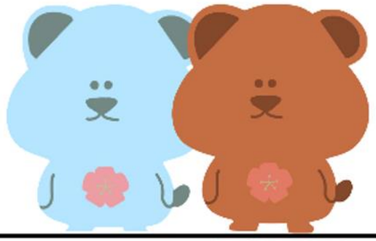
한국어 문장 관계 분류 경진대회

한예송 홍재령

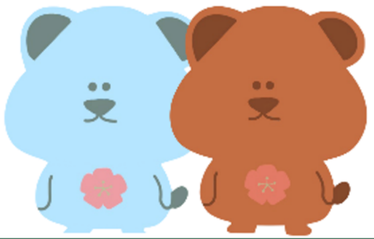
EWHA EURON

목차



EWHA EURON

대회 소개 & EDA



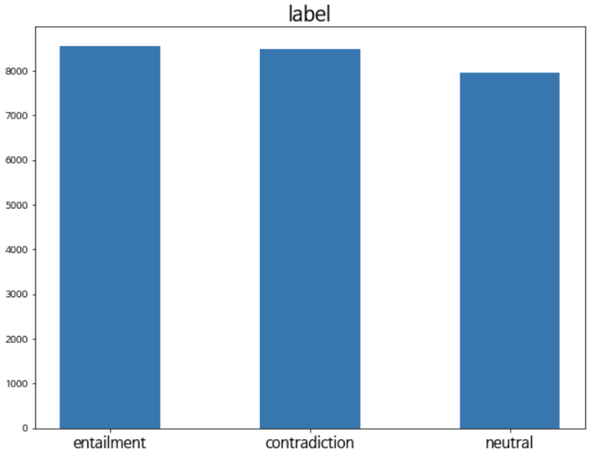
EWHA EURON

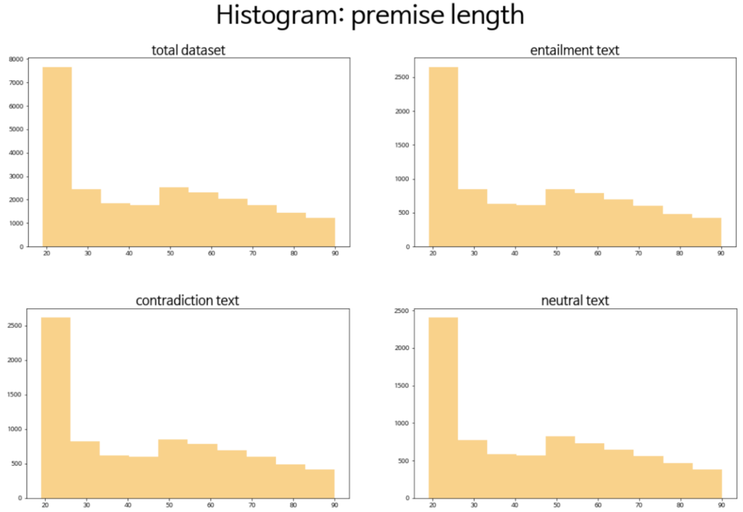
#01-1 대회 소개 & EDA

Hypothesis 문장이 참인지, 거짓인지 판별

EDA

	index	premise	hypothesis	label
1	0	씨름은 상고시대로부터 전해져 내려오는 남자들의 대표적인 놀이로서, 소년이나 장정들이 넓고 평평...	씨름의 여자들의 놀이이다.	contradiction
2	1	삼성은 자작극을 벌인 2명에게 형사 고소 등의 법적 대응을 검토 중이라고 하였으나, 중국 내에서의 ...	자작극을 벌인 이는 3명이다.	contradiction
3	2	이를 위해 예측적 범죄예방 시스템을 구축하고 고도화한다.	예측적 범죄예방 시스템 구축하고 고도화하는 것은 목적이 있기 때문이다.	entailment
4	3	광주광역시가 재개발 정비사업 원주민들에 대한 종합대책을 마련하는 등 원주민 보호에 적극 나섰다.	원주민들은 종합대책에 만족했다.	neutral
5	4	진정 소비자와 직원들에게 사랑 받는 기업으로 오래 지속되고 싶으면, 이런 상황에서는 책임 있는 모...	이런 상황에서 책임 있는 모습을 보여주는 기업은 아주 드물다.	neutral
6	5	이번 증설로 코오롱인더스트리는 기존 생산량 7만7000톤에서 1만6800톤이 늘어나 총 9만 380...	코오롱 인더스트리는 총 9만 3800톤의 생산 능력을 확보했다.	entailment
7	6	자신뿐만 아니라 남을 돕고자 하는 청년의 꿈과 열정에 모두가 주목하고 있다.	모든 청년은 꿈과 열정을 가지고 있다.	neutral
8	7	시대상황을 고려하는 현명한 시청태도가 요구되다.	시청태도에 특별한 주의점은 없다.	contradiction
9	8	사진과 차이없는 아기자기한 실내소품들과 분위기가 멋졌습니다.	아기자기한 실내소품들은 사진에서 본 것과 차이가 있었습니다.	contradiction
10	9	빠른 답장과 간편한 체크인, 깨끗한 집 좋았어요	체크인이 복잡했어요.	contradiction





entailment – 참
contradiction – 거짓
neutral – 참/거짓 여부 알 수 없음

워드 클라우드 시각화

true

false

neutral

1등 솔루션

EWHA EURON

#01-2 1등 솔루션

다음 두 개의 모델 결과값을 softvoting ensemble하여 제출

(1) Tunib's KoElectra-base finetuned with Arcface Head

(2) KLUE Roberta-
large finetuned with sentence pooling embeddings and special token embeddingsnts

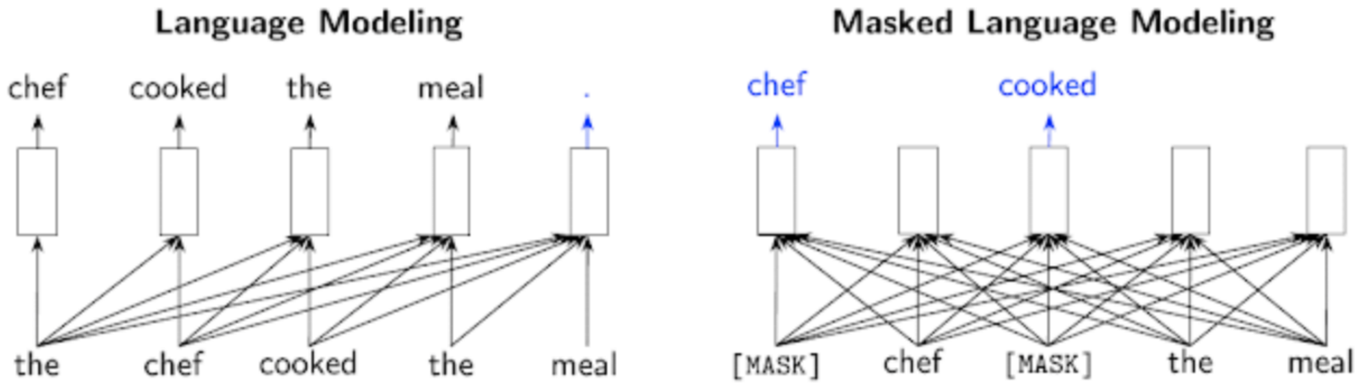
#01-3 기존의 NLP pre-training

LM(Language Modeling)

- GPT
- directional(왼쪽 -> 오른쪽)

MLM(Masked Language Modeling)

- BERT, RoBERTa, ALBERT
- input을 벗어나는 작은 양의 단어에대해 마스킹처리하여 identities를 예측
- bidirectional(왼쪽, 오른쪽)
- 모든 input token에 대해 학습 X



Existing pre-training methods and their disadvantages. Arrows indicate which tokens are used to produce a given output representation (rectangle). **Left:** Traditional language models (e.g., GPT) only use context to the left of the current word. **Right:** Masked language models (e.g., BERT) use context from both the left and right, but predict only a small subset of words for each input.

#01-4 ELECTRA(1)

- Efficiently Learning an Encoder that Classifies Token – Replacements Accurately
- 기존의 RoBERTa나 XLNet과 비슷한 성능, 훨씬 효율적

  ->language understanding에 대해 RoBERTa와 XLNet의 $\frac{1}{4}$만큼의 연산으로 비슷한 성능
- 1등 팀) ELECTRA, RoBERTa-large 모델 앙상블 -> 정확도 각각 0.896, 0.902
- -> ELECTRA 모델이 RoBERTa-large 모델보다 파라미터 수가 1/3정도의 작은 규모, attention layer수로 따지면 절반 정도 깊이인 모델, 비슷한 정확도

- 34GB의 한국어 text로 학습한 KoELECTRA 모델도 존재

#01-5 ELECTRA(2)

Generator
- MLM과 비슷
- Discriminator의 input을 만들기 위한 모델

Discriminator
- ELCTRA에 해당
- 전체 토큰들이 처음의 입력 토큰과 일치하는지 (이진)분류하는 모델

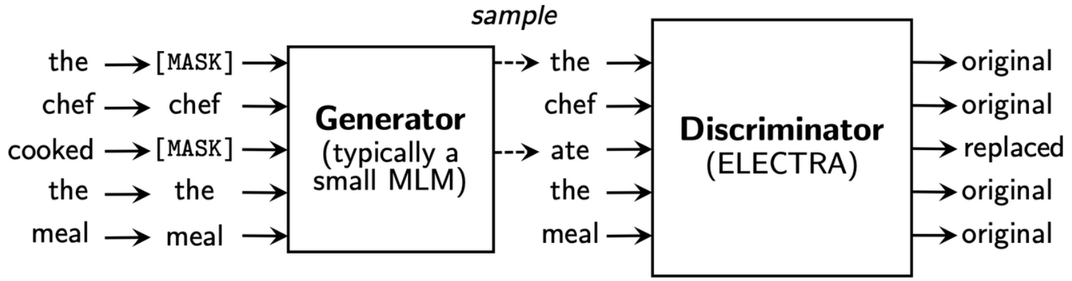


Figure 2: An overview of replaced token detection. The generator can be any model that produces an output distribution over tokens, but we usually use a small masked language model that is trained jointly with the discriminator. Although the models are structured like in a GAN, we train the generator with maximum likelihood rather than adversarially due to the difficulty of applying GANs to text. After pre-training, we throw out the generator and only fine-tune the discriminator (the ELECTRA model) on downstream tasks.

Loss: jointly training O / adversarial training X

$$\mathcal{L}_{\text{MLM}}(\boldsymbol{x}, \theta_G) = \mathbb{E}\left(\sum_{i \in \boldsymbol{m}} -\log p_G(x_i | \boldsymbol{x}^{\text{masked}})\right)$$

$$\mathcal{L}_{\text{Disc}}(\boldsymbol{x}, \theta_D) = \mathbb{E}\left(\sum_{t=1}^{n} -\mathbb{1}(x_t^{\text{corrupt}} = x_t) \log D(\boldsymbol{x}^{\text{corrupt}}, t) - \mathbb{1}(x_t^{\text{corrupt}} \neq x_t) \log(1 - D(\boldsymbol{x}^{\text{corrupt}}, t))\right)$$

$$\min_{\theta_G, \theta_D} \sum_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}_{\text{MLM}}(\boldsymbol{x}, \theta_G) + \lambda \mathcal{L}_{\text{Disc}}(\boldsymbol{x}, \theta_D)$$

#01-6 ArcFace loss

기존의 loss: 모델 예측 값(class)과 positive class 유사도 고려 -> 학습

Metric learning loss :
- 모델 예측 값(class)과 positive class 유사도, negative class 유사도 고려
  -> 학습
- Euclidean-distance-based, angular-margin-based

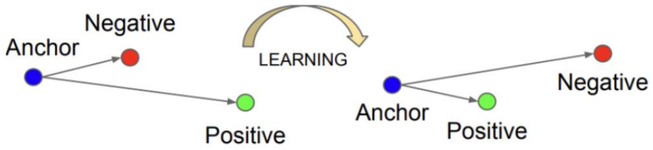
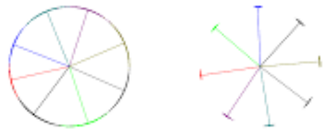


Figure 3. The **Triplet Loss** minimizes the distance between an *anchor* and a *positive*, both of which have the same identity, and maximizes the distance between the *anchor* and a *negative* of a different identity.

ArcFace loss:
- Angular 방법으로 distance 계산
- $\cos(\theta_1 + m) - \cos\theta_2$ => margin과 각도 유사성 ↑



(a) Softmax (b) ArcFace

Figure 3. Toy examples under the softmax and ArcFace loss on 8 identities with 2D features. Dots indicate samples and lines refer to the centre direction of each identity. Based on the feature normalisation, all face features are pushed to the arc space with a fixed radius. The geodesic distance gap between closest classes becomes evident as the additive angular margin penalty is incorporated.

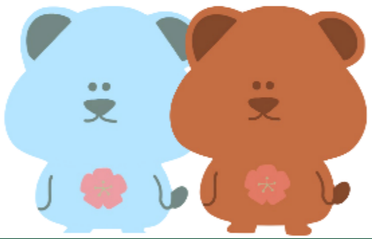
TABLE V

DECISION BOUNDARIES FOR CLASS 1 UNDER BINARY CLASSIFICATION CASE, WHERE $\hat{x}$ IS THE NORMALIZED FEATURE. [106]

Loss Functions	Decision Boundaries
Softmax	$(W_1 - W_2)\,x + b_1 - b_2 = 0$
L-Softmax [104]	$\|x\|\,(\|W_1\|\,cos(m\theta_1) - \|W_2\|\,cos(\theta_2)) > 0$
A-Softmax [84]	$\|x\|\,(cosm\theta_1 - cos\theta_2) = 0$
CosFace [105]	$\hat{x}\,(cos\theta_1 - m - cos\theta_2) = 0$
ArcFace [106]	$\hat{x}\,(cos\,(\theta_1 + m) - cos\theta_2) = 0$

RoBERTa

(1) RoBERTa란
(2) RoBERTa 모델을 활용한
    1등 노트북
(3) RoBERTa 모델을 활용한
    2등 노트북



EWHA EURON

#3-1 RoBERTa

Robustly Optimized BERT Pretraining Approach

BERT를 개선하기 위해 복제 연구를 진행

BERT에
1. 모델을 더 오래, 더 큰 배치로, 더 많은 데이터로 훈련시킴
2. Next sentence prediction objective 제거
3. Longer sequence를 넣어줌
4. Dynamic masking 적용
5. 새로운 큰 데이터셋 수집 (CC-NEWS)

- BERT & BERT 이후의 후속 모델보다 우월한 성능을 보임
- 이전에 간과되었던 설계의 중요성을 대두시킴

RoBERTa의 특징

- 160GB의 데이터
- Dynamic masking
- MLM만으로 pre-train
- Full-sentence 형식의 input
- BERT의 약 32배의 batch size
- Byte-level BPE tokenizer

① 더 큰 batch size로 좀 더 많은 데이터를 학습시킨 모델
② Dynamic Masking
③ NSP 방법 사용 X (=MLM만으로 pre-train)

EWHA EURON

#3-1 RoBERTa의 특성

1) 더 큰 batch size로 좀 더 많은 데이터를 학습시킨 모델

Batch Size
Batch size의 영향력을 확인하기 위한 실험 진행
→ 전체 step 수가 유지되도록 batch size & epoch 수 조정

∴ 같은 step 수더라도 batch size가 클수록 성능이 좋았음

bsz	steps	lr	ppl	MNLI-m	SST-2
256	1M	1e-4	3.99	84.7	92.7
2K	125K	7e-4	**3.68**	**85.2**	**92.9**
8K	31K	1e-3	3.77	84.6	92.8

Data
총 160GB ⇔ 기존모델(BERT) 16GB
데이터 크기가 클수록 성능이 좋아짐 → RoBERTa는 최대한 데이터를 많이 모으는 것에 집중함
➔ BookCorpus, EnglishWikipedia, CC-News, OpenWebText, Stories 총 5개의 데이터셋을 합침

+) 학습 시간을 길게 할수록 성능이 올라감

#3-1 RoBERTa의 특성

2) Dynamic Masking

Masking	SQuAD 2.0	MNLI-m	SST-2
reference	76.3	84.3	92.8
Our reimplementation:			
static	78.3	84.3	92.5
dynamic	78.7	84.0	92.9

Static masking

- 기존의 BERT가 pre-training에 사용한 MLM
- 학습을 시작하기 전에 무작위로 token에 mask를 씌워, 그것을 예측하는 방식
  → 매 학습 단계에서 똑같은 mask를 보게 됨
  ➔ 해결 방법 : 같은 단어의 maskin을 피하기 위해 같은 문장을 10번 복사한 뒤 각각의 sequence가 다른 방식으로 masking되도록 설정
  → 같은 마스크를 가진 sequence는 최대 반복횟수의 1/10만큼만 사용됨
  → 크기가 큰 데이터에 대해서는 비효율적 ☹

Dynamic masking

매 epoch마다 mask를 새로 씌움 (마스킹 패턴을 생성하는 방식)
→ 매번 masking을 새로 하여 학습 시간은 늘어나지만 메모리를 아낄 수 있음, 성능도 더 좋음
→ 더 많은 횟수를 반복하거나, 큰 데이터셋을 다루는 데 중요

#3-1 RoBERTa의 특성

3) NSP 방법 사용 X (=MLM만으로 pre-train)

NSP (Next Sentence Prediction)

기존의 BERT : 2개의 문장을 이어붙여 input을 만들고,
두 문장이 문맥상으로 연결된 문장인지 판단하는 NSP를 pre-training에서 사용
→ 엄청 짧은 input도 등장

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6
$\text{BERT}_{\text{BASE}}$	88.5/76.3	84.3	92.8	64.3
$\text{XLNet}_{\text{BASE}}$ (K = 7)	–/81.3	85.8	92.7	66.1
$\text{XLNet}_{\text{BASE}}$ (K = 6)	–/81.0	85.6	93.4	66.7

MLM만으로 pre-training

→ 꼭 두 문장을 이어붙인 형태의 input을 사용할 필요가 없어짐
➔ Token 수가 512를 넘어가지 않는 선에서 문장을 최대한 이어 붙여서 input 만듦
→ 모든 input들의 token 수가 약 512
→ 더 나은 성능

Training Objective

BERT

Masked Language Model (MLM)

Next Sentence Prediction (NSP)

RoBERTa

Masked Language Model (MLM)

~~Next Sentence Prediction (NSP)~~

#3-2 RoBERTa 모델을 활용한 1등 노트북

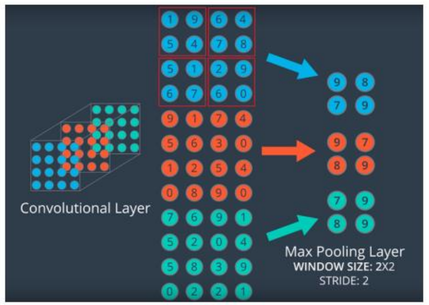
한국어 문장 관계 분류 경진대회

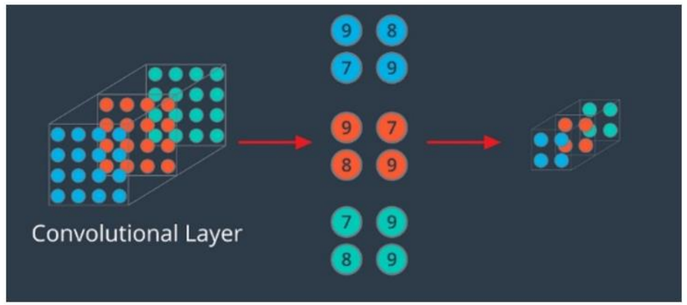
모델 2개 사용, 두 모델 결과값을 Softvoting Ensemble하여 제출
(1) Tunib's KoElectra-base finetuned with Arcface Head (Public LB 0.896)
(2) KLUE Roberta-large finetuned with sentence pooling embeddings and special token embeddings (Public LB 0.902)

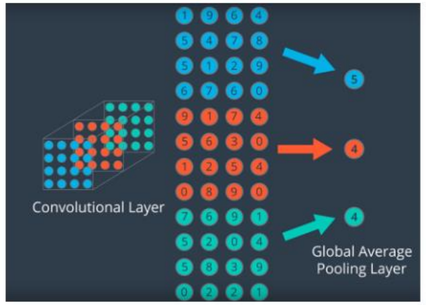
KLUE Roberta-large finetuned with sentence pooling embeddings and special token embeddings

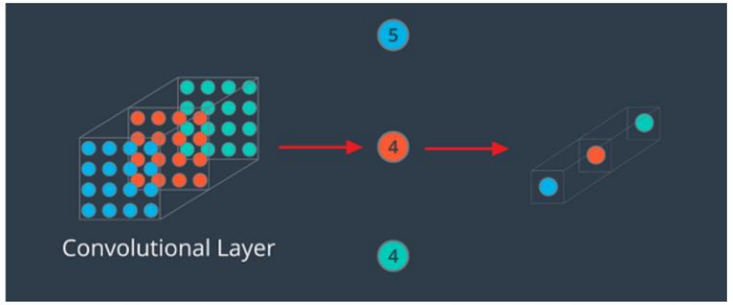
- Roberta를 finetuning 할 때 두 개의 논문을 혼합
- 모델 구조 요약 : 문장의 Global Average Pooling과 Special Token을 모두 사용

- Global Average Pooling : 목적 = feature를 1차원 벡터로 만들기







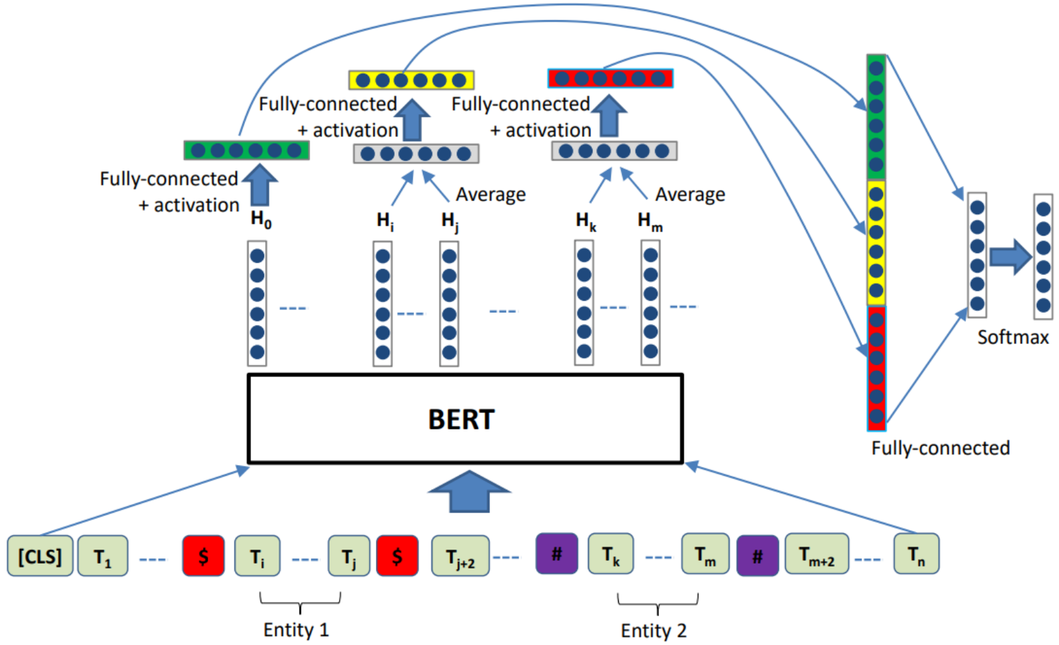


#3-2 RoBERTa 모델을 활용한 1등 노트북

Special Token

텍스트가 BERT를 거치기 전
1) 타겟 엔티티의 양 옆에 스페셜 토큰을 삽입
→ 두 타겟 엔티티의 위치를 잘 포착하여
두 엔티티의 정보를 BERT로 전달

2) BERT를 거치고 난 뒤의 출력되는 임베딩에서의
<u>두 개의 타겟 엔티티의 위치를 찾기</u>



이러한 방식으로 BERT는 relation classification task에
더 적합하도록 문장의 의미와 두 타겟 엔티티를 포착할 수 있습니다!😀

#3-2 RoBERTa 모델을 활용한 2등 노트북

한국어 문장 관계 분류 경진대회

모델 4개 사용, 4 모델 결과값을 Final Hard Voting Ensemble하여 제출
(1) Bart Noise 3-way+BackTrans (Public LB 0.902)
(2) Custom Model R-Roberta (Public LB 0.897)
(3) Randomly Masking Token (Public LB 0.892)
(4) Soft Ensemble 5-fold (Public LB 0.891)

모델의 차별화 전략:
- KLUE Official Dev Data를 학습에 추가 사용
- R-BERT 모델 구조에서 아이디어를 차용하여 모델 아키텍처를 수정
- 5-Fold Soft Ensemble 을 시도하려 했지만 Colab의 런타임 이슈로 인하여 4번째와 5번째 Fold만 학습 완료됨
- 이들 중 Public Score 0.897을 기록한 4번째 Fold 모델만을 Inference에 사용함

#3-2 RoBERTa 모델을 활용한 2등 노트북

R-BERT

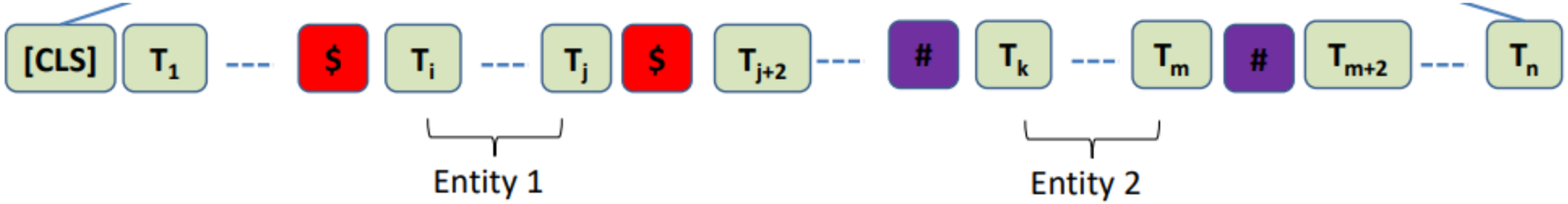
Relation classification task를 위해 고안된 BERT 모델

Relation classification

- 두 엔티티 사이의 관계를 추출하는 중요한 NLP 태스크

Ex. 주어진 시퀀스 s가 있고, 명사쌍 e1과 e2가 있을 때, e1과 e2 사이의 관계를 알아내는 것

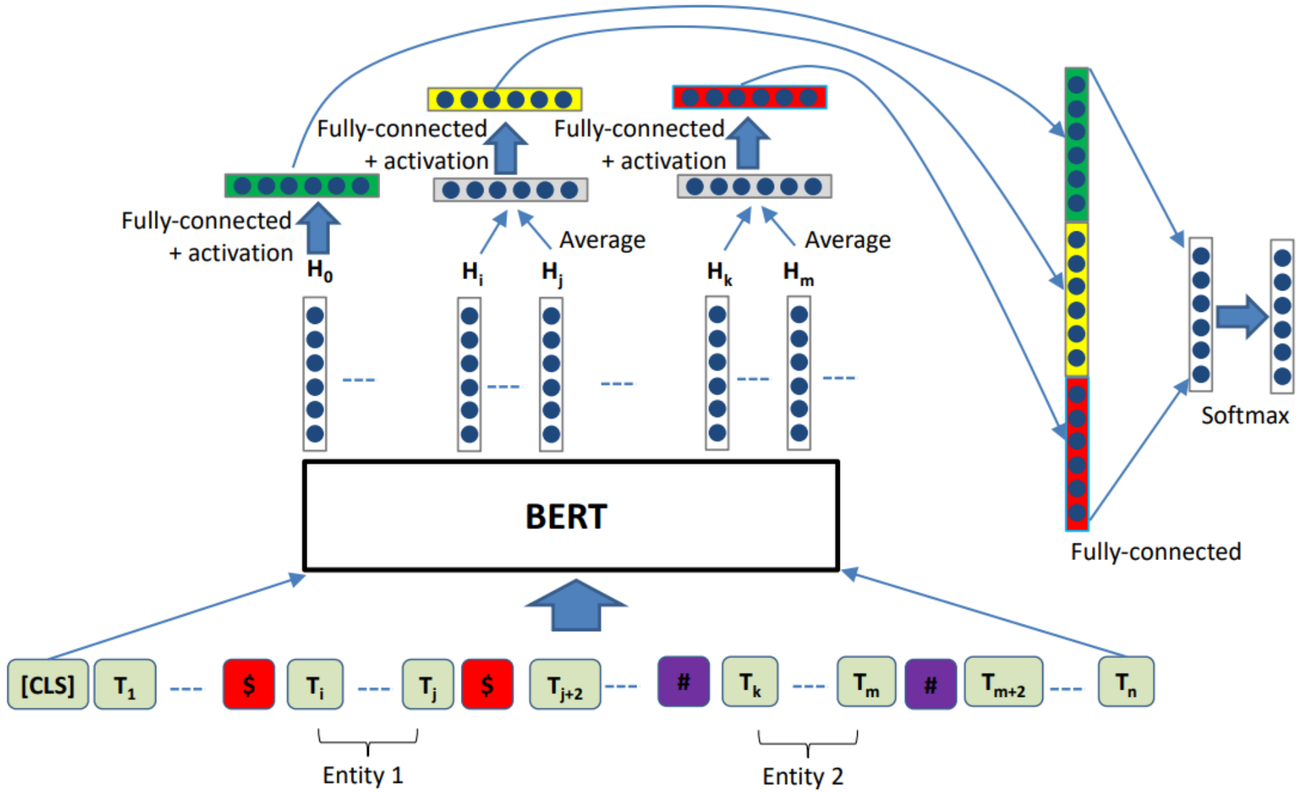
- 수행하기 위해 문장에 대한 정보 & 두 엔티티에 대한 정보 필요
  - **문장에 대한 정보** : BERT의 last hidden states의 출력값 [CLS]토큰에 담겨 있음
  - **두 타겟 엔티티** : 각각 양 옆에 스페셜 토큰 삽입하여 위치 잘 포착 가능



#3-2 RoBERTa 모델을 활용한 2등 노트북

모델 아키텍쳐

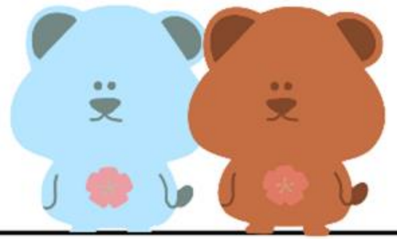
R-BERT 모델로부터 아이디어를 얻어 KLUE Roberta Large 모델의 아키텍쳐를 수정
Relation Extraction Task에서 R-BERT 모델은 CLS 토큰 뿐만 아니라 entity1과 entity2 임베딩 벡터를 같이 활용함으로써 그 성능을 높임



Relation classification 태스크를 해결하기 위해서는 BERT가 두 엔티티의 위치를 잘 포착해야 함
→ 각 토큰의 양 옆에 special token을 추가
→ 문장의 의미를 포착하기 위해 문장의 맨 처음에 [CLS]토큰을 추가

(entity 1에 $, entity 2에 #를 special token으로 사용)
Ex. *[CLS] The $ kitchen $ is the last renovated part of the # house #.

THANK YOU



EWHA EURON