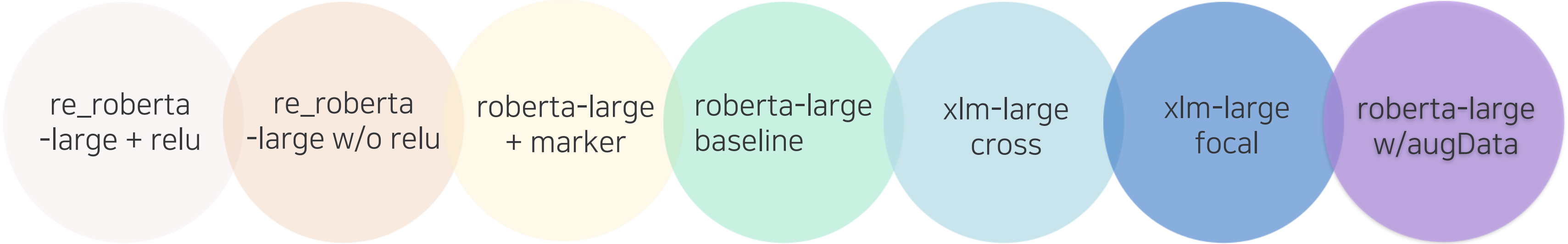
그의 외삼촌은 루이 13세이고 그녀의 남편 루이 14세는 외삼촌이 됩니다.  
기아 타이거즈 외야수 이창진이 롯데백화점 광주점에서 수여하는 9월 월간 MVP에 선...  
법포는 다시 최시형, 서병학, 손병희의 직통인 북부지부와 서장옥, 전봉준, 김가남.  
국토교통부가 실시한 '2019 교통문화지수 조사'에서 완도군(신우철 도지사)이 교통...  
중앙일보, JTBC 회장을 거쳐 중앙홀딩스 회장, 한반도평화재단 이사장, 고려기원 ...

더불어민주당은 7일 오전 9시부터 오후 5시까지 열린 원내대표·정책위원장 후보자 등...  
법포는 다시 직계 후손인 최시형·서병학·손병희·북쪽과 서장옥, 전봉준, 김가남을 지도자...  
국토교통부가 실시한 '2019년 교통문화지수 조사'에서 완도군(군수신우철군)이 A등...  
JTBC 중앙일보 회장을 거쳐 중앙홀딩스 회장, 한반도평화재단 이사장, 기원 회장을...  
화순군(구청장구충곤)은 17일 등면장 20여명이 코로나 예방을 위해 오동리 천운아파...



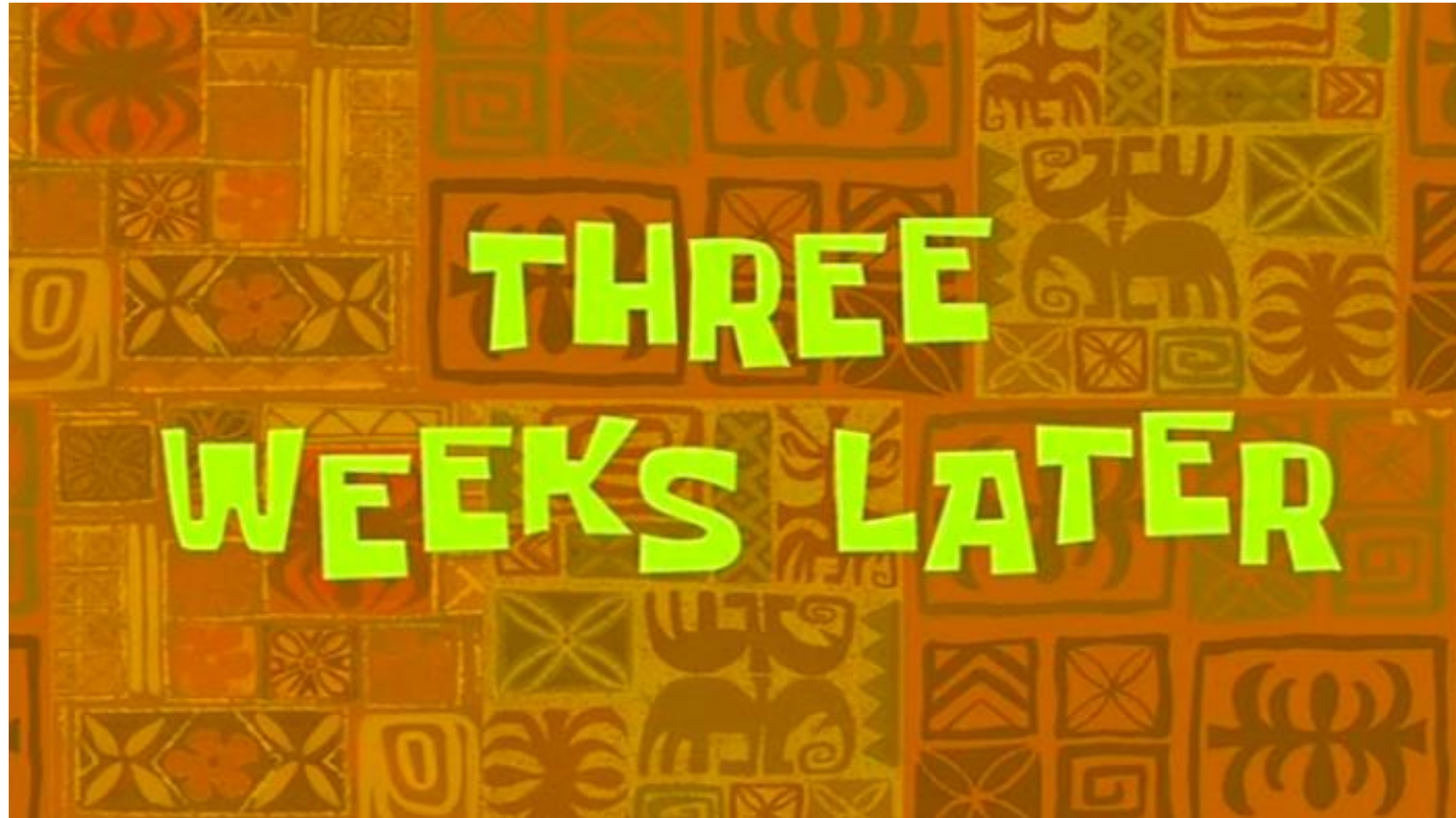
# ENSEMBLE

---



- |                           |                           |                           |                    |                    |                    |                    |
|---------------------------|---------------------------|---------------------------|--------------------|--------------------|--------------------|--------------------|
| ● original dataset        | ● 중복 제거                   | ● 중복 제거                   | ● original dataset | ● original dataset | ● original dataset | ● 중복제거 + backtrans |
| ● 전처리 X                   | ● 전처리 O                   | ● 전처리 O                   | ● 전처리 X            | ● 전처리 X            | ● 전처리 X            | ● 전처리 X            |
| ● entity marker embedding | ● entity marker embedding | ● entity marker embedding | ● Baseline 세팅      | ● Baseline 세팅      | ● Focal loss       | ● Focal loss       |
| ● Baseline 세팅             | ● 한글 질문(뒤)                | ● 한글 질문(뒤)                | ● Baseline 세팅      | ● Baseline 세팅      | ● Baseline 세팅      | ● 한글 질문(앞)         |
-

그 결과..



2

NLP\_07조



75.0999

83.5214

# SUMMARY

날짜	F1/micro	AUPRC	비고
22-04-02	69.9725 → 67.2611	71.6246 → 72.0717	klue/roberta-large+TAPT
22-04-03	73.5558 → 71.4286	75.9027 → 76.9896	LR tuning
22-04-04	74.0670 → 72.6614	77.8840 → 79.9564	Typed entity(punct) with question clue
22-04-05	75.5174 → 73.5146	81.0755 → 80.8328	Top-3 ensemble
22-04-07	77.0460 → 75.0999	82.4147 → 83.5214	Top-7 ensemble



# RETROSPECTION

## 회고

- + 체계적인 실험 기반은 실수 극복에도 도움된다
- + Metric에 대한 이해
  - + Loss, F1, Auprc 등을 종합적으로 보자
- + 일반화의 방향성에 대해 생각해보자



## + 하이퍼 파라미터란 무엇인가

- + 최고를 찾으려는 것보다, 상황 마다 최선의 선택을 하려 노력하자
- + 내가 쓰는 모델, 데이터, 작업의 특성을 충분히 이해하자
- + 하이퍼 파라미터별 특성을 이해하자
  - > 하이퍼 파라미터 튜닝에 쏟는 시간이 크게 줄어듦



# *THANK YOU!*

KLUE : 문장 내 개체간 관계 추출 대회  
SOLUTION 발표

7조 김준재 김현아 배현진 이강민 최성원