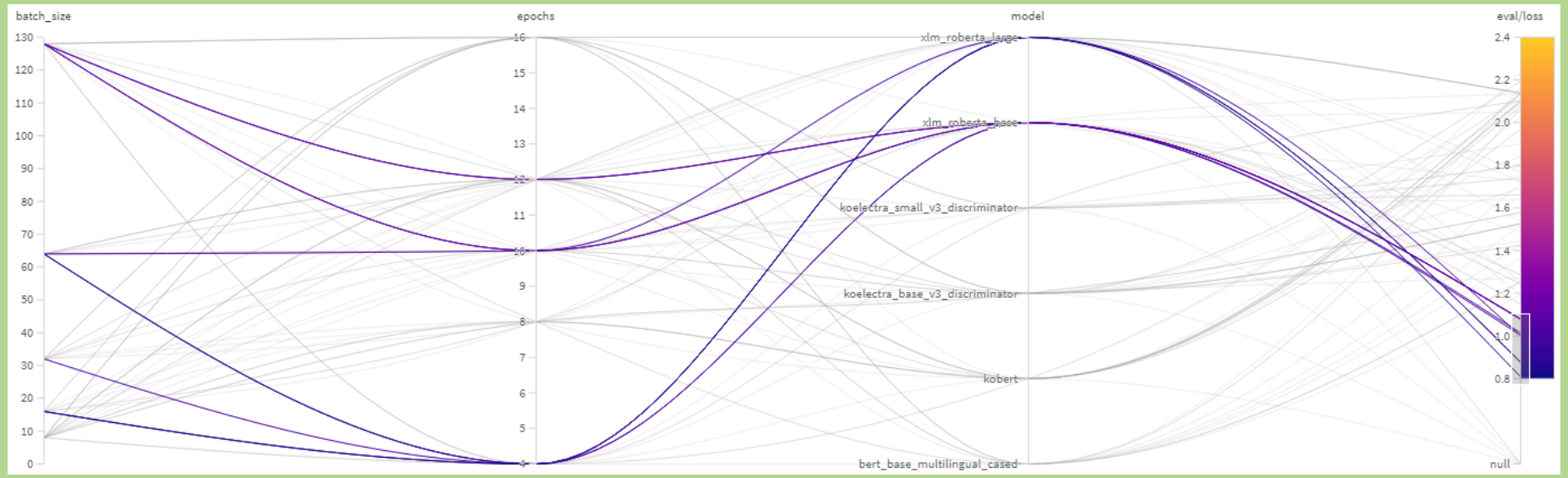
P-stage2 발표용

이상건

11	12	13	14	15	16	17
P-stage2대비 wandb Blog Post 상건 이	P-stage2 [Day1] Blog Post 상건 이	P-stage2 [Day2] Blog Post 상건 이	P-stage2 [Day3] Blog Post 상건 이	P-stage2 [Day4] Blog Post 상건 이	P-stage2 [Day5] Blog Post 상건 이	
18	19	20	21	22	23	24
	P-stage2 [Day6] autoML로 인... Blog Post 상건 이	P-stage2 [Day7] focalloss, 다... Blog Post 상건 이	P-stage2 [Day8] 외부 데이터, ... Blog Post 상건 이	P-stage2 [Day9] 외부 데이터, Blog Post 상건 이		

제출 day 1



- 주말동안 AutoML로 돌려 가장 성능이 좋았던 파라미터로 실험 (batch_size 32, epochs4, xlm_Roberta_large model.). LB 75.3%이 나옴.
- 이 이후로 모델은 모두 xlm_Roberta_lage를 사용

제출 day 2

비교표

batch_size: 64, epoch: 9	val loss	val acc	public acc
cross_entropy	1.042	0.7678	76.8%
cross_entropy*0.75 + focal*0.25	0.9354	0.7722	77.2%
cross_entropy*0.25 + focal*0.75	0.6798	0.7711	77.1%

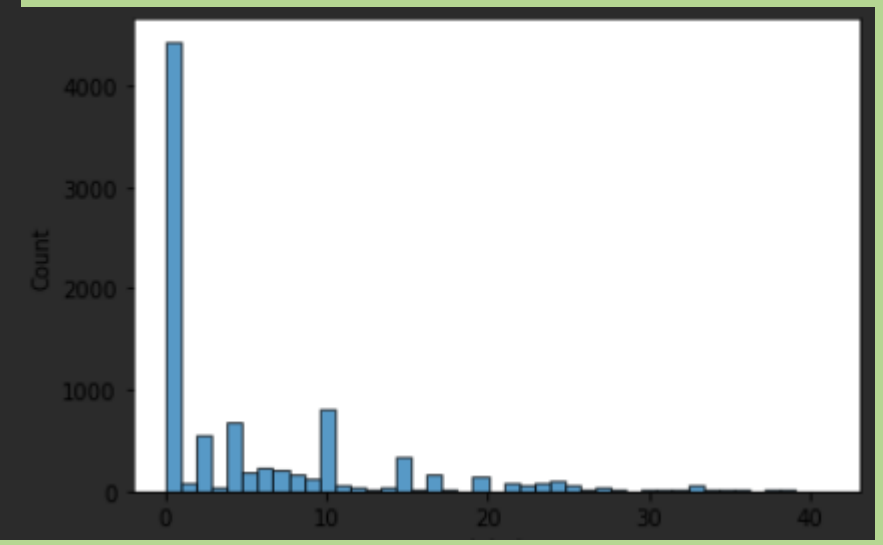
- Batch_size 64, epoch 9로 바꾸고 loss function을 cross entropy * 0.75 + focal loss 0.25 로 사용해서 LB 77.2%.
- 같은 조건에서 cross entropy만 사용할 때보다 LB 0.4% 상승

- Tokenizer를 roberta용이 아닌 다른걸로 써봤는데 acc이 0.5로 고정됨. 그냥 안 되는것 같다.

day 3

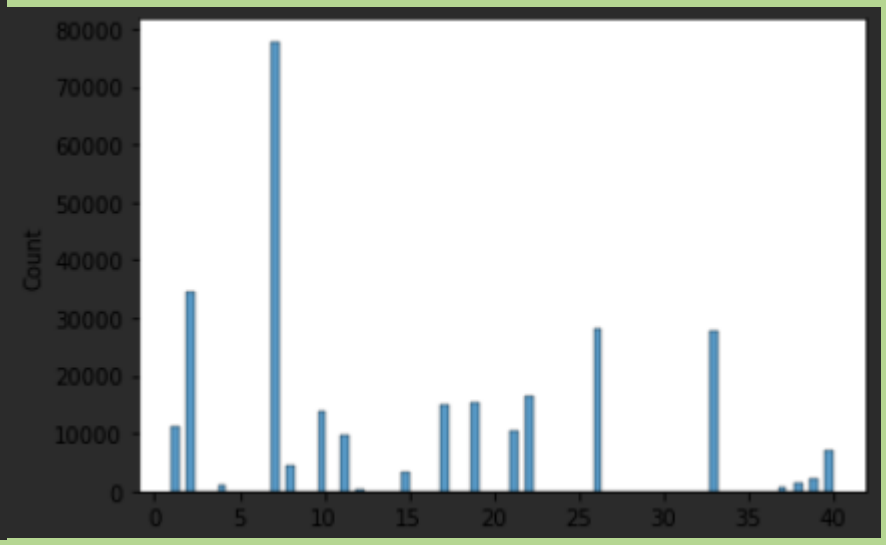
원본 훈련데이터 9000개

label	sentence
0	4438
1	92
2	553
3	44
4	679
5	186
6	231
7	209
8	164
9	115
10	815
11	58
12	45
13	11
14	45
15	335
16	23
17	169
18	7
19	4
20	136
21	84
22	56
23	79
24	103
25	52
26	8
27	36
28	9
29	5
30	12
31	15
32	26
33	67
34	15
35	27
36	11
37	3
38	18
39	9
40	1
41	5



외부 훈련데이터 280,510개

label	sentence
1	11205
2	34615
4	875
7	77850
8	4566
10	13779
11	9650
12	124
15	3436
17	15039
19	15402
21	10461
22	16601
26	28037
33	27757
37	551
38	1591
39	2062
40	6910

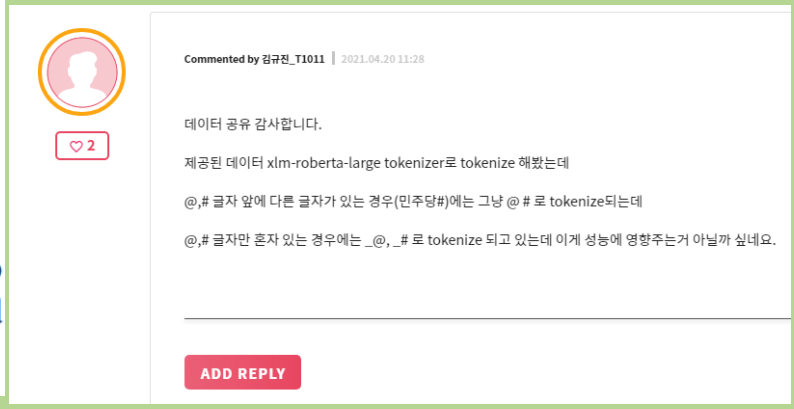


제출 day 3

https://paperswithcode.com/paper/an-improved-baseline-for-sentence-level/review/

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	`[E1-PERSON]` Bill `[/E1-PERSON]` was born in `[E2-CITY]` Seattle `[/E2-CITY]`.	**71.5**	**73.0**	71.0
Typed entity marker (punct)	@ * person * Bill @ was born in # ∧ city ∧ Seattle #.	70.9	72.7	**74.5**

Table 1: **Test F_1 (in %) of different entity representation methods on TACRED.** For each method, we also provide the processed input of an example sentence *"Bill was born in Seattle"*. Typed entity marker (original and punct) significantly outperforms other methods.



	C1
1	sentence
2	영국에서 사용되는 스포츠 유틸리티 @ α ARTIFACT α 자동차@의 브랜드로는 # β ORGANIZATION β 랜드로버#(Land Rover)와 지프(Jeep)가 있으며, 이
3	선거에서 # β ORGANIZATION β 민주당#은 해산 전 의석인 230석에 한참 못 미치는 57석(지역구 @ α QUANTITY α 27석@, 비례대표 30석)을 획득하는
4	# β ORGANIZATION β 유럽 축구 연맹#(@ α ORGANIZATION α UEFA@) 집행위원회는 2014년 1월 24일에 열린 회의를 통해 2017년 대회부터 UEFA U-2
5	용병 @ α CIVILIZATION α 공격수@ 챠디의 부진과 시즌 초 활약한 # β PERSON β 강수일#의 침체, 시즌 중반에 영입한 세르비아 출신 용병 미드필더 오
6	# β LOCATION β 람캄행# 왕은 1237년에서 1247년 사이 수코타이의 왕 @ α 0 α 퍼쿤 씨 인트라팃@과 쓰엉 부인 사이의 셋째 아들로 태어났다.

제출 day 3

- 외부데이터 약 280,000개를 그대로 훈련데이터 9,000개에 추가해서 했더니 LB 65%로 나왔다.
- Loss smoothing 0.5로 해봤는데 val acc은 별로 변화가 없었지만 val loss만 커졌다. 당연한 거지만 성능이 안높아졌으니 안썼음.
- 토론 게시판에 pororo library 를 이용하여 entity에 NER(named entity recognition)을 붙인 데이터를 사용했더니 val acc이 잘 나온 checkpoint가 있어서 그 checkpoint를 저장하기 위해 다시 훈련했으나 도저히 다시 안나와서 포기.

label	sentence	entity_01	entity_02	ratio	mul
0	4438	4438	4438	0.493111	1775.2
1	92	92	92	0.010222	36.8
2	553	553	553	0.061444	221.2
3	44	44	44	0.004889	17.6
4	679	679	679	0.075444	271.6
5	186	186	186	0.020667	74.4
6	231	231	231	0.025667	92.4
7	209	209	209	0.023222	83.6
8	164	164	164	0.018222	65.6
9	115	115	115	0.012778	46.0
10	815	815	815	0.090556	326.0
11	58	58	58	0.006444	23.2
12	45	45	45	0.005000	18.0
13	11	11	11	0.001222	4.4
14	45	45	45	0.005000	18.0
15	335	335	335	0.037222	134.0
16	23	23	23	0.002556	9.2
17	169	169	169	0.018778	67.6
18	7	7	7	0.000778	2.8
19	4	4	4	0.000444	1.6
20	136	136	136	0.015111	54.4
21	84	84	84	0.009333	33.6
22	56	56	56	0.006222	22.4
23	79	79	79	0.008778	31.6
24	103	103	103	0.011444	41.2
25	52	52	52	0.005778	20.8
26	8	8	8	0.000889	3.2
27	36	36	36	0.004000	14.4
28	9	9	9	0.001000	3.6
29	5	5	5	0.000556	2.0
30	12	12	12	0.001333	4.8
31	15	15	15	0.001667	6.0
32	26	26	26	0.002889	10.4
33	67	67	67	0.007444	26.8
34	15	15	15	0.001667	6.0
35	27	27	27	0.003000	10.8
36	11	11	11	0.001222	4.4
37	3	3	3	0.000333	1.2
38	18	18	18	0.002000	7.2
39	9	9	9	0.001000	3.6
40	1	1	1	0.000111	0.4
41	5	5	5	0.000556	2.0

원본 훈련데이터에서 원하는 비율만큼 각 label에서 빼가서 사용

외부데이터

label	sentence
1	11205
2	34615
4	875
7	77850
8	4566
10	13779
11	9650
12	124
15	3436
17	15039
19	15402
21	10461
22	16601
26	28037
33	27757
37	551
38	1591
39	2062
40	6910

제출 day 4

- 최적의 파라미터 (batch_size 64, epoch 10, warm_up step 300) 을 찾아내서
- 원본데이터로만 학습할 때 LB는 78.7%
- 외부데이터를 원본데이터 개수에 비율에 맞춰서 제출하기로 함.
  - 0.7 배율만큼 추가한걸로 학습하니 LB 76.0%
  - 0.3 배율만큼 추가한걸로 학습하니 LB 78.8%
  - 0.4 배율만큼 추가한걸로 학습하니 val acc이 80% 넘어 최고로 잘나왔지만 LB는 제출제한으로 확인하지 못함.
- Entity 앞뒤로 잘 안쓰는 기호 (α,β) 만 붙여봤는데 LB 78.6%이 나옴. 할려면 제대로 ner로 해야하는 듯.
- 원본데이터로만 학습한 모델 + 0.3배율 학습 모델 + 0.4배율 학습모델(*1.1) + 기호 (α,β) 학습모델 = 이 4개의 모델로 앙상블해서 최종점수 LB 80.2%