1	NLP_08조		75.7874	83.6859	139	1d

NLP 8조 MIML
Solution 발표

KLUE Relation Extraction

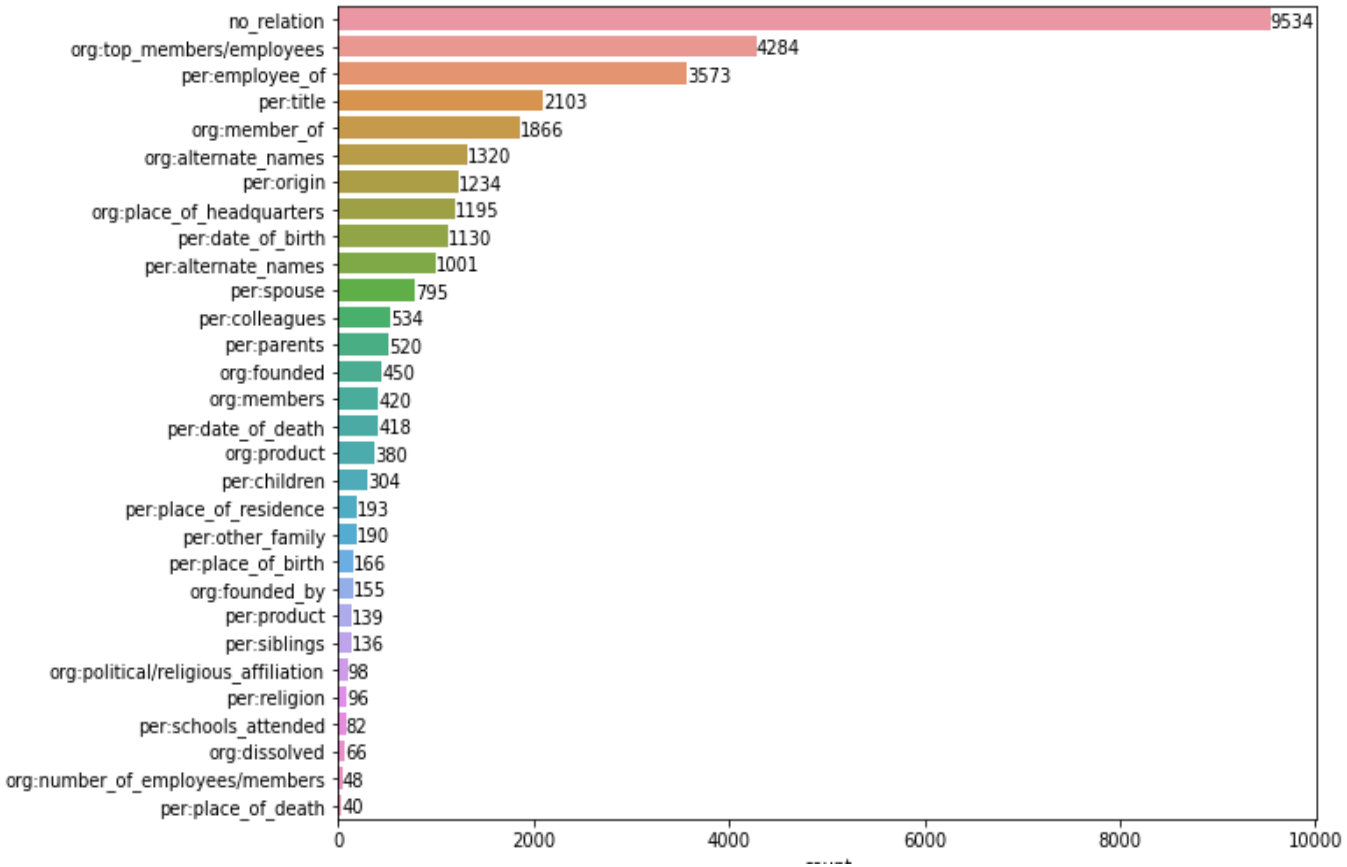
T3180 임동진
T3201 정재윤
T3205 조설아
T3238 허치영
T3246 이보림

Contents

Data

01

EDA(Exploratory Data Analysis)



이전에 진행했던 마스크 착용상태 분류대회때의 데이터보다 훨씬 심각한 Imbalanced Data

01

Data Augmentation

1. EDA(Easy Data Augmentation)

- SR : Synonym Replacement. 특정 단어를 유의어로 교체
- RI : Random Insertion. 임의의 단어를 삽입
- RS : Random Swap. 문장 내 임의의 두 단어의 위치를 바꿈
- RD : Random Deletion. 임의의 단어를 삭제

각 기법을 적용하기만 해도 데이터셋 4배 증강 가능!

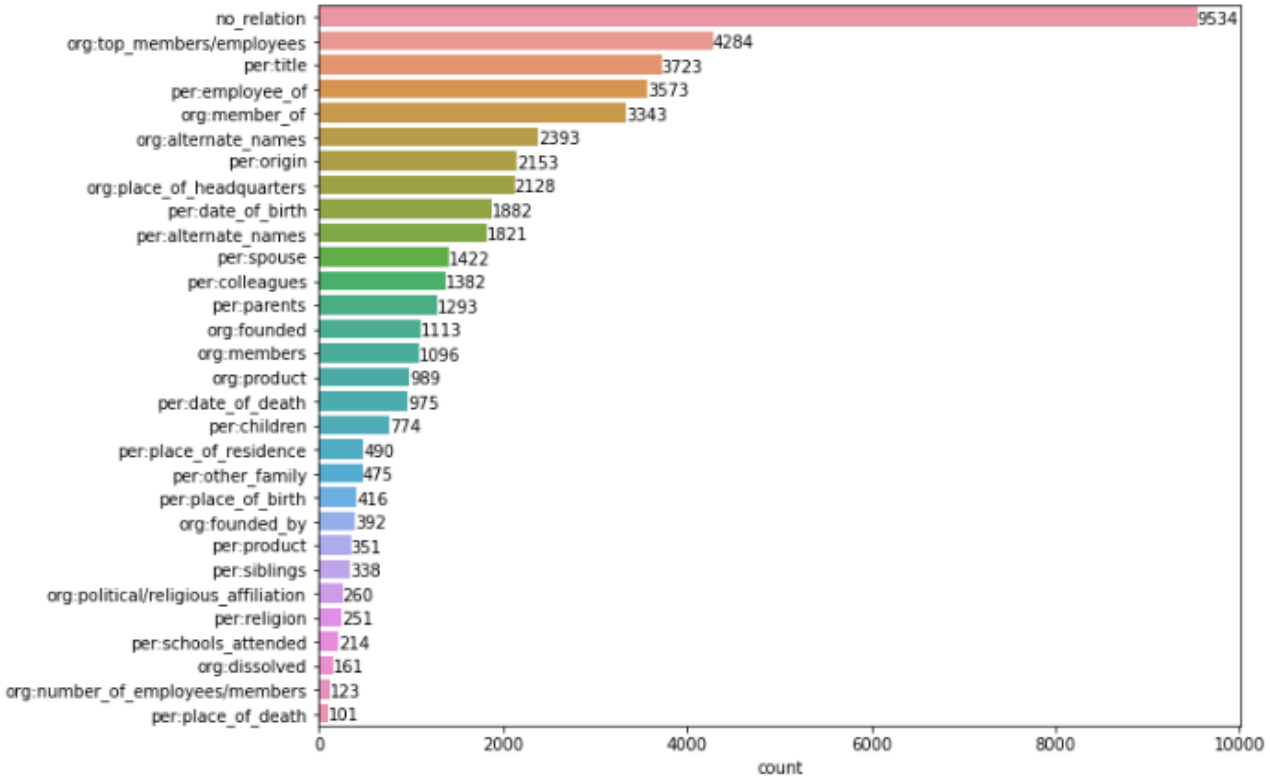
- WordNet의 단어들을 단순히 바꿔 변형을 하기때문에 의미가 변형됨
  - 한국어 특성상 원래 문장의 성질을 따르면서 적절히 변형하기엔 무리

안전하게 데이터를 증강하기 위해 RS, RD 기법만 사용

01

Data Augmentation

1. EDA(Easy Data Augmentaion)



32470 -> 47450

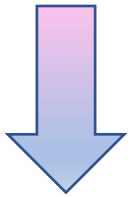
어느정도 데이터 증강이 되었으나 LB점수 하락

01

Data Augmentation

2. Back Translation

만화가 주호민 씨가 포털 사이트 '네이버'에게 받은 20주년 기념 선물을 공개했다.

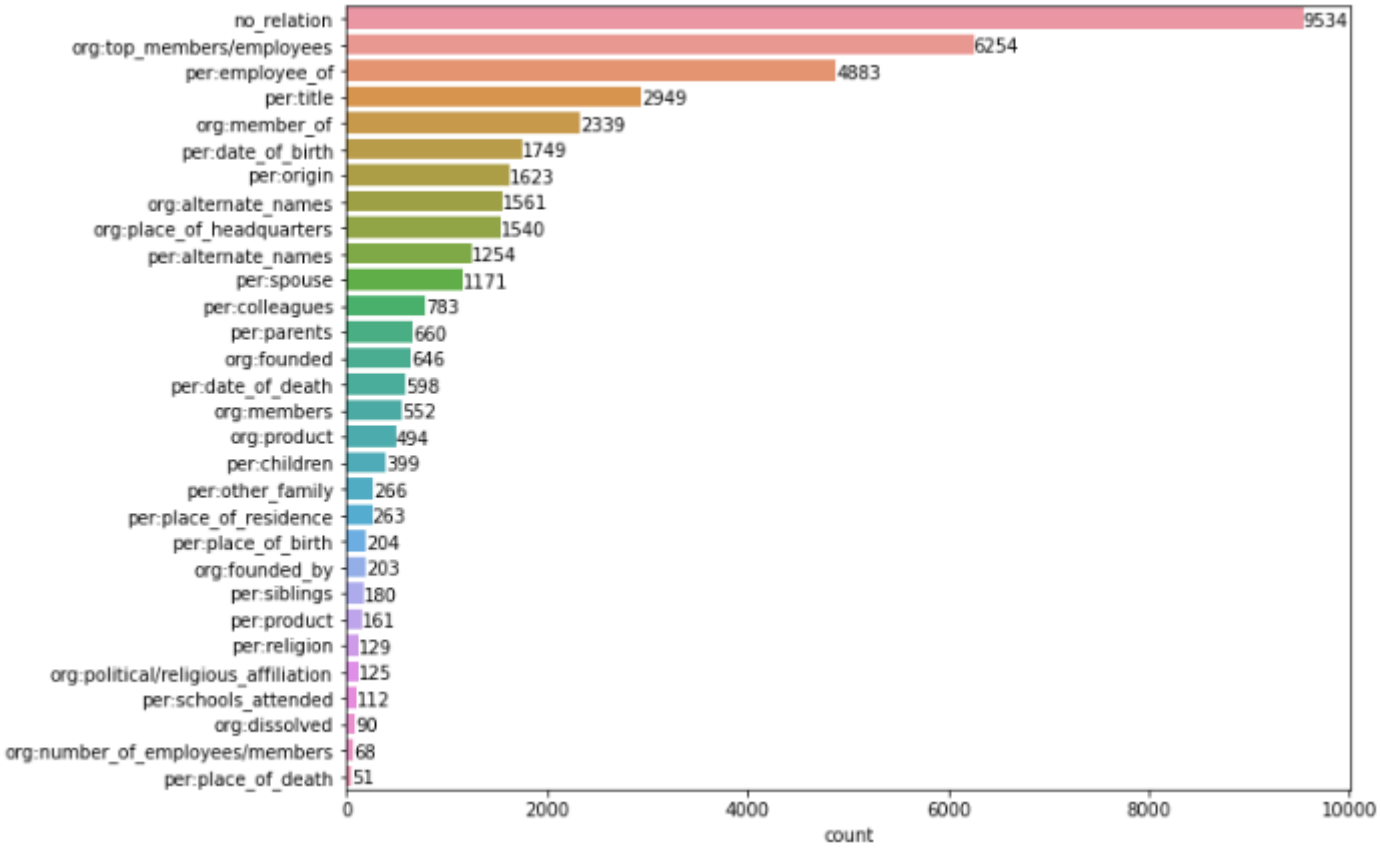


만화가 주호민이 포털사이트 네이버의 20주년 선물을 공개했다.

01

Data Augmentation

2. Back Translation (Ko-En-Ko)



No_relation 제외 증강 역시나 점수 하락(32470->40841)

01

Data Augmentation

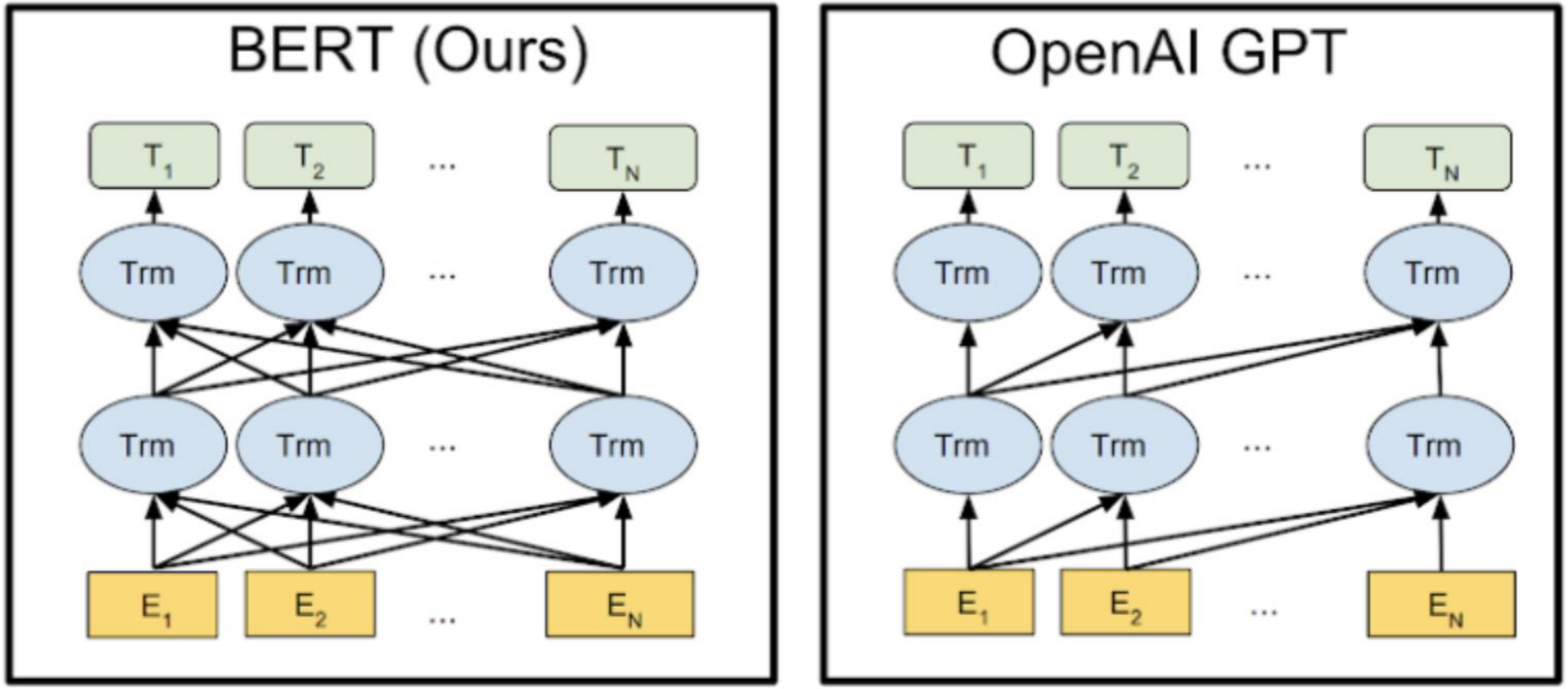
3. 기타 방법들

- Undersampling : 과도하게 많은 no_relation을 undersampling
- Oversampling, Undersampling 동시적용

Model

02

기본 모델을 어떤 것을 활용할 것인가?

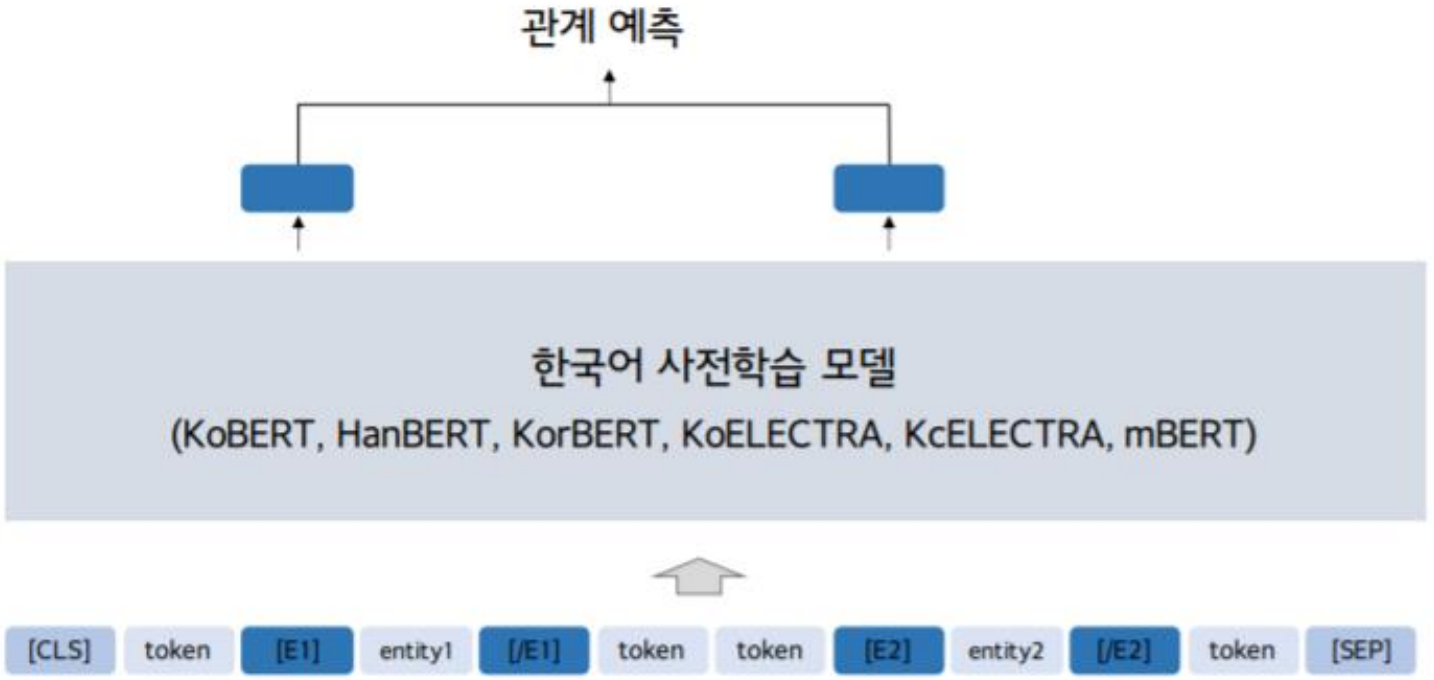


BERT: Pre-training of Deep Bidirectional Transformers for Language Modeling
(Devlin et al. 2018)

02 기본 모델을 어떤 것을 활용할 것인가?

- KoBERT, M-BERT(Multilinguial BERT)
  - 활용 이유 : 한국어 전문 BERT 모델이므로 성능 향상 기대
  - 결과 : 성능 향상 없음
- KLUE/Roberta-Large
  - 활용 이유 : BERT 논문에서 BERT Base보다 Bert Large가 더 성능이 좋다는 언급이 있어 활용
  - 결과 : 성능 매우 좋아짐(63.2541 → 67.7767)
- XLM-RoBERTa-Large
  - 활용 이유 : 한국어 NLU Task에서 활용되는 Large 모델 중 하나여서 혹시 좋아질까 싶은 마음에 시도
  - 결과 : 성능 향상 없음

02 모델의 Classifier를 어떻게 변경시킬 것인가?



출처 : https://www.koreascience.or.kr/article/CFKO202130060668824.pdf

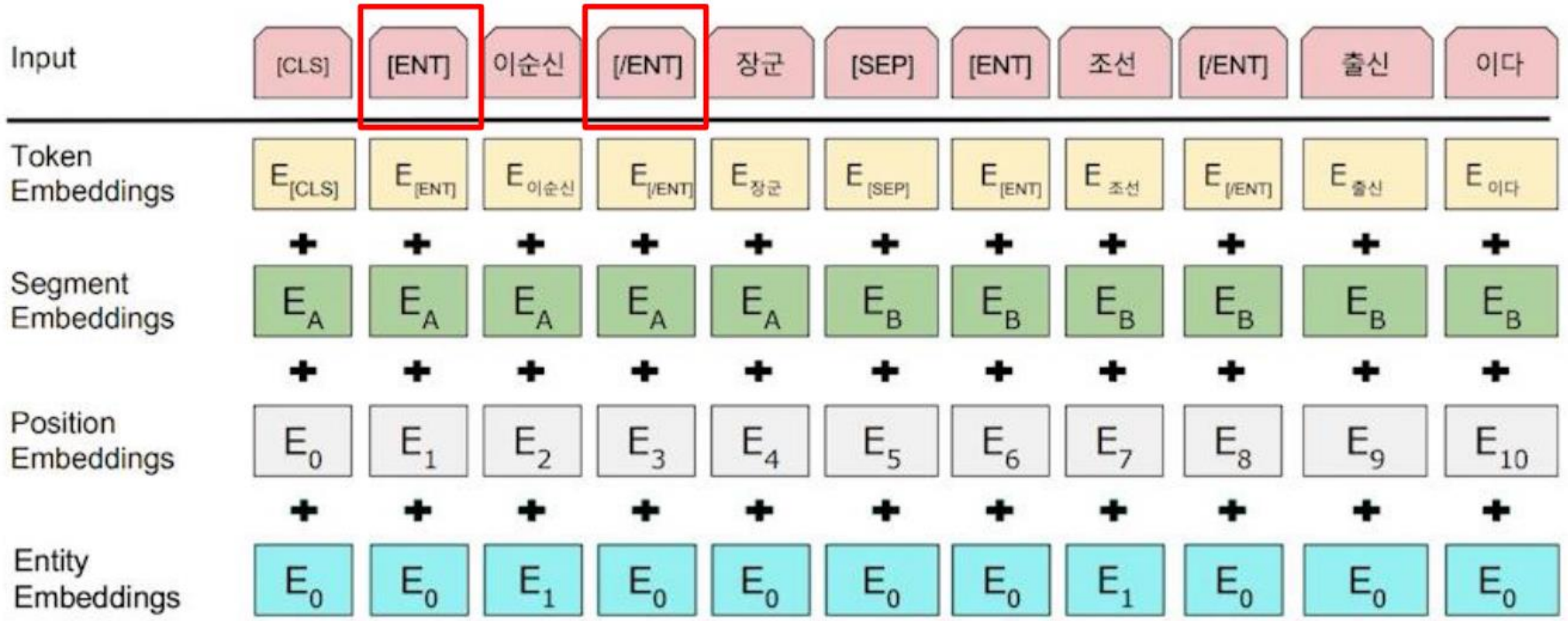
성능 향상 : 71.0110 → 75.0477

Entity Token

03

Why?

Entity를 더 강조하기 위해 문장의 Entity 위치를 표시



Why?

Entity를 더 강조하기 위해 문장의 Entity 위치를 표시

Method	Input Example	$\text{BERT}_{\text{BASE}}$	$\text{BERT}_{\text{LARGE}}$	$\text{RoBERTa}_{\text{LARGE}}$
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**

더 좋은 방법도 있지 않을까?

03 How?

- **Entity Marker**
  - [SUBJ]이순신[/SUBJ] 장군은 [OBJ]조선[/OBJ] 출신이다
- **Typed Entity Marker**
  - [SUBJ:PER]이순신[/SUBJ] 장군은 [OBJ:ORG]조선[/OBJ] 출신이다
- **Entity Mask**
  - [SUBJ:PER] 장군은 [OBJ:ORG] 출신이다
- **Type Marker**
  - [PER]이순신[/PER] 장군은 [ORG]조선[/ORG] 출신이다
- **Typed Entity Marker (punct)**
  - @*사람*이순신@ 장군은 #^ORG^조선# 출신이다
- **Typed Entity Marker (punct)** (순서에 따라)
  - @*사람*이순신@ 장군은 #^ORG^조선# 출신이다
- **Typed Entity Suffix**
  - *이순신*[TP]사람[/TP] 장군은 *조선*[TP]ORG[/TP] 출신이다

03 How?

- **Entity Marker**
  - [SUBJ]이순신[/SUBJ] 장군은 [OBJ]조선[/OBJ] 출신이다
- **Typed Entity Marker**
  - [SUBJ:PER]이순신[/SUBJ] 장군은 [OBJ:ORG]
- **Entity Mask**
  - [SUBJ:PER] 장군은 [OBJ:ORG] 출신이다
- **Type Marker**
  - [PER]이순신[/PER] 장군은 [ORG]조선[/ORG] 출신이다

Best Score

66.1128 -> 73.1450
Micro-F1 7점 상승

- **Typed Entity Marker (punct)**
  - @*사람*이순신@ 장군은 #^ORG^조선# 출신이다
- **Typed Entity Marker (punct) (순서에 따라)**
  - @*사람*이순신@ 장군은 #^ORG^조선# 출신이다
- **Typed Entity Suffix**
  - *이순신*[TP]사람[/TP] 장군은 *조선*[TP]ORG[/TP] 출신이다

03 Result

Special token을 추가하는 방식은 생각보다 좋지 않다

- 추가한 special token을 처음부터 학습해야 하기 때문

Subject/Object별 marker < 순서에 따라 marker

- Subject와 object의 구분보다 단어의 위치 정보에 따른 marker가 문맥 파악에 더 중요한 요소

Multi Sentence

04 Multi Sentence

Baseline

- Baseline Concat Entity : Subject Entity[SEP]Object Entity
- Baseline Score : f1 score – 62.3808 / auprc – 59.7112

● 원문: 이순신 장군은 조선 출신 이다. Entity_01: 이순신 Entity_02: 조선

- Multi
● [CLS] 이순신 장군은 조선 출신 이다. [SEP] 이 문장에서 이순신과 조선은 어떤 관계일까? [SEP]

- Single
● [CLS] 이순신 장군은 조선 출신 이다. 이 문장에서 이순신과 조선은 어떤 관계일까? [SEP]

● 기존 BERT의 Pretrain 방식과 유사한 input을 만들어 줄 수 있다

04 Multi Sentence

1. Token 제외

"이 문장에서 Subject Entity와 Object Entity의 관계를 구하시오."

f1 score - 62.3808 / auprc – 59.7112 ⇢ f1 score - 69.9848 / auprc - 64.2220

2. Entity 강조

"이 문장에서 *Subject Entity*와 *Object Entity*의 관계를 구하시오."

f1 score - 69.9848 / auprc - 64.2220 ⇢ f1 score - 74.5912 / auprc - 79.6785

04 Multi Sentence

3. Type 명시 추가

"이 문장에서 *Subject Entity[Subject Type]*와 *Object Entity[Object Type]*의 관계를 구하시오."

f1 score - 74.5912 / auprc - 79.6785 ⇢ f1 score - 75.0628 / auprc - 79.1046

4. Object를 주어로 + Subject와의 관계로서 Object 타입 명시

"이 문장에서 [Object Entity]는 [Subject Entity]의 [Object Type]이다."

f1 score - 75.0628 / auprc - 79.1046 ⇢ f1 score - 75.7209 / auprc – 82.7783

Private 최고점 : f1 score 76.1568 auprc 83.1799

04 Multi Sentence

Label의 특성에 따른 접근으로 성능 향상

1. Label의 특성 -> Subject에 대해서는 독립적인 타입만을 판단, Object에 대해서는 Subject와의 관계를 판단

2. Object가 무엇인지, Subject와의 관계 속에서의 성질이 어떠한 지를 파악하는 것이 중요 -> 관계 질문에서 관계 설명 평서문으로 문장 구성 바꿈

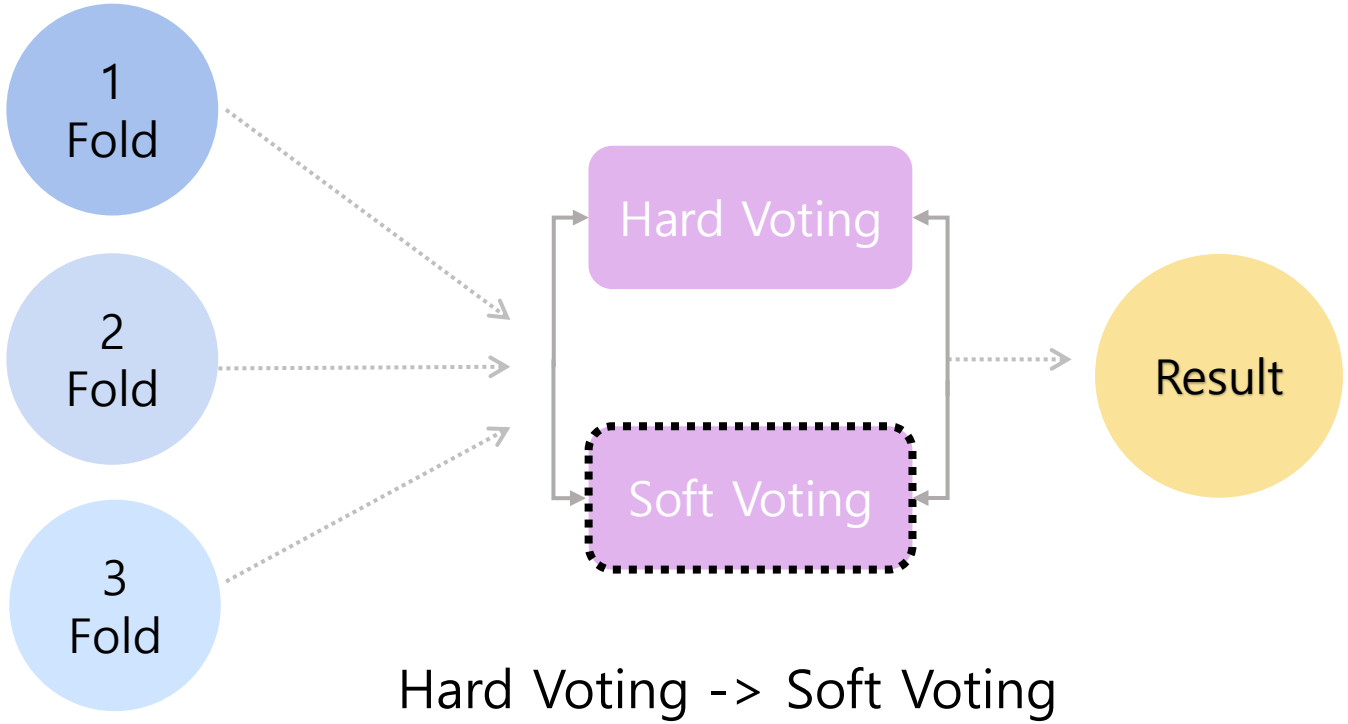
3. Object를 주어로 하고, Object의 Type만 추가하여 Object를 보다 강조

4. 모델이 한국어 문장구조를 제대로 학습했다면 앞선 문장 구조의 의미를 이해할 것이라 생각한 것이 성능 향상으로 이어짐

Ensemble

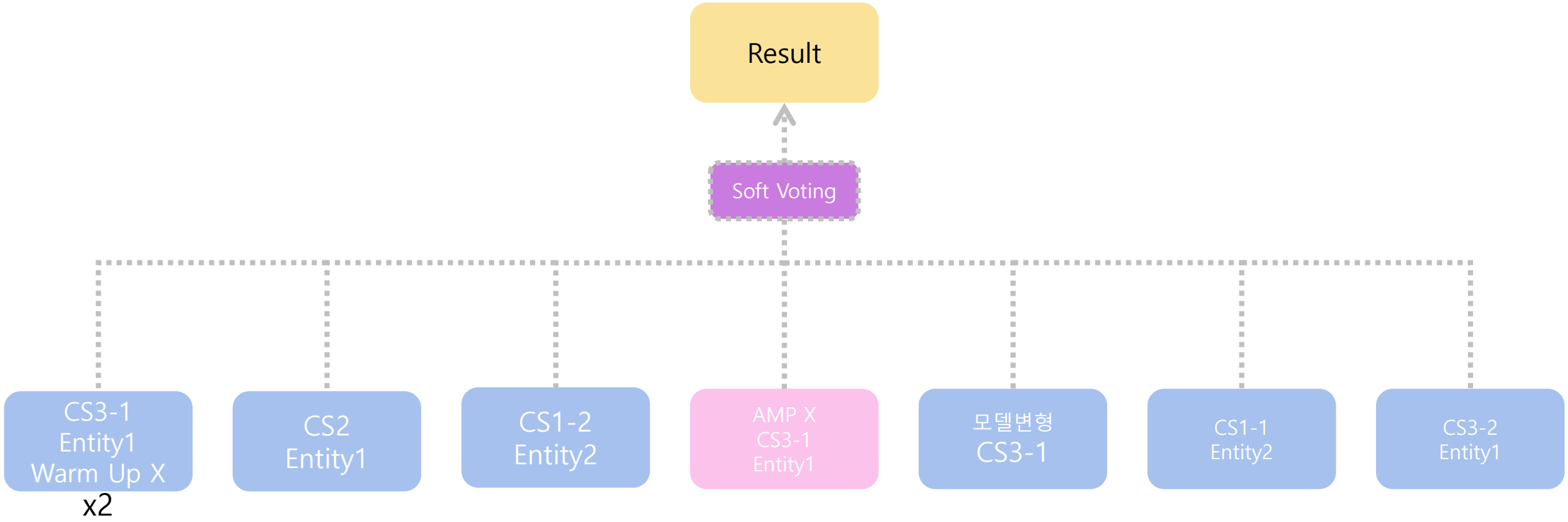
05 Ensemble

Stratified Shuffle Split



Hard Voting -> Soft Voting
micro f1 score : +0.81
auprc : +1.9

05 Ensemble



Base : klue/Roberta-large, 2e-5, 400(warmUp), 3(epoch), CrossEntropy(loss)
CS(Concat sentence)1-1 : *{sub}* 와 *{obj}*의 관계를 구하시오.
CS1-2 : *{sub}[{sub_type}]* 와 *{obj}[{obj_type}]* 의 관계를 구하시오.
CS2 : 이 문장에서 *{sub}*과 ^{obj}^은 어떤 관계일까?
CS3-1 : 이 문장에서 [{obj}]은 [{sub}]의 [{obj_type}]이다.
CS3-2 : 이 문장에서 [{obj}]은 [{obj_type}]인 [{sub}]의 [{obj_type}]이다.
Entity1(Sentence 안에): @*{TYPE[sub_type]}*{sub_word}@ ~ #^{TYPE[obj_type]}^{obj_word}#
Entity2: *{sub_word} [{TYPE[sub_type]}]* ~ *{obj_word}[{TYPE[obj_type]}]*
모델변형 : 토큰 concat

06 후기

	좋았던 점	아쉬웠던 점
임동진	모델 변경하면서 Huggingface 활용법을 많이 알게 되었습니다.	처음에 코드를 뜯어보려는 노력을 하지 않아 실패 사례가 너무 많았고, 시간적인 낭비가 너무 심했습니다. 또한 GitHub를 제대로 활용하지 못한 것 같습니다
정재윤	NLP에서 쓰이는 여러 augmentation기법들에 대해 알게되었습니다. 그리고 대회기간동안 hugging face 공식문서를 계속 찾아보면서 hugging face와 더욱 친숙해지게된 좋은 계기였습니다.	알게된 여러 기법들을 모두 사용해보지 못했고, 적용했던 augmentaion이 점수향상에 반영되지 않았던 것이 아쉬웠습니다.
조설아	Hugging face에 익숙해지는 시간을 가지게 되어 유익했습니다. 또한 국문법에 맞는 접근이 유의미한 효과를 거두었다는 것이 흥미로웠습니다.	entity embedding을 적용하는 코드를 짜지 못해서 활용하지 못한 점이 아쉬웠다. 모델을 task에 맞게 custom하는 능력이 필요함을 느꼈습니다.
허치영	팀원들과 많은 토론을 하며 대회가 원활하게 진행되어 좋았습니다.	TAPT 기법과 RECENT 기법을 제대로 사용해보지 못했습니다. 이번에는 Github를 처음부터 제대로 활용하지 못해서 다음에는 대회 시작부터 Github를 활용해보는 방향으로 진행해보고 싶습니다.
이보림	아이디어들을 직접 반영하고 그거에 따른 결과를 비교하고 분석해보는 시간이 정말 재미있었다.	여러 논문들을 더 많이 찾아보고 공부해봤으면 좋았을 것 같고 LDAM loss처럼 시간이 부족해서 못해본 것들이 있어 시간관리를 더 잘했으면 좋았을 것 같습니다.

경청해주셔서

감사합니다

QnA