

KiYOUNG2

P Stage 2: KLUE Relation Extraction Task Solution sharing

한국어 자연어 이해 관계 분류 데이터셋 1등 솔루션 공유

2021.09.27 ~ 2021.10.07

14조: 김대웅 김태욱 김채은 유영재 이하람 진명훈 허진규

We are KiYOUNG2 !!

01

Introduction

02

Baseline

03

Improvement

04

Conclusion



KiYOUNG2조에 대해 소개합니다!

Introduction



K i Y O U N G



안녕하세요! 기영이조 입니다 😊



01

KiYOUNG2조에 대해 소개합니다!

Introduction

K i Y O U N G

한 손엔 Dialogue



한 손엔 Korean

Korean is All YOU Need for dialoGuE



한 손엔 Dialogue



한 손엔 Korean

Korean is All YOU Need for dialoGuE



KiYOUNG2조에 대해 소개합니다!

Introduction

Ground Rule!

- Peer Session 시작 전 Small Talk 필수!
- Role Checking: 주간, 일일 목표 공유 및 진행 상황 체크
- 강의 질의응답 + 아이디어 회의
- 19:00~09:30 모각공 자율 참여

Communication

- 슬랙 확인 이모지 누르기! 🐟😊 소통 소통
- 급한 일은 카톡방에서 멘션 💬
- 늦으면 random으로 한명에게 커피 사기 ☕
- 아침 인사 및 TODO List 공유
- 시시콜콜한 대화나 TMI 자유롭게 방출하기

기영이 Notion!



회의록

💡 회의록 모음

📅 일정

Name	Assign	Date Created	Due Date	Priority
10차 피어세션	DaeUng Kim M MyungH	October 7, 2021 3:34 PM	October 7, 2021	High 🌟
09차 피어세션	DaeUng Kim M MyungH	October 6, 2021 3:58 PM	October 6, 2021	High 🌟
08차 피어세션	진규 혀 M MyungHoon	October 5, 2021 3:50 PM	October 5, 2021	
라벨링 오류 정제 작업 틀	T taeuk kim M 진규 혀	October 2, 2021 9:28 PM	October 3, 2021	
07차 피어세션	DaeUng Kim M MyungH	October 1, 2021 4:06 PM	October 1, 2021	High 🌟
06차 피어세션	DaeUng Kim M 이하람	September 30, 2021 4:06 PM	September 30, 2021	High 🌟
05차 피어세션	진규 혀 M MyungHoon	September 29, 2021 3:59 PM	September 29, 2021	High 🌟
04차 피어세션	DaeUng Kim M MyungH	September 28, 2021 4:03 PM	September 28, 2021	High 🌟
03차 피어세션	DaeUng Kim M MyungH	September 27, 2021 4:01 PM	September 27, 2021	High 🌟
1차 회의 - 역할분담	DaeUng Kim M MyungH	September 27, 2021 1:31 PM	September 27, 2021	High 🌟
02차 피어세션 : Ground Rule 정하기	DaeUng Kim M MyungH	September 24, 2021 4:32 PM	September 24, 2021	
01차 피어세션 & KLUE-RE Kick-off	DaeUng Kim M uyzae	September 23, 2021 4:22 PM	September 23, 2021	High 🌟
Kick off	DaeUng Kim M MyungH	September 13, 2021 4:12 PM	September 13, 2021 7:00 PM	High 🌟
팀 구성 완료 운영진에 보고하기	DaeUng Kim M MyungH	September 9, 2021 10:32 PM	September 16, 2021	High 🌟
04차 피어세션	DaeUng Kim T taeuk kir	September 28, 2021 4:03 PM	September 28, 2021	High 🌟
KLUE-RE 코드 정리 및 회고	DaeUng Kim M MyungH	October 7, 2021 9:04 PM	October 7, 2021	High 🌟 ?

The screenshot shows a Notion workspace titled "Kiyoung2". The left sidebar lists "Shared" pages: 목표, 공지사항, 팀원 소개, 일정, 팀 규칙, 회의록, TMI, KLUE-RE, KLUE-MRC, 스크랩, 김대웅_T2022, 김재은_T2064, 김태우_T2065, 유영재_T2140, 이하람_T2175, 진영호_T2216, 허인규_T2242, NLP News, NLP 논문 읽기, MLops, 피어세션. The main area has sections for "Meeting Minutes" (회의록), "Calendar" (일정), and "Personal" (개인 공간). The "Meeting Minutes" section contains a parrot icon and a list of meeting minutes from October 7, 2021, to September 9, 2021. The "Calendar" section shows a grid of tasks with columns for Name, Assign, Date Created, Due Date, and Priority. The "Personal" section shows a list of team members: 김대웅_T2022, 김재은_T2064, 김태우_T2065.



KIYOUNG2조에 대해 소개합니다!

Introduction

Ground Rule!

- Peer Session 시작 전 Small Talk 필수!
- Role Checking: 주간, 일일 목표 공유 및 진행 상황 체크
- 강의 질의응답 + 아이디어 회의
- 19:00~09:30 모각공 자율 참여

Communication

- 슬랙 확인 이모지 누르기! 🐟😊 소통 소통
- 급한 일은 카톡방에서 멘션 💬
- 늦으면 random으로 한명에게 커피 사기 ☕
- 아침 인사 및 TODO List 공유
- 시시콜콜한 대화나 TMI 자유롭게 방출하기

- 진명훈 : Modeling, Project Manager, RECENT, Data voyager, TAPT, data version control
- 김대웅 : PTR, Debugging Mento, HyperParameter Search, Ensemble, Model Analysis
- 김채은 : Data Manager, Electra, Roberta, LSTM Classifier, HyperParameter Search, EDA
- 허진규 : Modeling, Mental Care, Custom loss, XLM, Electra, Data Handling, EDA
- 이하람 : Code Manager, EDA, Data Handling, K-Fold, Model Analysis, MLOps
- 김유재 : Entity Embedding, EDA, Data Handling, Mood Manager, Experiment Management
- 유영재 : Data Augmentation, PORORO, XLM, Named Entity Recognition, Modeling, MLOps

M 진명훈_T2216 오전 9:59

오늘도 다들 화이팅입니다~
각자 어떤 하루 보내실지 짧게 공유해주세요!!
필요한 role은 아무래도

- 어제 회의 때 나온 내용인 Further EDA - POH 등 기준에 대해 생각 정하기
- 추가적인 baseline 실험 (모델 팀별 진행)
- HP Search 코드 통합

이 있을 것 같습니다 !! (예시입니다) (편집됨)

4 😊

이하람_T2175 오전 10:08

어제 마저 한 일

- 데이터셋 관련 셋팅 변경해서 BERT 실험 후 리더보드 제출
- 강의 리스닝만 1~7강까지 + 실습 1~2강까지

오늘 할 일

- bert multilingual 모델 사용할 수 있는지 확인해보기
- confusion matrix 등 모델 결과 시각화 그래프를 wandb에 추가 출력해보기
- imbalance sampler 사용해보기 + 결과 비교하기 + 모듈화하기
- Attention Visualization 알아보고 시도해보기
- 실습 3~4강

5 😊

김채은_T2064 오전 10:04

좋은 아침입니다~! 오늘 하루도 화이팅해보아요💪

- 강의 - 5강
- 데이터증복 인덱스 전달
- 어제 회의대로 baseline 맞추어 bert 실험
- ray tune붙이기
- entity error(?) 모호한 데이터 문제 생각해보기

5 😊

유영재_T2140 오전 10:11

- 강의 5,6강 수강
- 어제 회의 토대로 baseline 둘려보기
- 추가적인 EDA 시도 + 회의에서 얘기한 augmentation 기법 고민해보기

6 😊

김대웅_T2022 오전 10:20

- 어제 한 일
 - hyperparameter search 코드 작성 및 테스트 완료
 - 강의 2강 + 실습 + 과제
 - 베이스라인 코드 정리
- 오늘 할 일
 - 베이스라인 코드 정리 마무리 및 커밋
 - 강의 3강 + 실습 + 과제
 - 적절한 Dev set 탐색
 - 대회 베이스라인 입력 양식으로 실험
 - EDA 시도하기

6 😊

허진규_T2242 오전 10:30

- 어제 한 일
 - 명훈님 방식 baseline에 반영 후, 제출
 - 강의 4강 + 학습정리
- 오늘 할 일
 - 코드 정리 (반영하면서 많이 더러워졌다)
 - 강의 6강까지 듣기

6 😊

김태욱_T2065 오전 10:06

- 가지고 있는 코드 깔끔하게 정리하기
- 추가적인 EDA, 전처리 할 수 있는데까지 하기
- 강의 5,6 강
- entity, label issue 정리하기

추가적으로 어제 실험해 봤는데 성능이 새로 적용한 데이터셋을 사용할 때 조금 더 잘나왔습니다! (편집됨)

5 😊

1개의 담글 9일 전



KIYOUNG2조에 대해 소개합니다!

Introduction

개발 협업

- 기능별 Branch 만들어서 추가 후 Merge
- GitHub Issue로 PR 적극 관리
- Huggingface Datasets로 Versioning

The screenshot shows a GitHub repository page for 'klue-level2-nlp-p-14'. A modal window titled 'Switch branches/tags' is open, showing a list of branches and tags. A red box highlights the commit history on the right side of the page, which lists several commits made by 'boostcampai2tech2' over the past few days. A red box also highlights the commit 'fix: .gitignore'.

Recent commits (highlighted by a red box):

- fix: .gitignore (3 days ago)
- parameter search (9 days ago)
- ADME.md (9 days ago)
- nce Dataset Sampler with Trainer (6 days ago)
- nce Dataset Sampler with Trainer (3 days ago)
- nce Dataset Sampler with Trainer (3 days ago)

Text overlay: 대회 종료 후 최종 Merge 대기 중이에요 ㅎㅎ

Branches (list from modal):

- baseline
- bert_all
- entity_layer_tmp
- experiments/hrlee
- feature/cm
- feature/imsam
- feature/kfold
- feature/valid_hr
- fix-entity-tagging
- main_chaeun

Code tab is selected.

Usage section:

- run klue task

```
python new_run.py configs/{YOUR_CONFIG}.yaml
```

About section:

- klue-level2-nlp-p-14 created by GitHub Classroom
- Readme
- MIT License

Releases section:

- No releases published
- Create a new release

Packages section:

- No packages published
- Publish your first package

Contributors section:

- 5 contributors (avatars shown)

Languages section:

- Python 100.0%



KIYOUNG2조에 대해 소개합니다!

Introduction

개발 협업

- 기능별 Branch 만들어서 추가 후 Merge
- GitHub Issue로 PR 적극 관리
- Huggingface Datasets로 Versioning

<input type="checkbox"/> [Model] Ensemble - Top 6 Model Soft Voting modeling	#33 opened 2 days ago by KimDaeUng			
<input type="checkbox"/> [Data] entity, label 수정 ver.final data	#31 opened 4 days ago by taeukkkim			
<input type="checkbox"/> [Dev] Add Model Analysis Feature, SentencePiece Tokenizer modeling	#30 opened 4 days ago by KimDaeUng 2 tasks done			
<input type="checkbox"/> [Dev] Classification Layer 적용 modeling	#27 opened 7 days ago by taeukkkim 1 task done			
<input type="checkbox"/> [Dev] Custom loss 추가 및 모듈화 develop	#26 opened 7 days ago by JeangyuHeo 2 tasks done			
<input type="checkbox"/> [Dev] Optuna Hyperparameter Search code error develop	#25 opened 7 days ago by taeukkkim			
<input type="checkbox"/> [Dev] Pororo NER 모듈 huggingface로 porting develop	#24 opened 7 days ago by jinmang2			
<input type="checkbox"/> [Dev] Entity Special Token & Add Entity Embedding Layer modeling	#23 opened 7 days ago by KimDaeUng 2 tasks done			
<input type="checkbox"/> [Data] entity, label 수정 ver.1 data	#21 opened 8 days ago by taeukkkim			

<input type="checkbox"/> [DEV] Implement TAPT modeling	#20 opened 8 days ago by jinmang2			
<input type="checkbox"/> [Data] duplicated data 처리 data	#17 opened 10 days ago by Amber-Chaeeunk			
<input type="checkbox"/> [Data] Augmentation with NER data	#16 opened 10 days ago by uyeongjae			
<input type="checkbox"/> [Dev] K-Fold Reference code dev	#15 opened 10 days ago by KimDaeUng			
<input type="checkbox"/> [Data] EDA, Preprocessing data	#9 opened 11 days ago by taeukkkim			
<input type="checkbox"/> [Dev] Load train.csv to Huggingface Datasets dataset data	#8 opened 11 days ago by KimDaeUng			
<input type="checkbox"/> [Dev] Hyper-parameter search code & search space of baseline dev	#7 opened 11 days ago by jinmang2			
<input type="checkbox"/> CI/CD by GitHub Actions dev	#5 opened 11 days ago by jinmang2			
<input type="checkbox"/> Active Learning to fix entity labeling error data	#4 opened 11 days ago by jinmang2 3 tasks			
<input type="checkbox"/> [SOTA] Curriculum Learning for improving sentence-level RE modeling	#3 opened 11 days ago by jinmang2			
<input type="checkbox"/> [SOTA] RECENT: Model Agnostic approach for RE modeling	#2 opened 11 days ago by jinmang2			
<input type="checkbox"/> [Model] PTR: Prompt Tuning with Rules for Text Classification modeling	#1 opened 11 days ago by KimDaeUng			



KIYOUNG2조에 대해 소개합니다!

Introduction

개발 협업

- 기능별 Branch 만들어서 추가 후 Merge
- GitHub Issue로 PR 적극 관리
- Huggingface Datasets로 Versioning

새로운 Validation Alignment 혹은
Augmentation Data 추가될 때마다 Version으로 관리!

The screenshot shows a GitHub repository named 'load_klue_re' with a main branch. The repository has 27 commits. A specific commit by 'jinmang2' is highlighted, showing it was published to v3.0.1. The commit message is 're4 publish v3.0.1'. Below the commit, the file history is listed:

File	Size	Last Commit
.gitattributes	1.19 kB	train_test_split seed 42 t... last month
README.md	935 Bytes	Update README.md 2 days ago
dataset_infos.json	3.63 kB	re4 publish v3.0.1 2 days ago
klue_re.zip	7.07 MB	re4 publish v3.0.1 2 days ago
load_klue_re.py	6.84 kB	re4 publish v3.0.0 2 days ago



Baseline: EDA

Baseline

0	no_relation	서로 어떠한 관계도 없음
1	org:top_members/employees	지정된 조직(subject ORG)의 대표자 또는 구성원(object PER)
2	org:members	지정된 조직(subject ORG)에 속한 조직(object ORG)
3	org:product	지정된 조직(subject ORG)에서 생산하는 제품 또는 상품(object ?)
4	per:title	지정된 사람(subject PER)의 직위를 나타내는 공식 또는 비공식 이름(object POH)
5	org:alternate_names	지정된 조직(subject ORG)을 지칭하기 위해 공식 명칭 대신에 부르는 대체 명칭(object POH, others?)
6	per:employee_of	지정된 사람(subject PER)이 일하는 조직(object ORG)
7	org:place_of_headquarters	지정된 조직(subject ORG)의 본부가 위치한 장소(object LOC)
8	per:product	지정된 사람(subject PER)이 제작한 제품 또는 작품(object ?)
9	org:number_of_employees/mem	지정된 조직(subject ORG)에 소속된 총 구성원의 수(object NOH)
10	per:children	지정된 사람(subject PER)의 자녀(object PER)
11	per:place_of_residence	지정된 사람(subject PER)이 살았던 장소(object LOC)
12	per:alternate_names	지정된 사람(subject PER)을 지칭하기 위해 공식 이름 대신에 부르는 대체 이름(object POH)
13	per:other_family	지정된 사람(subject PER)의 부모, 자녀, 형제자매 및 배우자가 아닌 가족 관계의 사람(object PER)
14	per:colleagues	지정된 사람(subject PER)과 함께 일하는 사람들(object PER)
15	per:origin	지정된 사람(subject PER)의 출신 또는 국적(object LOC, ORG)
16	per:siblings	지정된 사람(subject PER)의 형제 혹은 자매(object PER)
17	per:spouse	지정된 사람(subject PER)의 배우자(object PER)
18	org:founded	지정된 조직(subject ORG)가 수립된 날짜(object DAT)
19	org:political/religious_affiliation	지정된 조직(subject ORG)이 소속된 정치/종교 단체(object ORG, others?)
20	org:member_of	지정된 조직(subject ORG)이 속한 조직(object ORG)
21	per:parents	지정된 사람(subject PER)의 부모(object PER)
22	org:dissolved	지정된 조직(subject ORG)가 해산된 날짜(object DAT)
23	per:schools_attended	지정된 사람(subject PER)이 다녔던 학교(object LOC, ORG)
24	per:date_of_death	지정된 사람(subject PER)이 죽은 날짜(object DAT)
25	per:date_of_birth	지정된 사람(subject PER)이 태어난 날짜(object DAT)
26	per:place_of_birth	지정된 사람(subject PER)이 태어난 장소(object LOC)
27	per:place_of_death	지정된 사람(subject PER)이 죽은 장소(object LOC)
28	org:founded_by	지정된 조직(subject ORG)을 설립한 사람 또는 조직(object ORG, PER)
29	per:religion	지정된 사람(subject PER)이 믿는 종교(object ?)



Baseline: EDA

Baseline

29357 | per:date_of_death

"요기 베라:PER는 자신의 메이저 데뷔 일과 같은 날로 내보였던 2015년 9월 22일에는 뉴저지주:LOC 웨스트 칼드웰에 있는 고령자 집합 주택에서 자는 도중에 노환으로 자연사(별세)했다."

14958 | org:top_members/employees

"그룹 2AM:PER 출신 정진운:PER 씨와 나인뮤지스 출신 경리가 열애를 인정한 가운데 두 사람이 듀엣곡을 작업하면서 인연을 맺게 됐다는 최근 증언이 전해졌다."

18065 | org:members

"2001년 3월 28일부터 2003년 12월 24일까지 바실리예프는 러시아:ORG 내무부 차관직을 수행하였고, 모스크바:DAT 극장 인질극 사건 때 인질들의 구출을 위한 책임을 맡았다."

26777 | org:political/religious_affiliation

"일찍이 독일에는 바이마르 공화국 시대의 독일 국가방위군이나 나치 독일:DAT 시대의 독일 국방군:ORG(Wehrmacht) 등의 군대가 있었다."

19531 | per:date_of_death

"페르시아 제국의 다리우스 3세:PER가 알렉산드로스 대왕에게 패하자 그를 살해한 박트리아:LOC의 총독 베수스는 민족적 저항을 조직하려고 하였다."

32284 | per:colleagues

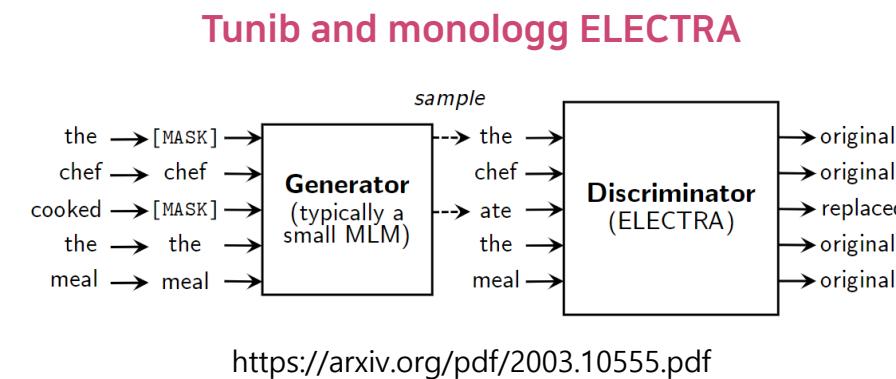
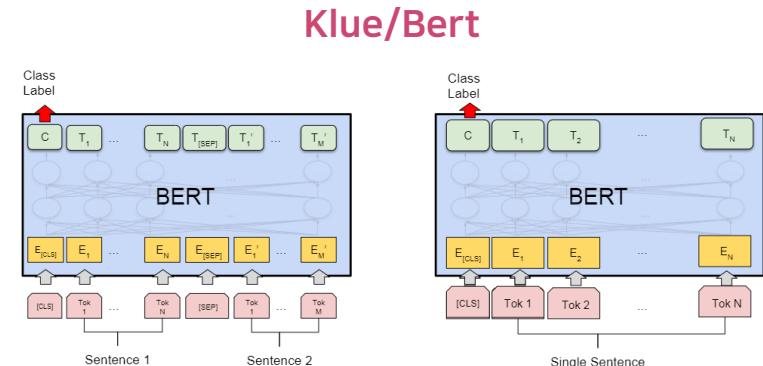
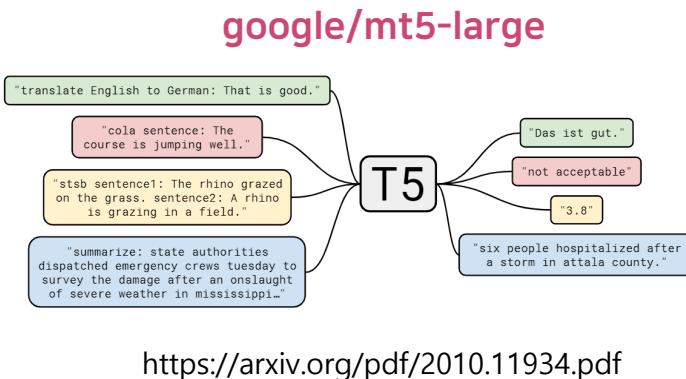
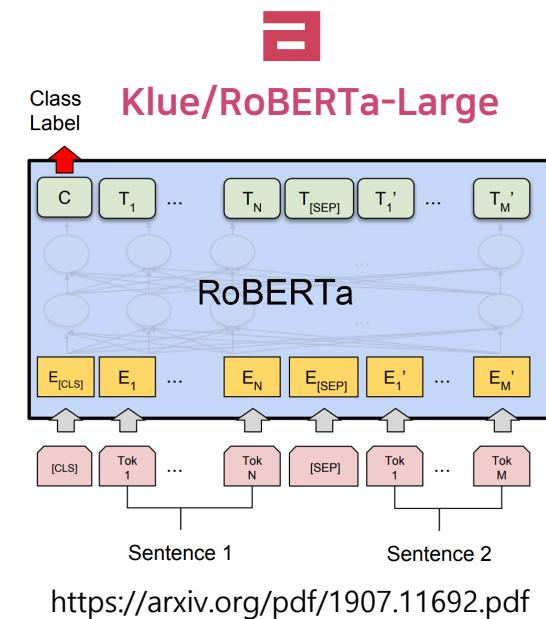
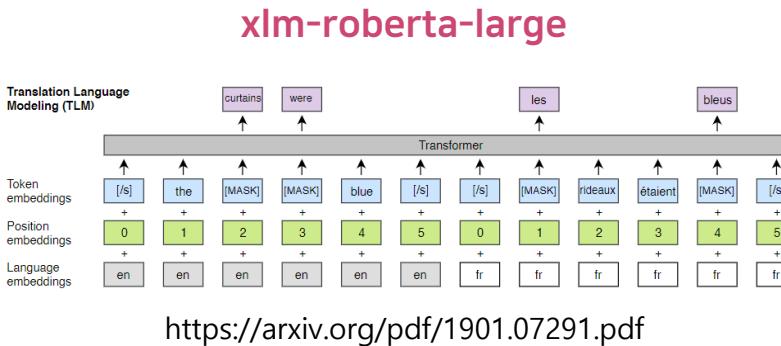
"초반에는 도쿄에서 한 마을인 히나미자와:LOC 마을에 전학 온지 얼마 안된 소년 마에바라 케이이치:PER와 개성적인 동아리의 동료들과의 일상을 그리지만 중반 이후에 갑자기 발생하는 괴사건과 서스펜스가 공포 분위기를 두드러지게 한다."

띄용? 이번에도 어김없이 데이터에 노이즈가...!! (생각보다 많아요!)

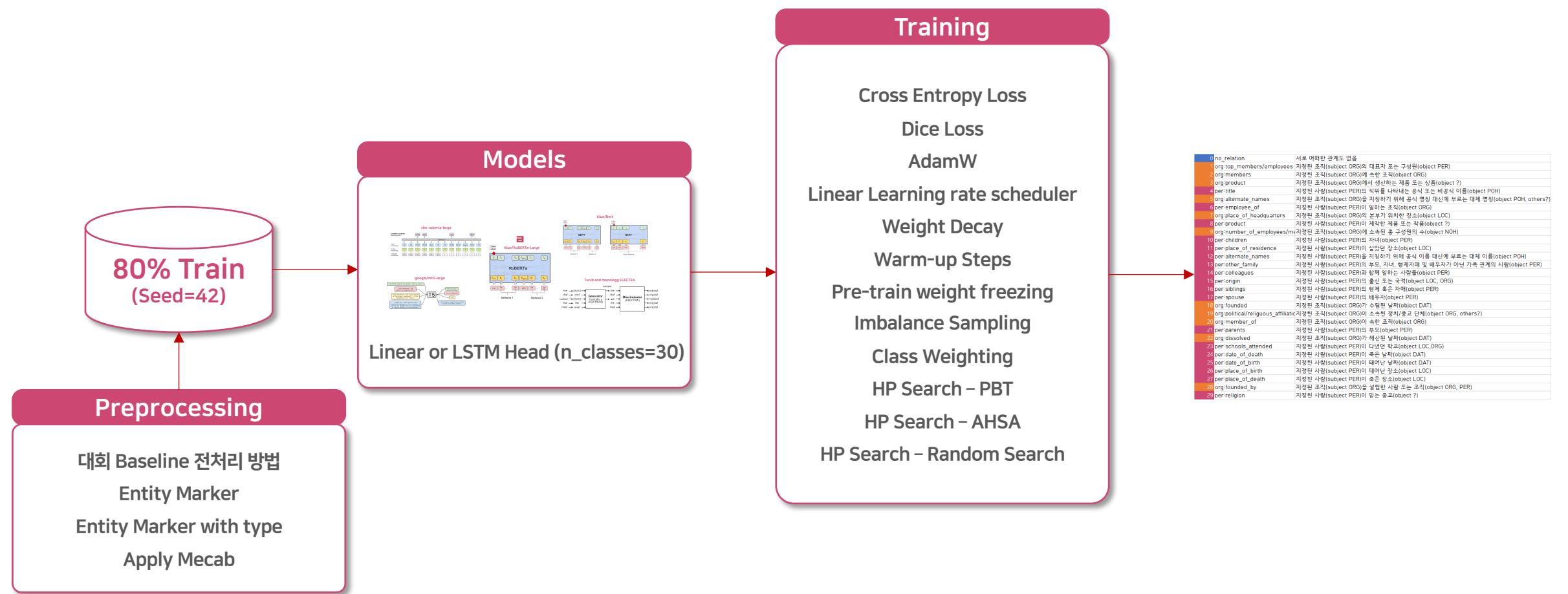


02 Baseline: Models

Baseline



한국어 특화 모델 + Multilingual 모델로 초기 실험 진행



30개의 클래스를 분류하는 모델을 다양한 조건에서 실험!



Baseline: Insights

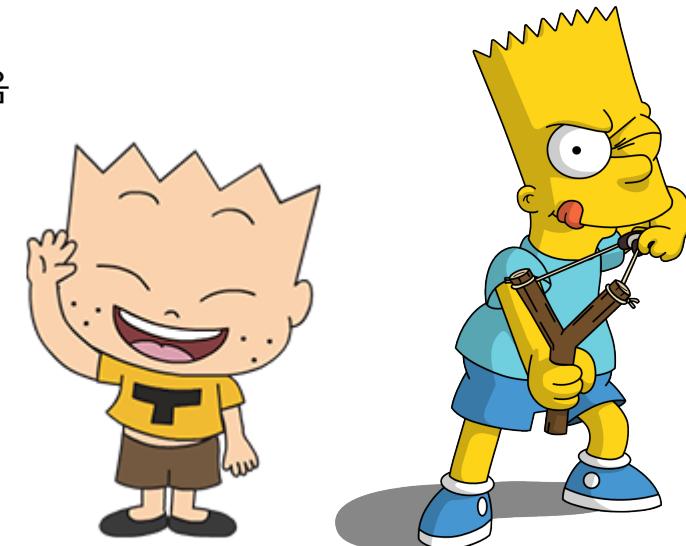
Baseline

누가 누가 잘했나?

- RoBERTa 기반의 모델들의 성능이 가장 좋았어요!
- 하지만... 계속해서 밀려나는 LB 순위... 특단의 조치가 필요했습니다!
- Baseline을 구축하면서 저희가 얻은 Insights
 1. Labeling은 생각보다 잘 되어있음 (물론 완전히 깔끔하진 않음)
 2. Entity Type에 굉장히 많은 noise가 있음
 3. 꼭 30진 분류로 모델을 풀어야 하는가? Label들을 보면 hierarchical한 관계가 있음
 4. Class Imbalance 문제는 Level 1 대비 훨씬 고난이도 □ □

이를 어떻게 개선할까?

- 3장에서 언급드릴게요!!





성능 개선: ① 데이터 Denoising

Improvement

- 앞에서 언급된 데이터 cleaning
- Active Learning을 적용하고자 했으나 모델 성능이 95%이상으로 올리기가 힘들어서 직접 수작업
- Entity Pair별 계층 관계를 확인하고 re-tagging

id	sentence	subject_entity	object_entity	subject_type	object_type	source
125	그룹 '소녀시대' 출신 제시카가 중국 매니먼트에 20억	소녀시대	제시카	PER	PER	wikitree
243	트래비스 재러드 블랙리(Travis Jarrod Blackley, 1982) 퍼시픽리그	일본 프로 야구	ORG	LOC	wikipedia	
1016	노르망디 뇌부르(현재의 외르주) 출신으로, 프랑스 혁명 노르망디	프랑스	ORG	POH	wikipedia	
1101	2015-16 오스트레일리아 W 리그 시즌에서 8골을 기록한 시드니	오스트레일리아	ORG	POH	wikipedia	
1216	마다 왕국 치하의 마을은 그라마카스라는 마을 이장과 같다	판자아트	ORG	LOC	wikipedia	
1848	이에 아웃사이더 죽은 엠넷 '쇼미더머니' 출연 간과 관련 아싸 커뮤니케이	아웃사이더	ORG	ORG	wikipedia	
2250	2007년 12월 25일에 일본 프로 야구 센트럴 리그	일본 프로 야구	ORG	LOC	wikipedia	
2464	과거에는 대한민국 육군의 육군훈련소에서 카투사	대한민국 육군	ORG	LOC	wikipedia	
2585	데이즈 앤 라이브(daze alive) 제리케이가 2012년 2월 제리케이	daze alive	ORG	ORG	wikipedia	
2739	도쿠가와 씨는 1800년대 중반에 에도 막부가 문을 닫을 때도 막부	도쿠가와	ORG	ORG	wikipedia	
3118	제4대 쟁차, 씨네, 호연, 유리, 윤아로 소녀시대의 두 번째 소녀시대	소녀시대-Oh!GG	PER	PER	wikipedia	
3175	정식 명칭은 일본 프로 야구 조직 센트럴 리그 운동부이 센트럴 리그	일본 프로 야구	ORG	LOC	wikipedia	
3194	미군정 하에서 창설된 대한민국 국군에 참여하여 대한민국 국군	대한민국 국군	ORG	LOC	wikipedia	
3297	경기도 종합체육대회 유치를 위해 가평군, 용인 가평군	경기도	ORG	POH	wikitree	
3607	제4수송사령부는 미국 육군의 수송대 사령부이다. 제4수송사령부	미국 육군	ORG	LOC	wikipedia	
3925	그러나 2010년 6월 말 인천교통공사에서 월미동하버일 인천교통공사	인천광역시	ORG	POH	wikipedia	
4197	재단법인 오뚜기한태재단은 6월 27일 인천 송도컨벤 오뚜기	한태로	ORG	ORG	wikitree	
4245	리아드를 연고로 하는 사우디 프로리그 클럽으로는 1941 알나스르	사우디 프로리그	ORG	LOC	wikipedia	
4264	2018년 1월 30일 우즈베키스탄 리그의 FK 토코모티즈 E 우즈베키스탄 리	FK 토코모티즈 E 우즈베키스탄	리	ORG	wikipedia	
4428	쾰른의 브루노(1130년 - 1101년 10월 6일)은 로마 가톨릭교회 쾰른의 브루노	쾰른의 브루노	ORG	LOC	wikipedia	
4513	임지선 대표는 세계 최고의 브랜드 베를리가 보해양조 임지선	보해양조	PER	ORG	wikitree	
4758	2021년 경기도 종합체육대회 유치를 위해 최종판 판주 가평군	경기도	ORG	POH	wikitree	
5053	소울 컴퍼니의 Kebee와 함께 이름 프로젝트 알리, "Elu" 소울 컴퍼니	Kebee	ORG	ORG	wikipedia	
5362	해군특수전전단(海軍特殊戰戰團)은 대한민국 해군(ROK 해군특수전전단) 대한민국 해군	대한민국 해군	ORG	LOC	wikipedia	
5527	국립과천과학관(國立果川科學館)은 기초과학·융합과학 국립과천과학관 대한민국 과학기	대한민국 과학기	ORG	LOC	wikipedia	
5955	횡성 고씨 고구려 왕실종친회는 2015년 대한민국 통계: 고구려 왕실종친 횡성 고씨	ORG	LOC	wikipedia		
6232	국립과천과학관(國立果川科學館)은 기초과학·융합과학 과천과학관	대한민국 과학기	ORG	LOC	wikipedia	
6277	사법연수원(司法研修院)은 판사의 연수와 사법연수생과 사법연수원	대한민국 대법원	ORG	LOC	wikipedia	
6298	부산 군기지(釜山軍基址)는 부산광역시 남구에 있는 부산 해군기지	대한민국 해군	ORG	LOC	wikipedia	
6583	김두봉은 같은해 7월 정치 조직으로 조선독립운동을, 김길두봉	조선의종군	PER	ORG	wikipedia	
6713	국립나주병원(國立羅州病院)은 정신질환을 가진 사람에 국립나주병원	대한민국 보건복지	ORG	LOC	wikipedia	
6920	그러나 1125년, 요나라는 또 다른 불복계 민족인 여진족 금나라	여진족	ORG	POH	wikipedia	
6937	제5공중기동비행단(第五空中機動飛行團)은 김해국제공항제5공중기동비행단 대한민국 공군	대한민국 공군	ORG	LOC	wikipedia	
7024	1010년 6월 17일 중국 구미다 전투에서 카자흐의 유타를 학살한	대한민국 육군	ORG	LOC	wikipedia	

Entity Type별 클래스 분포

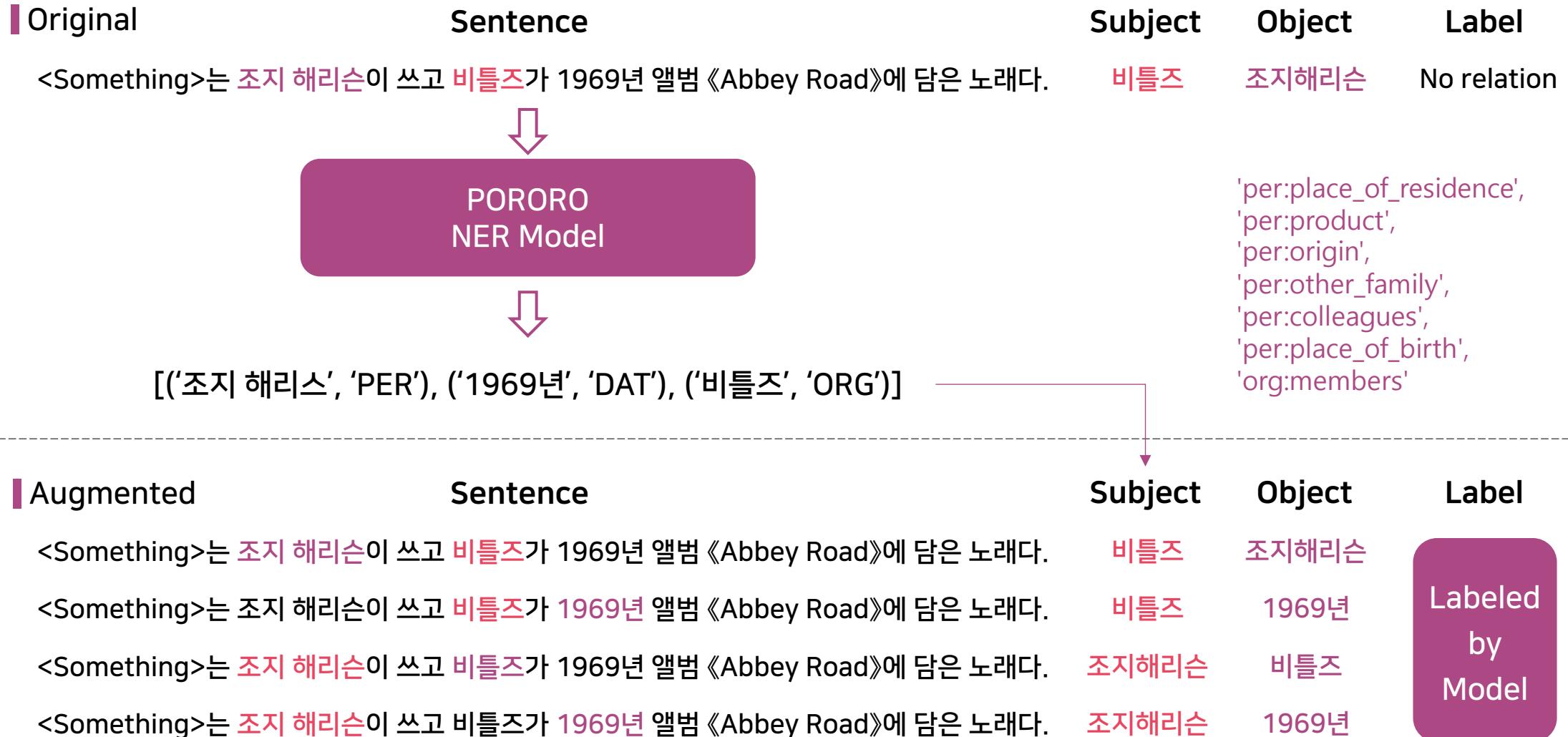
Aa 이름	태그
ORG_PER	org:founded_by org:top_members/employees
ORG_ORG	org:founded_by org:member_of org:political/religious_affiliation org:members
ORG_DAT	org:dissolved org:founded
ORG_LOC	org:member_of org:place_of_headquarters org:members
ORG_POH	org:alternate_names org:product org:top_members/employees
ORG_NOH	org:number_of_employees/members
PER_PER	per:colleagues per:spouse per:children per:parents per:other_family per:siblings
PER_ORG	per:employee_of per:schools_attended per:origin per:religion
PER_DAT	per:date_of_birth per:date_of_death
PER_LOC	per:place_of_birth per:place_of_residence per:place_of_death
PER_POH	per:title per:alternate_names per:product
PER_NOH	



03

성능 개선: ② Pororo NER을 활용한 Data Augmentation

Improvement





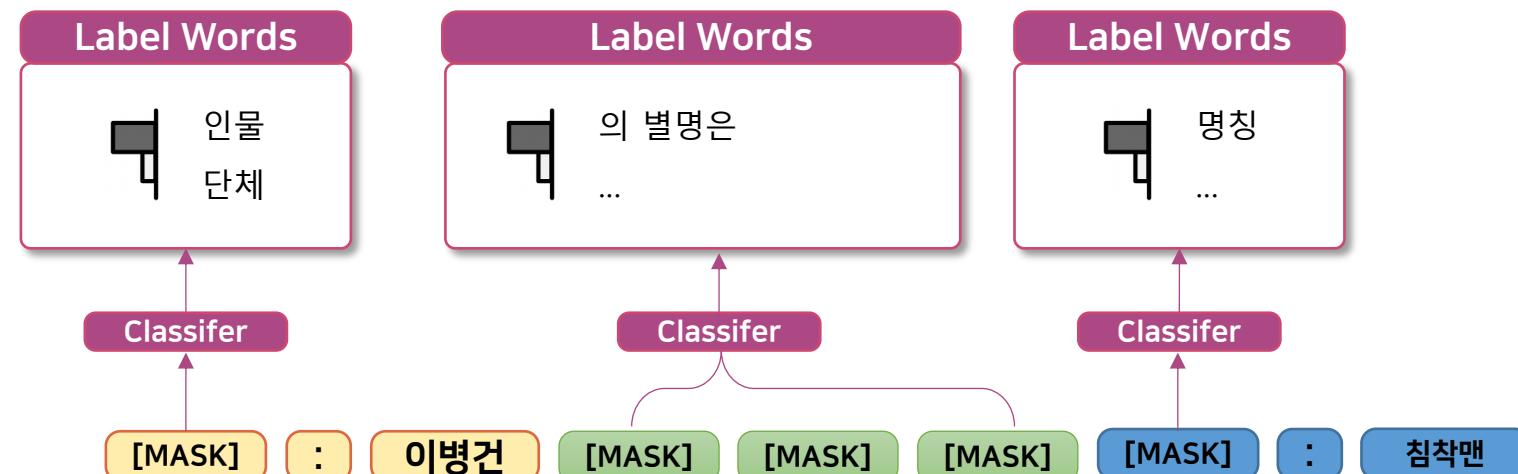
03

성능 개선: ③ Prompt Tuning with rules for Text Classification

Improvement

Prompt Engineering을 이용한 성능 개선 시도

- 인간의 논리적 추론 방식을 반영한 분류 방식
 - e.g) "per: alternative_name" relation인지 판단할 때
 - 각 entity의 속성이 특정 조건을 만족하는지 두 가지 조건을 확인함
 - entity의 속성 분류: 각 entity가 인물, 명칭에 해당하는지
 - Entity간의 관계 분류: 두 entity간의 관계가 인물과 다른 이름에 대한 관계인지
- 각 Entity의 속성을 분류하는 Prompt와 Entity간 관계를 분류하는 Prompt를 결합하여 관계 분류
 - e.g.) per:alternative_name



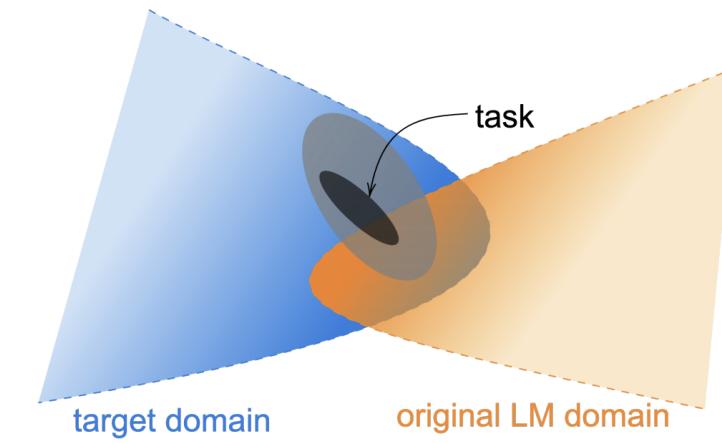


03

성능 개선: ④ Input Representation & Further Pre-training

Improvement

Input	[CLS]	[ENT]	이순신	[ENT]	장군	[SEP]	[ENT]	조선	[ENT]	출신	이다
Token Embeddings	$E_{[CLS]}$	$E_{[ENT]}$	$E_{이순신}$	$E_{[ENT]}$	$E_{장군}$	$E_{[SEP]}$	$E_{[ENT]}$	$E_{조선}$	$E_{[ENT]}$	$E_{출신}$	$E_{이다}$
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B	E_B
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
Entity Embeddings	E_0	E_0	E_1	E_0	E_0	E_0	E_0	E_1	E_0	E_0	E_0



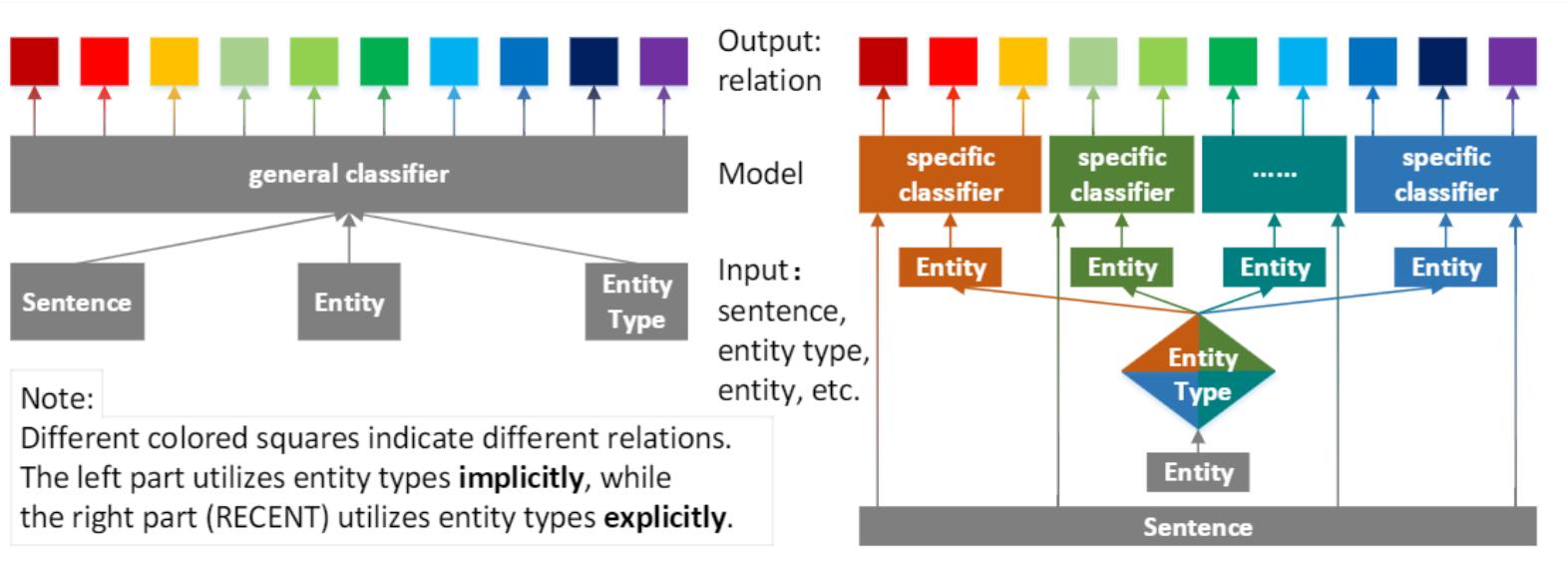
Entity Embedding + TAPT



03

성능 개선: ⑤ Relation Classification with Entity Type Restriction

Improvement



Algorithm 1 RECENT

Learning Phase:

Input: $\mathcal{D} = \{(se_i, s_i, o_i, ts_i, to_i, r_i) | i = 1, 2, \dots, N\}$ where the subscript i indicates the i th sample, se is sentence, s is subject entity, o is object entity, ts is type of subject entity, to is type of object entity, r is relation.

Output: Multiple classifiers.

- 1: Group sentences by entity types.
- 2: **for** each group g (entity types (ts, to)) **do**
- 3: aggregate relations in the group as candidate relations $R_{(ts,to)}$ defined in Eq. 3.
- 4: learn a classifier (marked as f_g) on the group that maps $\{(se_i, s_i, o_i) \in g\}$ to $R_{(ts,to)}$.
- 5: **end for**

Prediction Phase:

Input: A new sample $\{se, s, o, ts, to\}$, each specific classifier for each pair of entity types.

Output: A relation.

- 6: match the sample to a group (marked as g') according to the entity types (ts, to) .
- 7: Use the classifier ($f_{g'}$) learned on the group to map (se, s, o) to a relation.
- 8: **return** the relation.

최신 TACRED SOTA인 RECENT 적용!



Our Solution

Conclusion

- 이하 모델에 대한 양상을 결과
 - Dataset : Ver1.0.0(full), 1.0.1b(train), 2.0.1(full2), 2.0.1b(train2), 3.0.1(full3), 3.0.1b(train3)
 - 대회 제공 베이스라인과 KLUE 베이스라인 상이
 - 순위 / 제출자 / F1 / AUPRC / 모델 / 데이터셋 / 상세보기 설명 및 기타 세팅 / 사용 베이스라인 / 기타
1. 진명훈 / 76.360 / 73.679 / klue-roberta-large + TAPT / full3 + aug1 / KLUE 베이스라인 / final_jinmang2.csv
 2. 김태욱 / 72.467 / 74.668 / klue-roberta-large + TAPT / LSTM-classifier, focal loss / full3 + aug1 (5-fold) / KLUE 베이스라인/ Roberta_lstm_kfold.csv
 3. 이하람 / 72.466 / 72.643 / klue-roberta-large / full2 (5-fold) / KLUE 베이스라인/ submission_roberta_large_fold_complete.csv
 4. 진명훈 / 71.681 / 74.031 / klue-roberta-large / full / KLUE 베이스라인 / submission_finetune_tapt_0930.csv
 5. 김대웅 / 71.100 / 74.771 / klue-roberta-large / full / KLUE 베이스라인 / submission-baseline-0929-roberta-large-warmup0.2_lr3e-5-newdata.csv, epoch 4
 6. 진명훈 / 70.953 / 72.579 / klue-roberta-large / full / KLUE 베이스라인 / submission2.csv
 7. 유영재 / 70.728 / 74.084 / xlm-roberta-large / full / 대회 제공 베이스라인 / max_len=128 epoch4~5
 8. 허진규 / 70.440 / 74.911 / xlm-roberta-large / full / 대회 제공 베이스라인 / xlm-roberta-large batch_64 3e-5, epoch=10 몇 번째 epoch인지 모름
 9. 유영재 / 70.146 / 75.625 / xlm-roberta-large / train / 대회 제공 베이스라인 / base_setting xlm-roberta-large batch_64 max_len=128 epoch4~5

총 9개 모델에 대한 Soft Voting 결과!



04 아쉬웠던 점 + 중요했던 점

Conclusion

아쉬웠던 점

- Optimizer에 대한 깊은 고민을 하지 않았다
- Linear Scheduler 외 다른 실험을 진행하지 않았다
- Sortish Sampler + Curriculum Learning 아이디어 적용 못함
- 역할 분배의 쏠림 현상

중요했던 점

- Dataset Versioning으로 빠른 협업
- SOTA 모델 반영을 위한 데이터 처리 과정
- 데이터 셋에 대한 깊은 이해도를 모델링에 반영
- NER을 활용한 Data Augmentation (성능 개선)
- Focal Loss 및 Imbalance Sampler로 Class Imbalance 개선

감사합니다!