Project. 2

TEAM 3조 벤쿠버
이현지, 김응진, 유한솔, 김나현, 김준철

목차

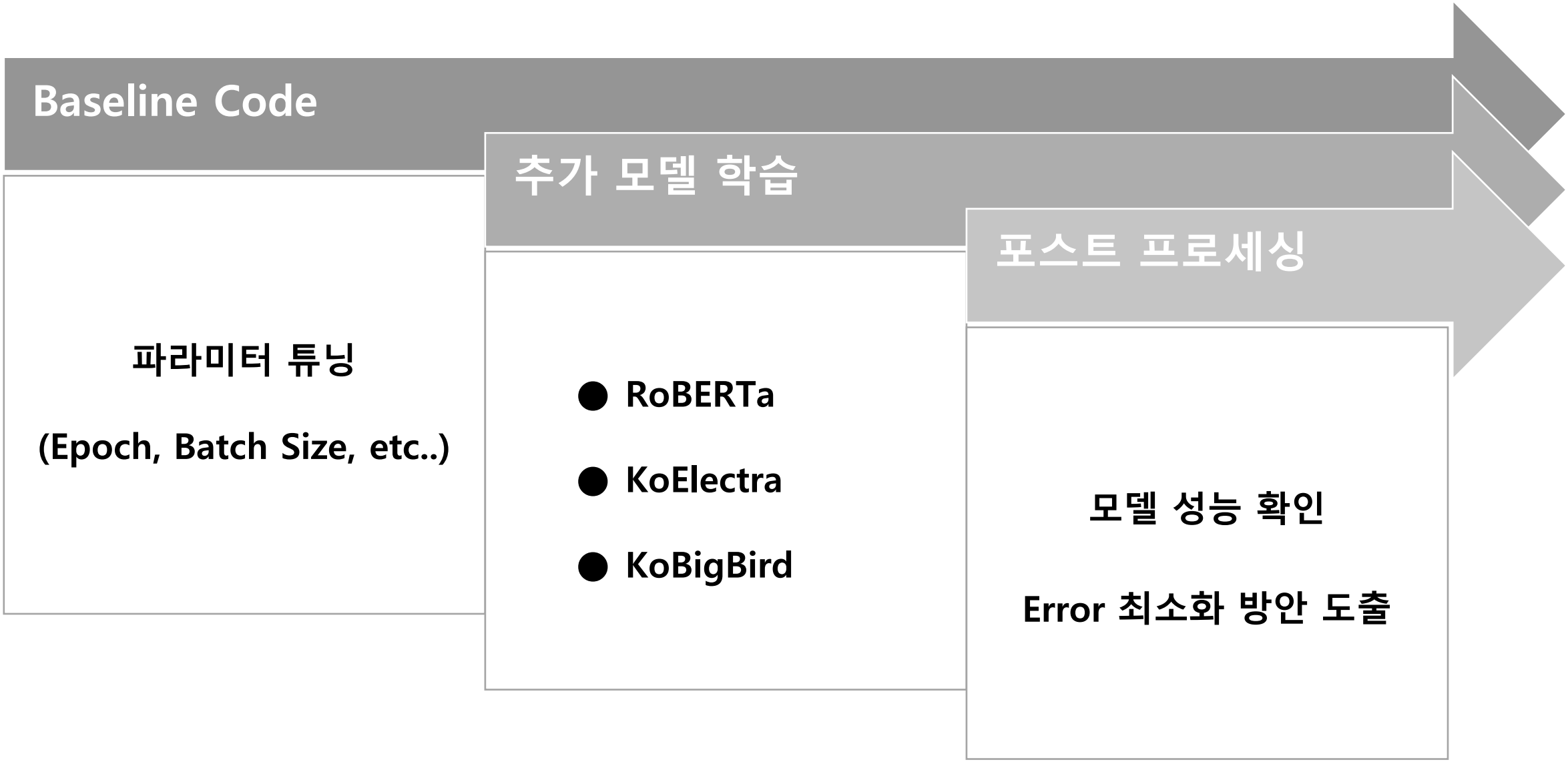
1. 프로젝트 개요

- Klue data의 Machine Reading Comprehension Task 수행
  - 주어진 문단을 학습 후, context 내 존재하는 예상 답안 도출
  - 언어 : 한국어
  - 데이터 타입 : JSON
  - 크기 : Train(17663), Test(4008)
- Levenshtein Distance 기반의 점수 산출

2. 프로젝트 팀 구성 및 역할

훈련생	역할	담당업무
이현지	팀장	- 노션 페이지 관리 - AI-HUB 데이터셋 추가
김응진	팀원	- 프로젝트 기획 (모델 등) 관리 - 자체평가 Metric 확보
유한솔	팀원	- Q&A 대비 - 긴 context 문제 해결
김나현	팀원	- 프로젝트 발표 - 추가 모델 자료 조사
김준철	팀원	- 발표 자료 제작 - 추가 모델 적용
공동 역할 : 모델 개선을 위한 정보 공유 및 코딩		

3. 프로젝트 진행 프로세스



4-1. 프로젝트 결과 (Baseline)

	A	B	C	D	E	F	G	H
1	Baseline		Batch Size 변경(split=0.1)				Data split 변경(BS=64	
2	Id	실제 답안	BS=16	BS=64	BS=32	BS=128	split = 0.25	split=0.2
3	d14cb73158624cf	뉴 740 Li 25주년	'BMW 코리아 25	뉴	'BMW 코리아 25	'BMW 코리아 25주	말라카이트	1988년
4	906631384e9149	독일 뒤셀도르프	150명의 사망자를	의도적으로 여객	장치를 확인한 결	스페인 바르셀로나	스페인 바르셀로	24일
5	35e61dcb479643	페이스북	18개의 질문으로	미국	6가지 특징으로	페이스북 등 정보가	3위, 5위를 차지하	18개의 질문으로
6	075e761b370040	페이스북	18개	"톱 50 가운데 22	50만명의 직장인	소통할 수 있다는	11월부터 지난달	50 가운데 22개가
7	e67ed38f3dd944	마드리드	백혈병 재단'을 설	백혈병	치료비	카탈루냐	마드리드에 위치	백혈병
8	80cacfdfe76442b	국제 원자재 값 하	서형종 대신증권	36.5% 감소했다.	1804억원) 대비	7939	1조3836억원으로	50% 미만으로 하
9	78a80cca941c463	5조원	지난 15일 발표한	25일 코레일에 '각	25%에서 57%로	1조4000억원을 수	부대사업을 추진	25일 코레일에 '각
10	5c19b9781f8a4f0	운영허가 인증서	숙박시설을 이용	숙박시설에서는	'운영허가 인증서'	인증서'를 발급 받	인증서를 받은 숙	외국인 배우자 등
11	d2844b7141cb4a	중합시켜 사슬 모	무명천	37	평면 구조를 넓게	1분자마다 평균 10	산류	폴리실록산은 실
12	be39b91f52a04b	노르웨이	무르만스크	독일군	무르만스크	노르웨이	노르웨이	모스크바 평화 조
13	98fe72c173f642e	22개	338억원	23개 제품의 포장	'착한 포장' 프로젝	338억원	338	7개 들어 있던 마
14	1756f8a643124b5	필리핀 바콜로드		시우, 다온' 7인으	Dustin		1004	더스틴(DUSTIN)'
15	1f1ef6951050490	김교성	〈에라 좋구나〉, 전	〈	주현미	서양음악을 접했	〈황성의 적〉이 큰	태평양 전쟁 종전
16	acbc4be5f9d04b	79달러	'아쿠아수르스 스	37달러에서 36달	'브라이트닝 클렌	'브라이트닝 클렌징	'아쿠아수르스 스	'아쿠아수르스 스
17	393657048abd43	존 위클리프	에라스무스	라틴어	에라스무스	에라스무스는《우	에라스무스	교회의 운영은 성
18	08ee8dae9f6744c	얀 후스	우신예찬	에라스무스	에라스무스	에라스무스	에라스무스는《우	교회
19	0fd8aca8b9844e3	지기스문트	교황이 존재하는	1409	1409	1377	1377	로마에서는 교황
20	fa773f05fd224cb	복정역	'위례 아이파크 1	'위례 아이파크 1	'위례 아이파크 1	'위례 아이파크 1차	'위례 아이파크 1	'위례 아이파크 1
21	89810801a2f446e	한살림	경기	김현종	변호사 업계에 구	다져온 한결이 이합	26일자	법칙'이란 별칭을
22	30a8f6b62fa1422	한예종 부설 한국	이경선	1977년 경희대 음	거른	스코티시	갈까	10여년 전부터는
23	4957b7070c3a46	로사다운자켓	'샤벳다운자켓'	'피어론	'샤벳다운자켓'	벳다운자켓'은 후드	'로사다운자켓'	'샤벳다운자켓'은
24	28b135ddc9374d	킨프로스 선	폭탄으로 라돔프	표의 재물을 탐내	무역	해 일을 때 라돔프	털어 상품을 사서	무역을 하기 위해

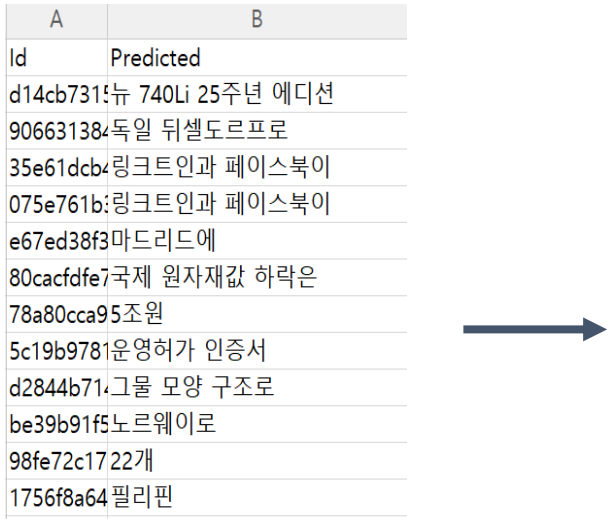
- **Kaggle Score : 183 점**
- **제대로 된 학습이 이루어 지지 않음**
- **Data split 변경으로 개선되는 모습에도 여전히 높은 오답률을 확인**

4-2. 추가 모델 적용

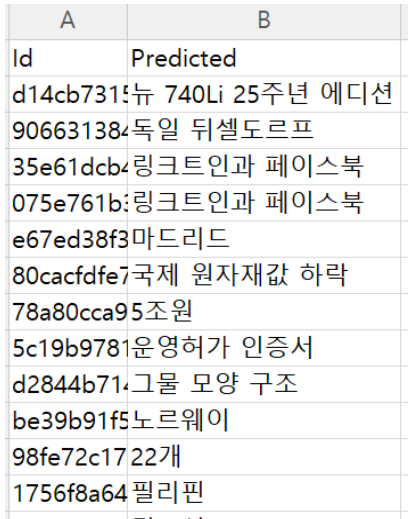
모델명	적용 배경	결과 (Kaggle Score)
BERT (Baseline)	• Pre-train 후 Fine-tuning • Masked Language Modeling (random masking) • Deep bi-direction	183.23
RoBERTa	• 큰 배치사이즈 • 긴 시퀀스 • Masking 동적 할당	4.14
KoBigBird	• Sparse attention (BERT의 8배의 토큰(512→4096) 처리) • 한국어 데이터로 학습 (모두의 말뭉치, 위키, Commom crawl 등)	4.17
KoElectra	• 정확도 및 학습 효율성 개선 모델 • MASK 외에도 모든 토큰 학습 • Replaced Token Detection (generated token → discriminator → real or fake token → train)	3.69

4-3. Post-Processing

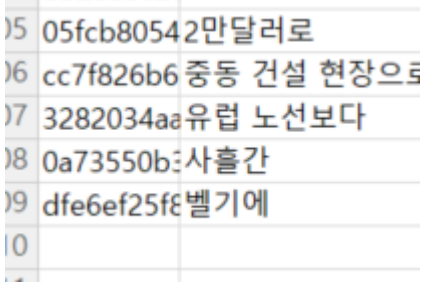
- 대상 : Ko-Electra
- 방식 : 조사 및 어미 후처리
  - 3글자 (까지다 ,이었다, 한다고, 이라고)
  - 2글자 (에서, 으로, 이나, 에게, 까지 ,처럼, 라고, 였다, 이다, 부터, 이면, 에는, 되고, 보다)
  - 1글자 (이, 로, 과, 와, 의, 는, 은, 을, 를, 에, 가, 도, 께)
- 결과 : Kaggle score 3.69점 → 7.77점



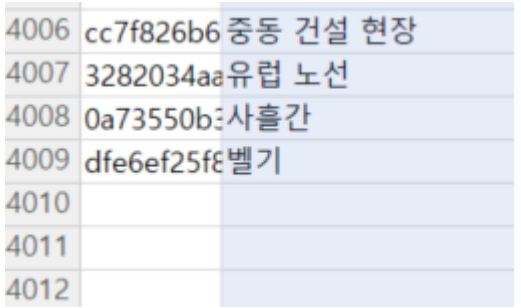
	A	B
1	Id	Predicted
2	d14cb7315	뉴 740Li 25주년 에디션
3	90663138	독일 뒤셀도르프로
4	35e61dcb	링크트인과 페이스북이
5	075e761b	링크트인과 페이스북이
6	e67ed38f3	마드리드에
7	80cacfdfe7	국제 원자재값 하락은
8	78a80cca95	조원
9	5c19b9781	운영허가 인증서
10	d2844b71	그물 모양 구조로
11	be39b91f5	노르웨이로
12	98fe72c17	22개
13	1756f8a64	필리핀



	A	B
1	Id	Predicted
2	d14cb7315	뉴 740Li 25주년 에디션
3	90663138	독일 뒤셀도르프
4	35e61dcb	링크트인과 페이스북
5	075e761b	링크트인과 페이스북
6	e67ed38f3	마드리드
7	80cacfdfe7	국제 원자재값 하락
8	78a80cca95	조원
9	5c19b9781	운영허가 인증서
10	d2844b71	그물 모양 구조
11	be39b91f5	노르웨이
12	98fe72c17	22개
13	1756f8a64	필리핀



5	05fcb8054	2만달러로
6	cc7f826b6	중동 건설 현장으로
7	3282034aa	유럽 노선보다
8	0a73550b3	사흘간
9	dfe6ef25f8	벨기에
10		



4006	cc7f826b6	중동 건설 현장
4007	3282034aa	유럽 노선
4008	0a73550b3	사흘간
4009	dfe6ef25f8	벨기
4010		
4011		
4012		

4-4. 결과

모델명	결과 (Kaggle Score)
BERT (Baseline)	183.23
KoElectra	3.69
KoElectra (Post-Processed)	7.77
RoBERTa	4.14
KoBigBird	4.17

- 모델 변경을 통한 성능 개선

  (183.23 → 3.69)

- BERT 대비 roBERTa의 높은 성능 확인

- 모든 input token을 학습한 Electra의 다른 모델 대비 높은 성능 확인

5. 자체 평가 및 보완

- GPU 이슈로 인한 한계점
- 1차 프로젝트와는 다른 접근 (다양한 모델 적용)
- 포스트 프로세싱의 복잡성

6. 팀별 공통 의견

- 모델 조사 과정에서의 연구 트렌드 탐구 기회

Q & A

감사합니다

Appendix. 참고문헌

- *Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova(24 May 2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding 1810.04805.pdf (arxiv.org)*

- *Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov(26 Jul 2019), RoBERTa: A Robustly Optimized BERT Pretraining Approach*

  https://arxiv.org/pdf/1907.11692.pdf

- *Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning, ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS ,ICLR 2020*

  https://openreview.net/pdf?id=r1xMH1BtvB

- *딥러닝 모델 압축 방법론과BERT 압축 , March 6, 2020*

  https://blog.est.ai/2020/03/%EB%94%A5%EB%9F%AC%EB%8B%9D-%EB%AA%A8%EB%8D%B8-%EC%95%95%EC%B6%95-%EB%B0%A9%EB%B2%95%EB%A1%A0%EA%B3%BC-bert-%EC%95%95%EC%B6%95/