Between Spaces

#Bert #한국어 전처리 #띄어쓰기 #아버지가방에들어가신다

NLP-03
Bumblebee

boostcamp ai tech

목차

1. Overview

5W1H – Why

아버지가방에들어가신다

아버지 가 방 에 들어가 신다
NNG JKS NNG JKB VV EP+EC

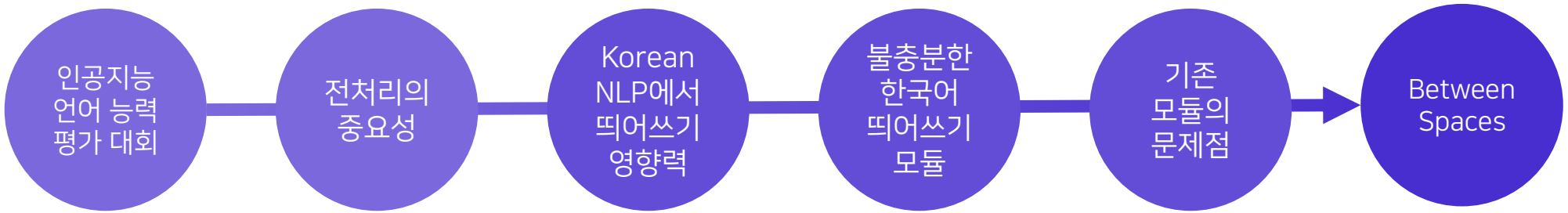
아버지가방에들어가신다
NNP

아버지 가방 에 들어가신 다
Noun Noun Josa Verb Eomi

Mecab	
NNG	일반명사
JKS	주격 조사
JKB	부사격 조사
VV	동사
EP+EC	선어말 어미+ 연결 어미

Komoran	
NNP	일반명사

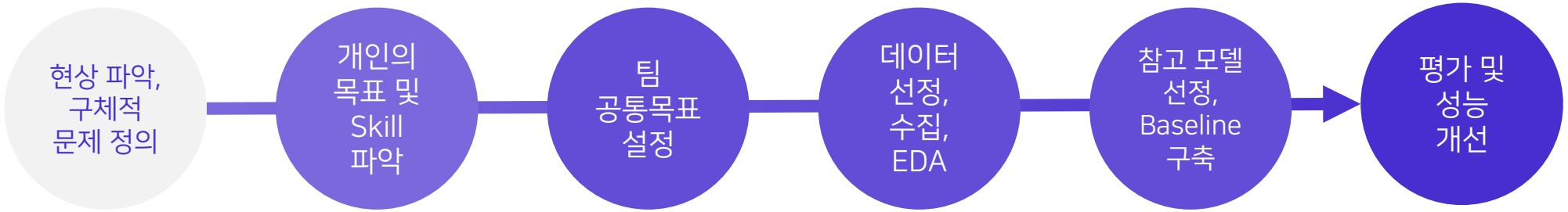
Twitter	
Noun	명사
Josa	조사
Verb	동사
Eomi	어미



5W1H – How

1. 현상 파악

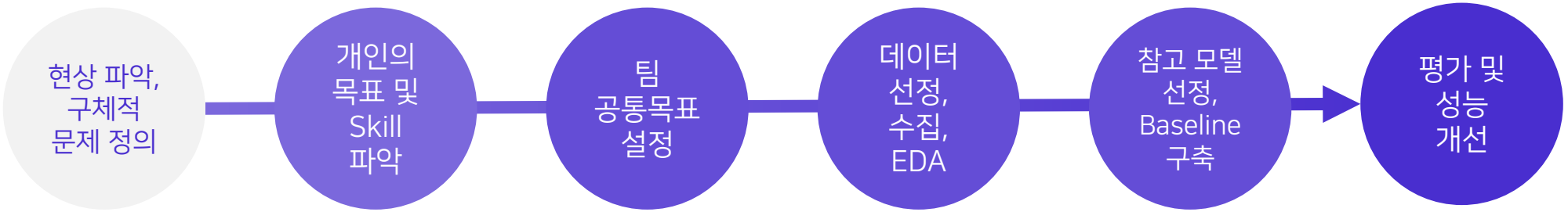
- **어떤 일이 발생하고 있는가?**
  - 룰베이스, 비표준어(신조어, 인터넷), 어순, 빈번한 문장 오류, 교정 필요성
- **해당 일에서 어려움은 무엇인가?**
  - 모든 룰 체크, 얕은 도메인지식(문법, 기호, 규칙 다양), noise train, 룰 변화
- **해당 일에서 해결하면 좋은 것은 무엇인가?**
  - 데이터로 룰 해결(지도학습), 자소서 문법 판정, 신조어 예외처리, 문맥 고려, newly-fashioned words, 타 언어 확장, 지도학습 모델



5W1H - How

1. 현상 파악

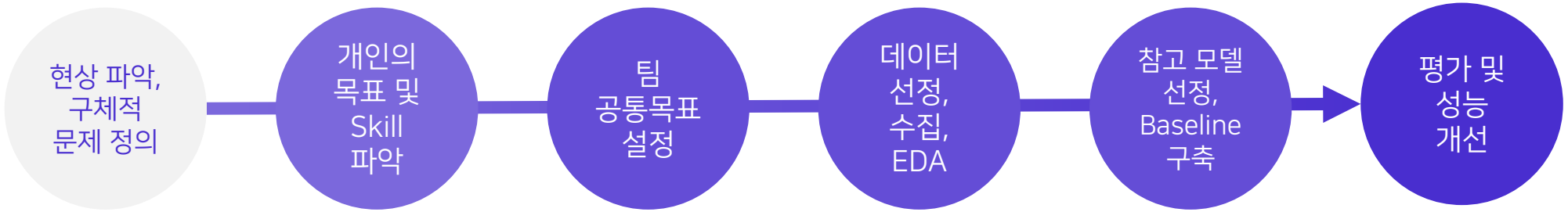
- **추가적으로 무엇을 해볼 수 있을까?**
  - 타 언어 적용, 어플 for foreigner, 가독성 안 좋은 문장 해결, 선호 높은 문구(정제된 문장), 신조어 판단, 단어 교정, 존댓말, rule-base 한계
- **어떤 가설을 만들어 볼 수 있을까?**
  - 정제된 문장으로 교정, 언어 역사성 수치 확인
- **어떤 데이터가 있을까?**
  - 기존 데이터 출처(말뭉치, aihub, exobrain), 기존 데이터 수정



5W1H – How

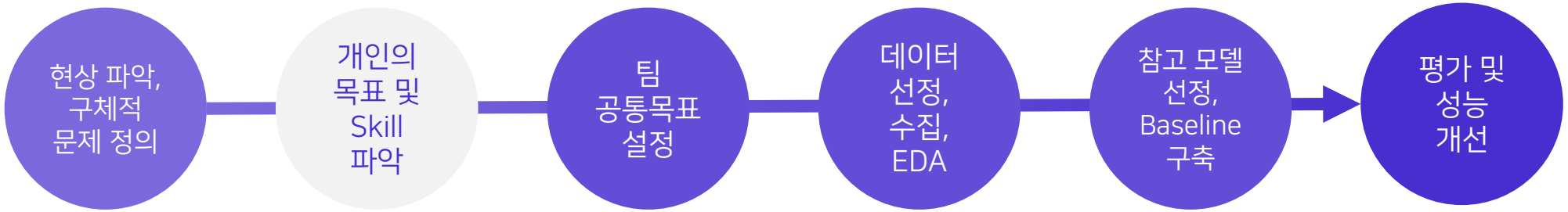
2. 구체적 문제 정의

- **무엇을 해결하고 싶은가?**
  - 룰베이스 한계 극복, 글 단위 한 번에 해결, 교정
- **무엇을 알고 싶은가?**
  - 룰베이스와의 차이점, 구문 어순 오류 차이, 타 언어 확장



5W1H - How

1. 각자 이번 프로젝트에서 달성하고 싶은 최종 목표와 각각의 우선순위
2. 원하는 팀 & Role
3. 하루에 할애할 수 있는 시간
4. 본인이 할 수 있는 Skill Set



5W1H – How

Purpose (우리 팀 공통의 목표)

달성 기준 baseline

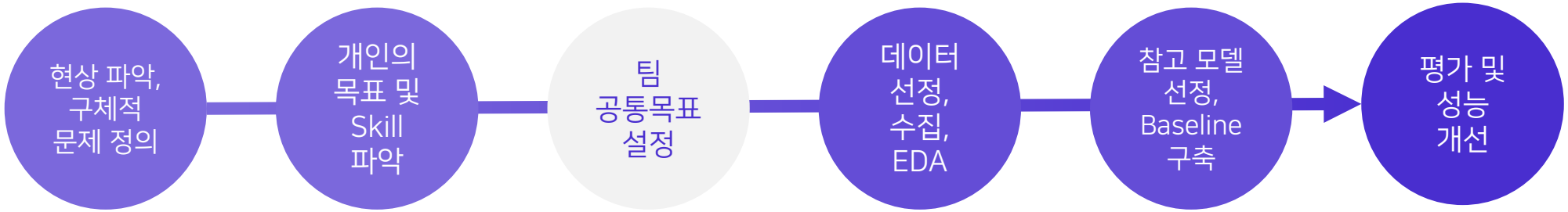
데이터 측면: 10만 쌍 이상의 train 데이터 사용
모델 측면: Accuracy 95%
테스트 측면: 100글자당 Inference 0.1s 이내
서빙 측면: 웹으로 구축, streamlit 사용

우선 순위

1. 모델 구축하기
2. 서빙
3. 데이터 / 테스트

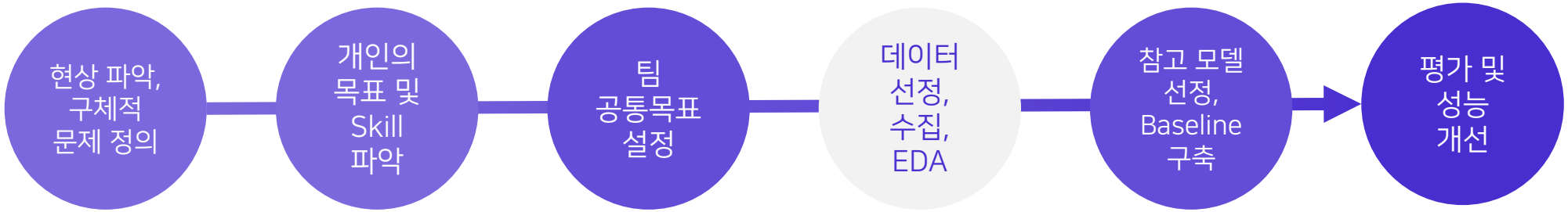
기간: 2021.12.09~2021.12.20.

목표 모델: 띄어쓰기 correction 모델



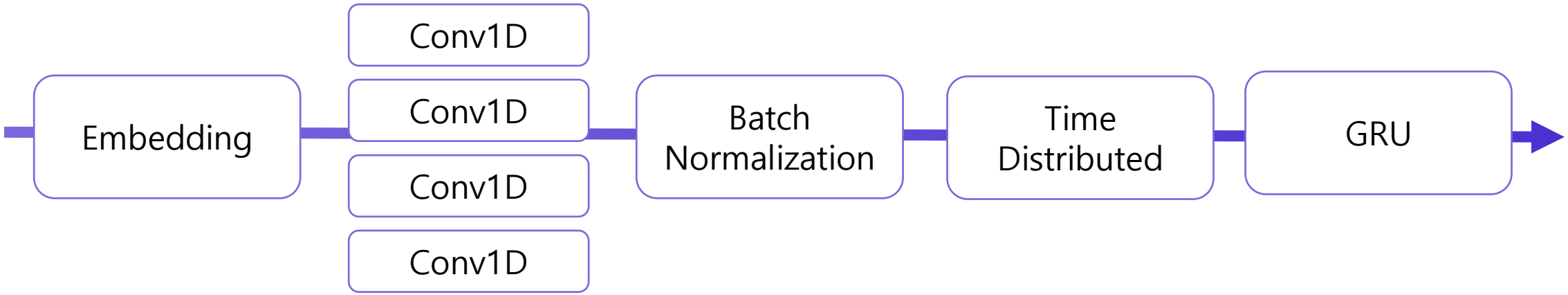
5W1H - How

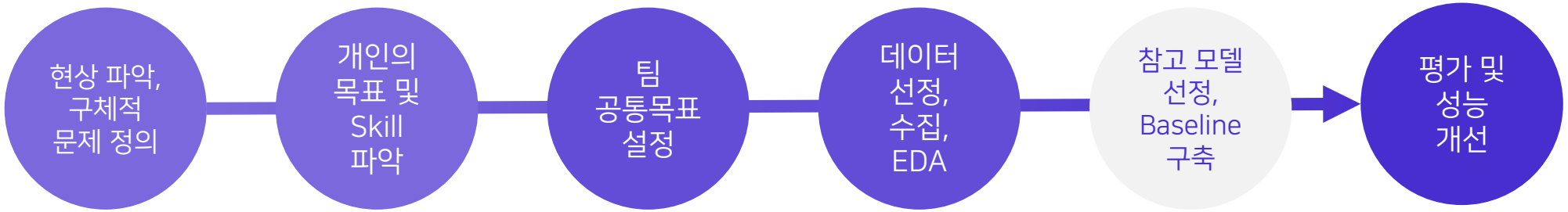
우선순위	데이터 이름	데이터 코멘트	데이터 개수
1	문법성 판단 말뭉치	문법적으로 올바른 데이터로만 이루어져 있어 추가 데이터 정제가 많지 않아 초기 사용 데이터로 적합함	8,983 문장
2	신문 말뭉치	신문 기사 원문 자료 수집 및 정제된 데이터	2,693,991 문장
3	개체명 분석 말뭉치 2020	데이터에서 문장만 쓰기에 좋아보임	대화 223,962 문장 신문 150,082 문장
4	비출판물 말뭉치	테스트 데이터 (교정을 거치지 않음. 시(동시), 일기, 편지 글, 소설(동화), 감상문, 기타)	2,100,000 어절



5W1H – How

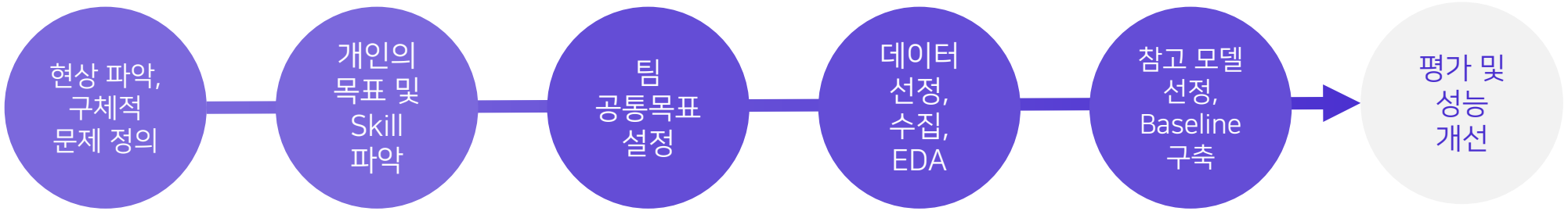
- 성능 기준 참고 모델: KoSpacing





5W1H – How

- Metrics
  - $\boldsymbol{Accuracy} = \frac{\#\ of\ correct\ BI\ taggings}{\#\ of\ chracters}$
  - $\boldsymbol{Boolean\ Accuracy} = (prediction == ground\ truth)$
  - $\boldsymbol{F1\ Score} = \frac{2*(precision*recall)}{(precision+recall)}, Precision\ and\ recall\ are\ calculated\ by\ tag\ 'B'$
  - $\boldsymbol{Word\ Error\ Rate(WER)} = \frac{\#\ of\ Substitutions+\#\ of\ Deletions+\#\ of\ Insertions}{\#\ of\ Words\ in\ the\ ground\ truth}$



5W1H – How

- Metrics 계산 예시

Ground Truth

안녕 나는 범블비야 반가워

B I B I B I I I B I I

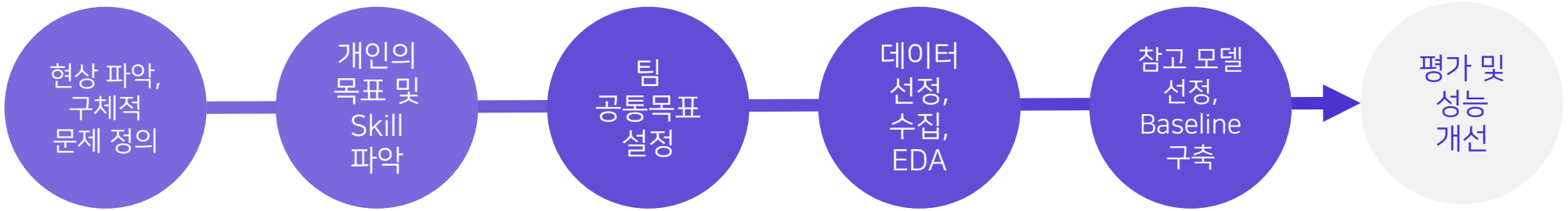
Prediction

안녕 나는 범 블비야반 가워

B I B I B B I I I B I

- $Accuracy = \frac{8}{11}$
- $Boolean\ Accuracy = 0$
- $F1\ Score = \frac{2}{3}$
- $Word\ Error\ Rate = \frac{3}{4}$

- 성능 개선
  - Improve accuracy
  - Decrease inference time

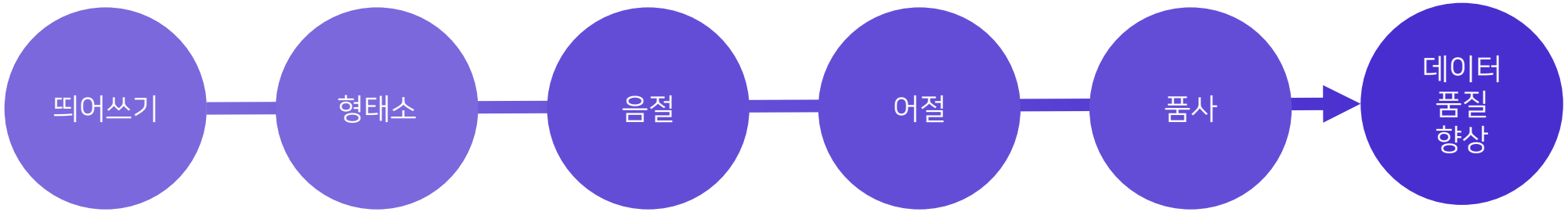


5W1H – What

- 띄어쓰기 교정 모델
- Web Page
- Accuracy > 95 %
- 100글자 당 < 0.1s

5W1H – Where

- 기대 효과
  - 한국어 전처리 및 후처리
    - 띄어쓰기는 형태소 분석 이전 반드시 수행해야 하는 중요 전처리 과정
    - Speech To Text 후처리
  - Data augmentation
  - 데이터 품질 향상



5W1H - Who

Data

Web, Data 전처리 및 제작
#풀스택 #디자인을곁들인

Data 전처리 및 제작
#아이디어뱅크 #분위기메이커

Data 전처리 및 제작
#은둔고수 #속전속결코딩

Modeling

Modeling 총괄, Fine tuning
#코딩왕 #모델링의신 #갓다인

PM, Modeling
#정리왕 #인간스케쥴러 #발표자

Modeling
#공유왕 #팀장님 #추진력

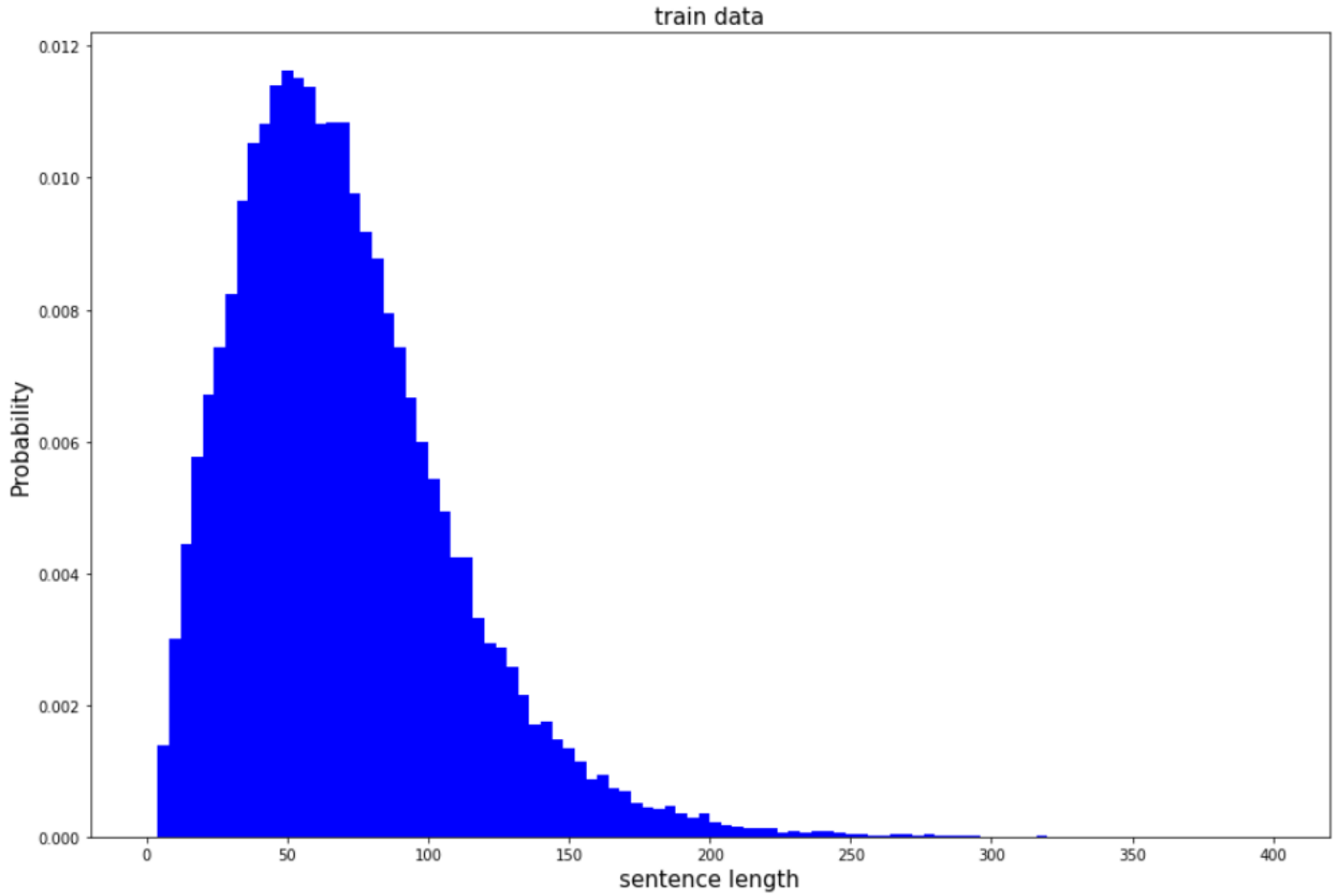
2. Data

EDA

- 특이 데이터
  - 시상식 소감
    - " MBC 드라마를 어렸을 때부터 쭉 지켜봤었는데 아 저 상을 받으면 얼마나 기분이 좋을까 근데 제가 받았습니다. 예, 기분이 되게 좋네요. 아, 무엇보다 저는 제가 사랑하는 연기를 하고 있습니다. 어, 그 연기, 그 사랑하는 연기를 이렇게 해 가지고 상도 받고 돈도 벌고 있어서 너무나도 시청자 여러분들께 책임감으로 앞으로도 조금 더 자만하지 않고 열심히 노력하는 그런 배우가 되도록 하겠습니다."
  - 기사 제목
    - 대구, 시흥 이어 안산도...홍역 전국 동시다발 '비상'
    - '확 달라진 베트남 축구' 박항서 매직은 현재 진행형
    - '홈' 창원에서 3년만에 왕좌 되찾은 3점왕과 덩크왕
  - 유니코드 및 특수문자
    - '₩u200b티웨이항공은 지난해 총 13개의 신규 노선을 개척했으며, 5대의 신규 항공기를 도입해 420만명의 국제선 이용객을 수송해 '
    - '2017년 328만명보다 90만여명 증가하는 실적을 거뒀다.₩n'

EDA

- 문장 길이
  - mean 69
  - min 5
  - max 327



EDA

- 말뭉치마다 다른 규칙
  - 온점 구분
    - 1: "그러던 어느 날, 하녀는 부자와 낙타 관리인이 집을 비운 사이 아기 낙타를 풀어 주었어요."
    - 2: "아기 낙타는 눈물을 흘리면서 다시 엄마 낙타가 떠난 쪽으로 달려갔어요."
  - 문단 구분
    - 1: "화자를 처음 만나 이야기를 들으러 왔다고 하자 서슴없이 꺼낸 첫 이야기이다. 화자로서 가장 쉽게 기억해낸 이야기인 셈이다. 설화 앞뒤에 교훈적 해석을 덧붙이고 있음은 화자의 습관화된 태도의 한 모습이기도 하다. 어려서 조모로부터 들었다고 했다."
  - 기준 없음
    - 1: "어 표현을 하고 들을 때는 저 사람이 무슨 말 어떤 식으로 자기를 표현하든 간에 그 말 뒤에 있는 그 사람의 느낌과 저 사람이 진정으로 원하는 게 뭘까를 들어 주는"
    - 2: "그런 태도. 그래서 마음과 마음이 연결되는"

EDA

Train/Val/Test	데이터 이름	데이터 개수	실제 사용한 문장 개수
8:1:1	문법성 판단 말뭉치	8,983 개	8,983 문장
8:1:1	신문 말뭉치	2,693,991 개	317,516 문장
8:1:1	개체명 분석 말뭉치 2020	대화 223,962 개 신문 150,082 개	대화 78,041 문장 신문 146,842 문장
Test only	비출판물 말뭉치	2,100,000 어절	56,881 문장

- **사용된 문장 수**
  - For Training: 441,106 문장
  - For Validation: 55,138 문장
  - For Test: 112,019 문장
  - 총 608,263 문장

달성 기준 baseline

[√] 데이터 측면: 10만 쌍 이상의 train 데이터 사용

[ ] 모델 측면: Accuracy 95%

[ ] 테스트 측면: 100글자당 Inference 0.1s 이내

[ ] 서빙 측면: 웹으로 구축, streamlit 사용

Data 전처리

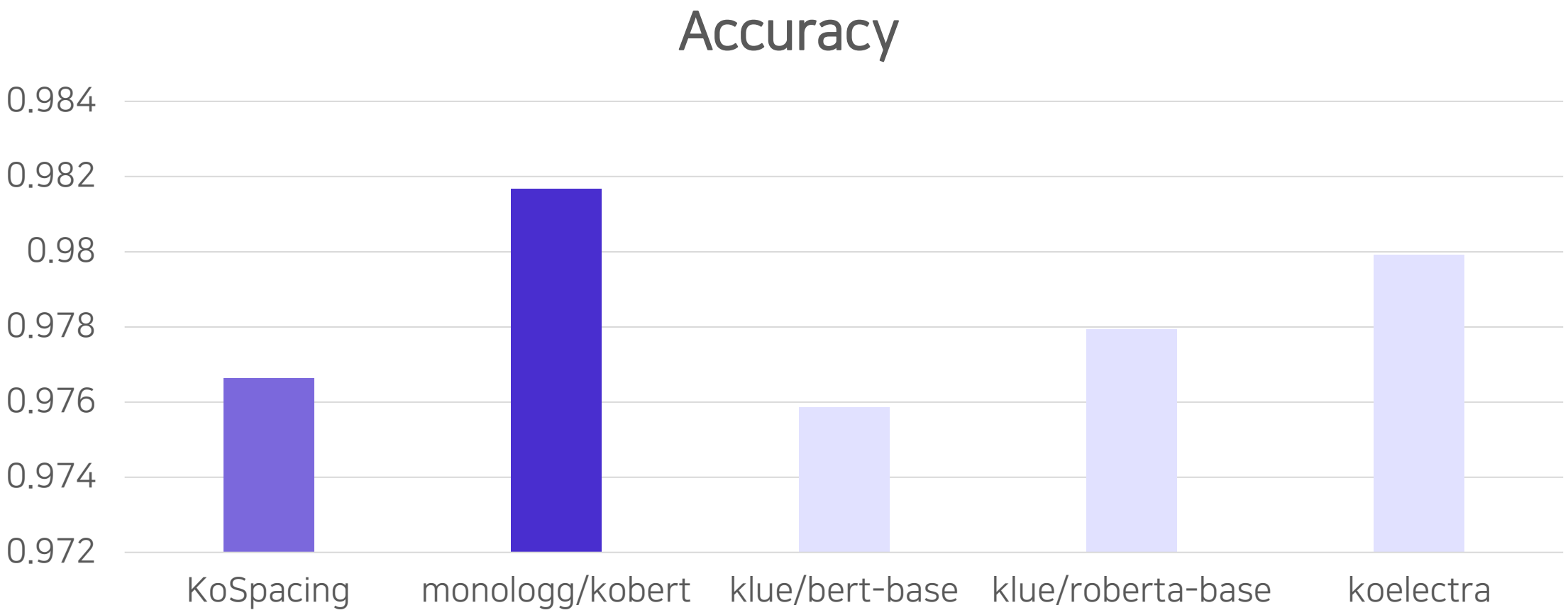
- 전처리 Issue 및 관련 합의 사항
  - 이름 마스킹
    - '화자A', '김모씨', '윤xx' : 그대로 사용
    - 'nameN' : '김xx' 형태로 변환
  - 특수문자
    - 슬래시 /, 겹괄호 ((,)) 등의 특수문자 삭제
    - 단괄호 (,) 는 유지
    - 유니코드 삭제
  - 외래어, 외국어 표기
    - 영문 및 외국어 표기, 외래어 유지

Data 전처리

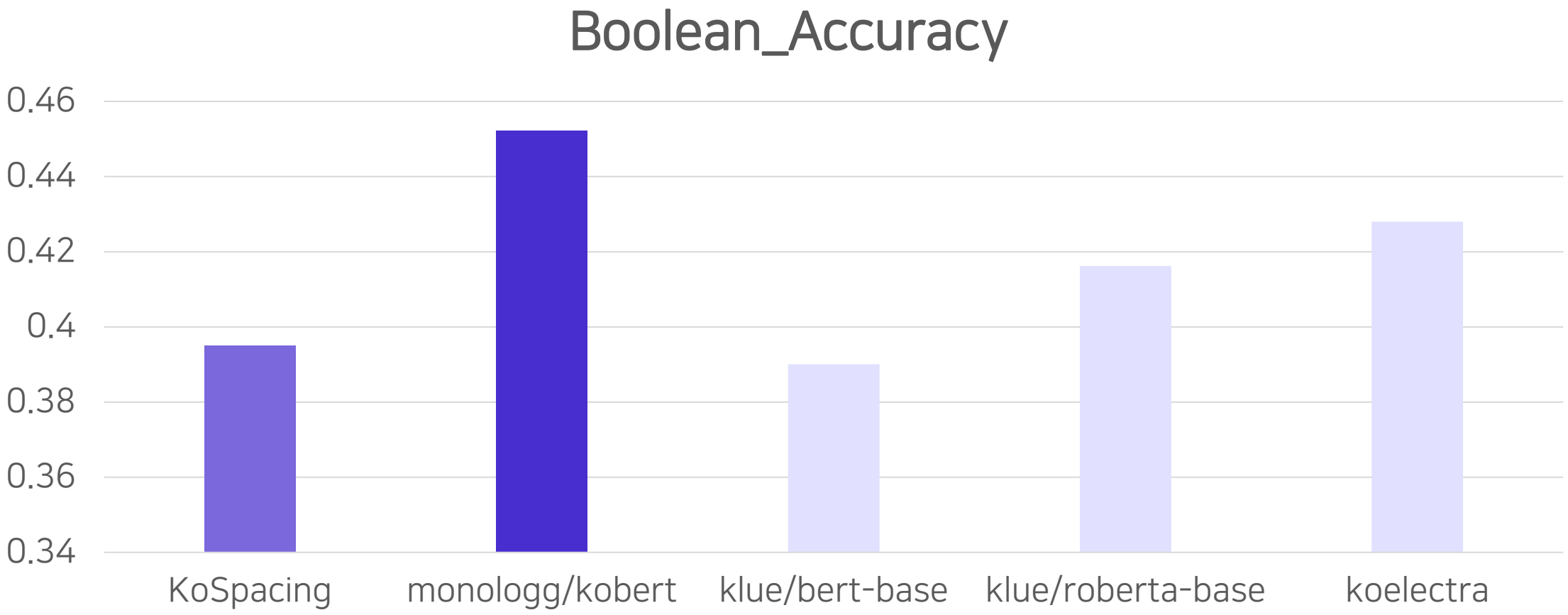
- **공백 삽입 방법**
  - 문법성 판단 말뭉치
    - 문장 길이가 상대적으로 짧아 4개의 랜덤 위치에 공백 삽입
  - 신문 말뭉치
    - 문장의 길이의 1/3개의 랜덤 위치에 공백 삽입
  - 개체명 분석 말뭉치, 비출판물 말뭉치
    - 문장의 각 위치마다 0.4의 확률로 공백 삽입

2. DL Model

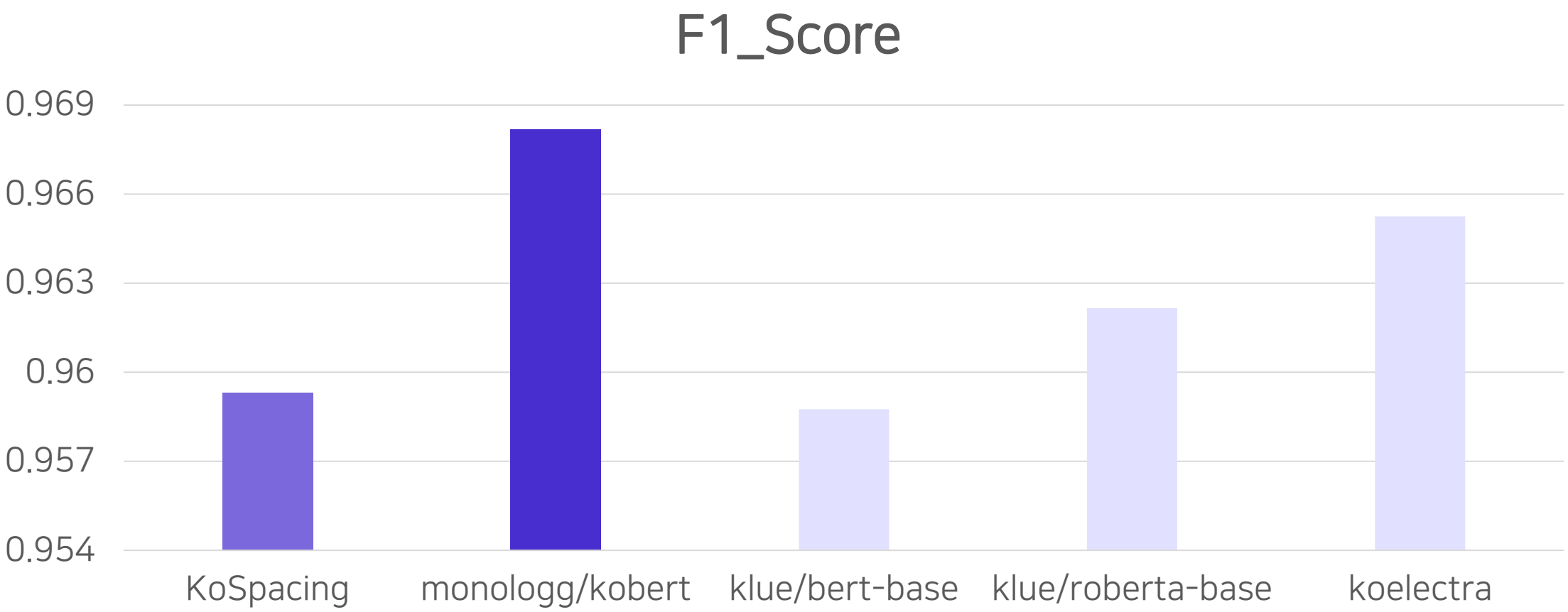
Backbone 선정


Accuracy
0.984
0.982
0.98
0.978
0.976
0.974
0.972
KoSpacing
monologg/kobert
klue/bert-base
klue/roberta-base
koelectra

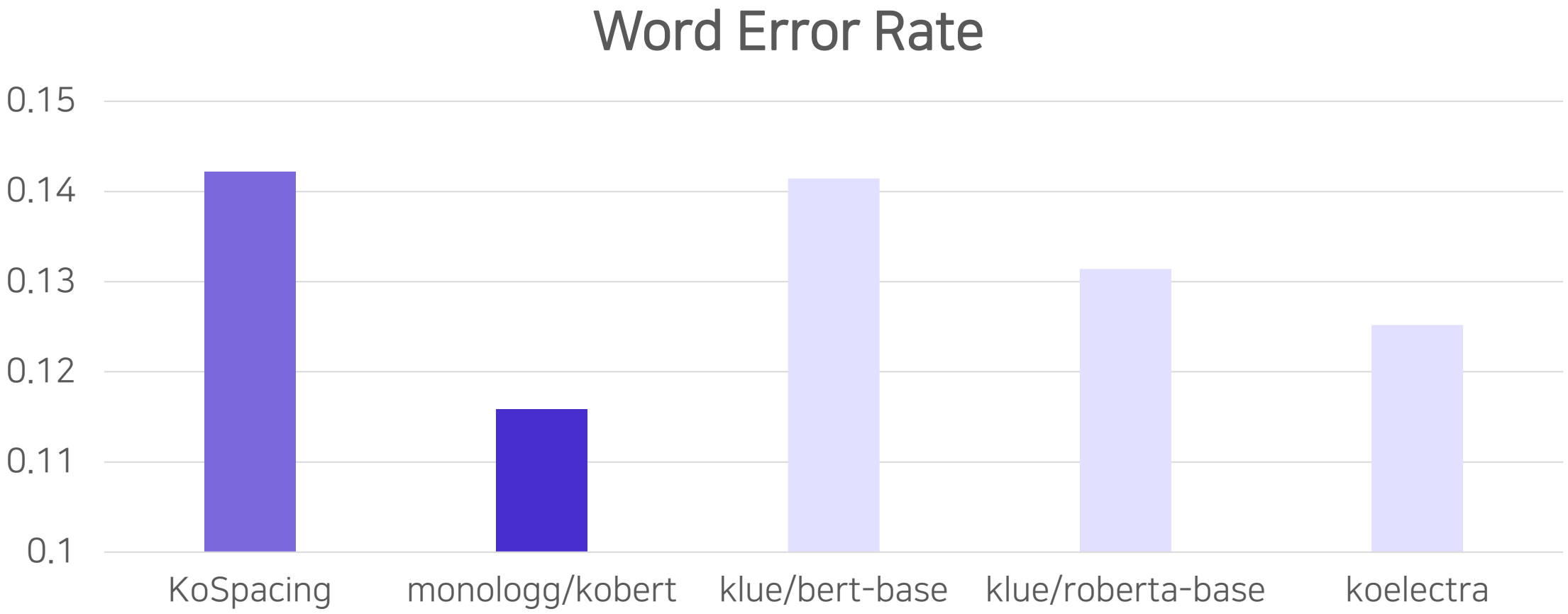
Backbone 선정


Boolean_Accuracy
0.46
0.44
0.42
0.4
0.38
0.36
0.34
KoSpacing
monologg/kobert
klue/bert-base
klue/roberta-base
koelectra

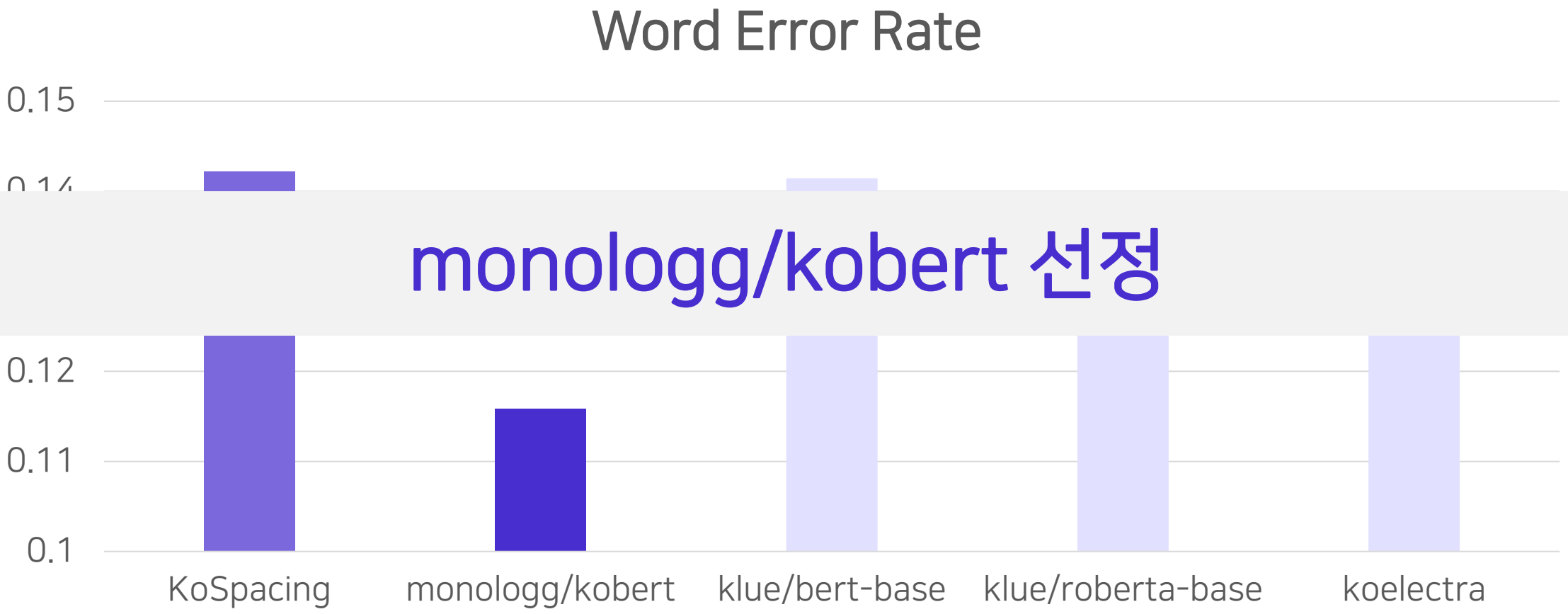
Backbone 선정


F1_Score
0.969
0.966
0.963
0.96
0.957
0.954
KoSpacing
monologg/kobert
klue/bert-base
klue/roberta-base
koelectra

Backbone 선정

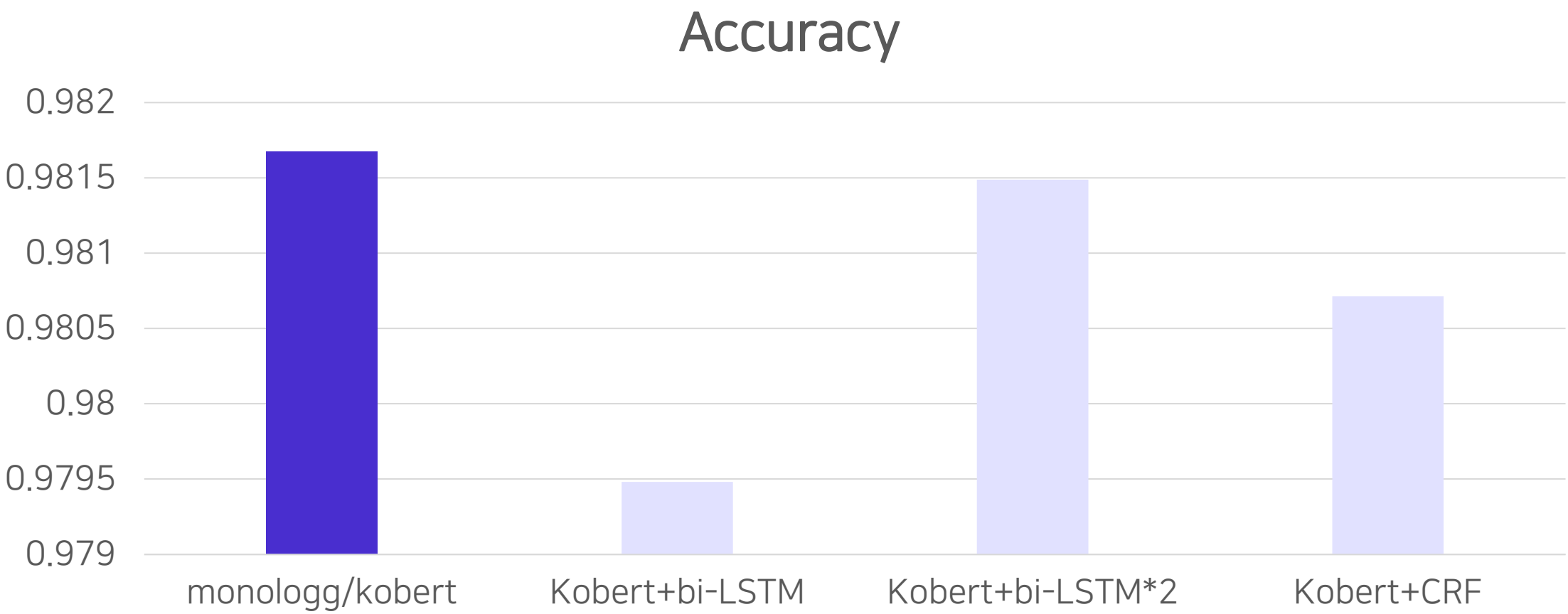

Word Error Rate
0.15
0.14
0.13
0.12
0.11
0.1
KoSpacing
monologg/kobert
klue/bert-base
klue/roberta-base
koelectra

Backbone 선정

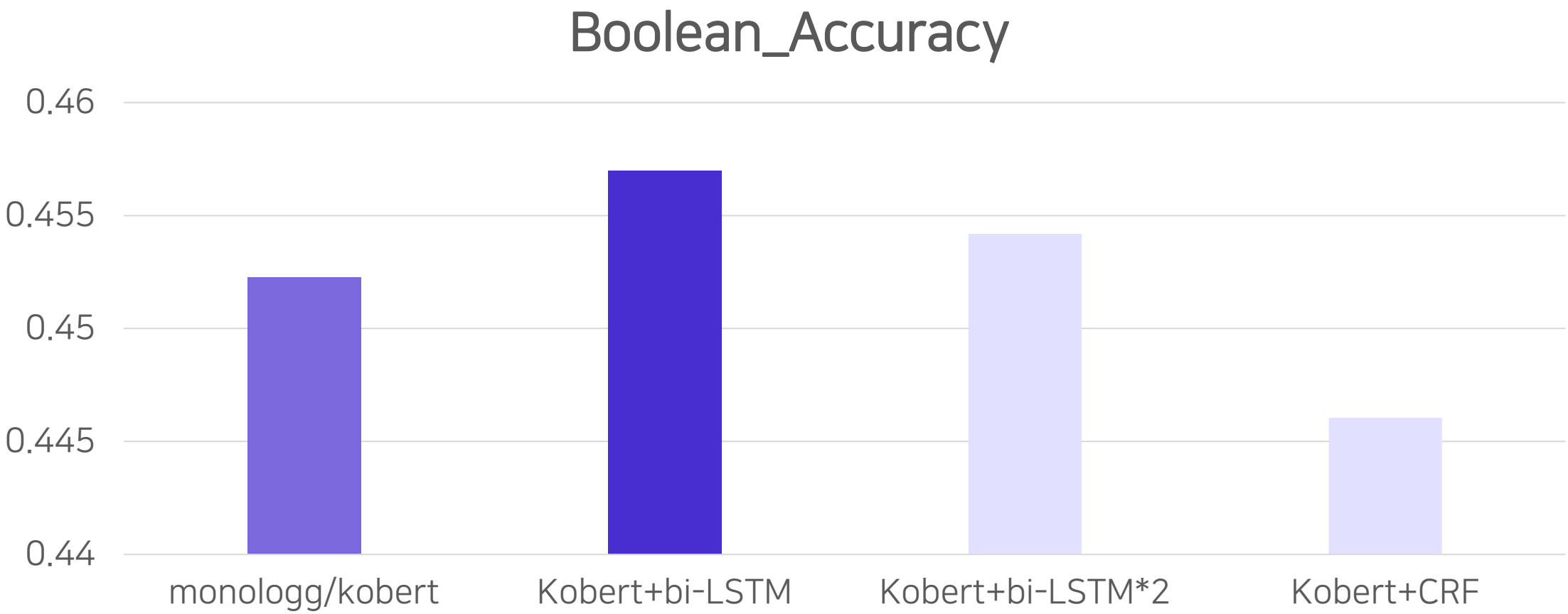


monologg/kobert 선정

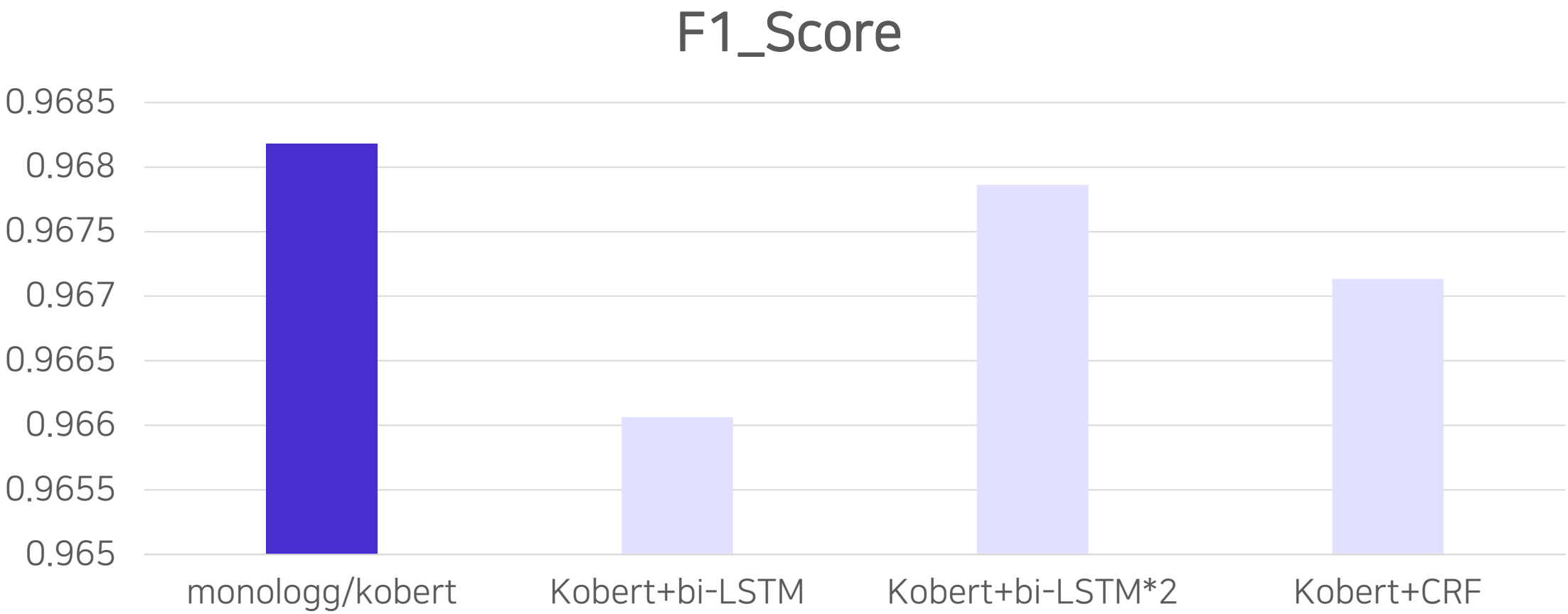
Layer 추가


Accuracy
0.982
0.9815
0.981
0.9805
0.98
0.9795
0.979
monologg/kobert
Kobert+bi-LSTM
Kobert+bi-LSTM*2
Kobert+CRF

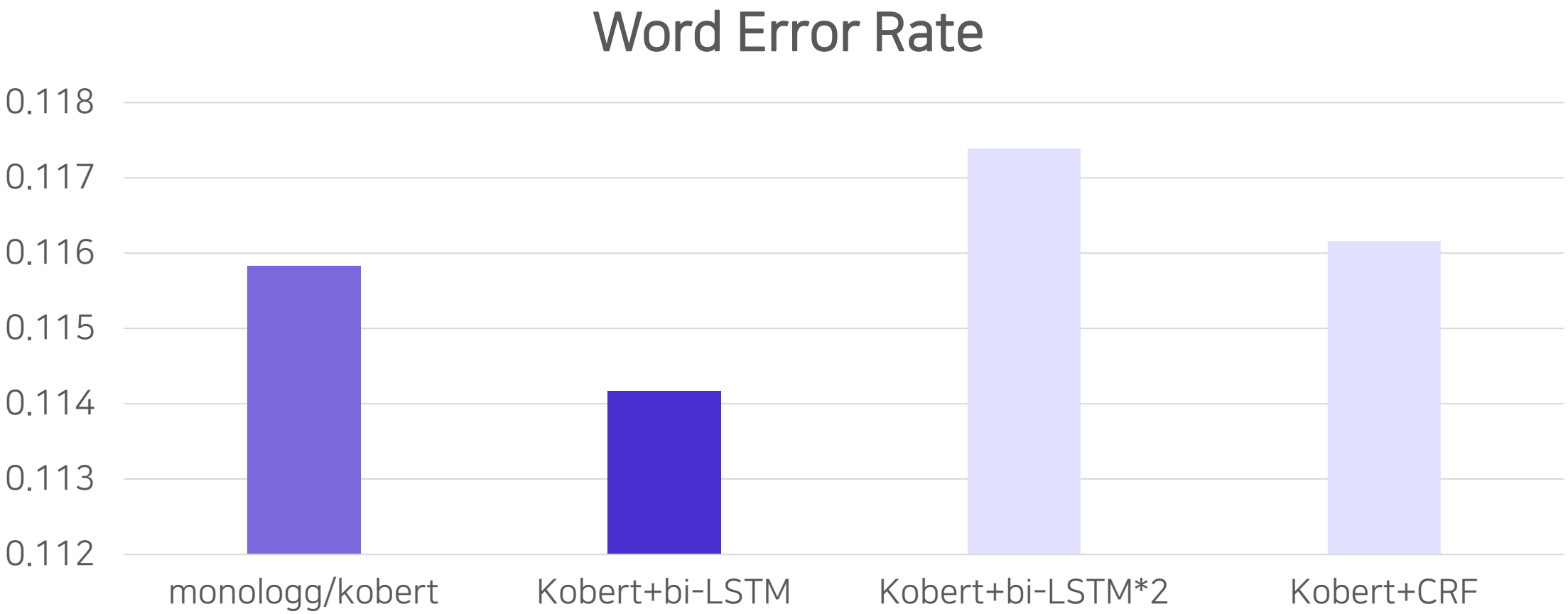
Layer 추가


Boolean_Accuracy
0.46
0.455
0.45
0.445
0.44
monologg/kobert
Kobert+bi-LSTM
Kobert+bi-LSTM*2
Kobert+CRF

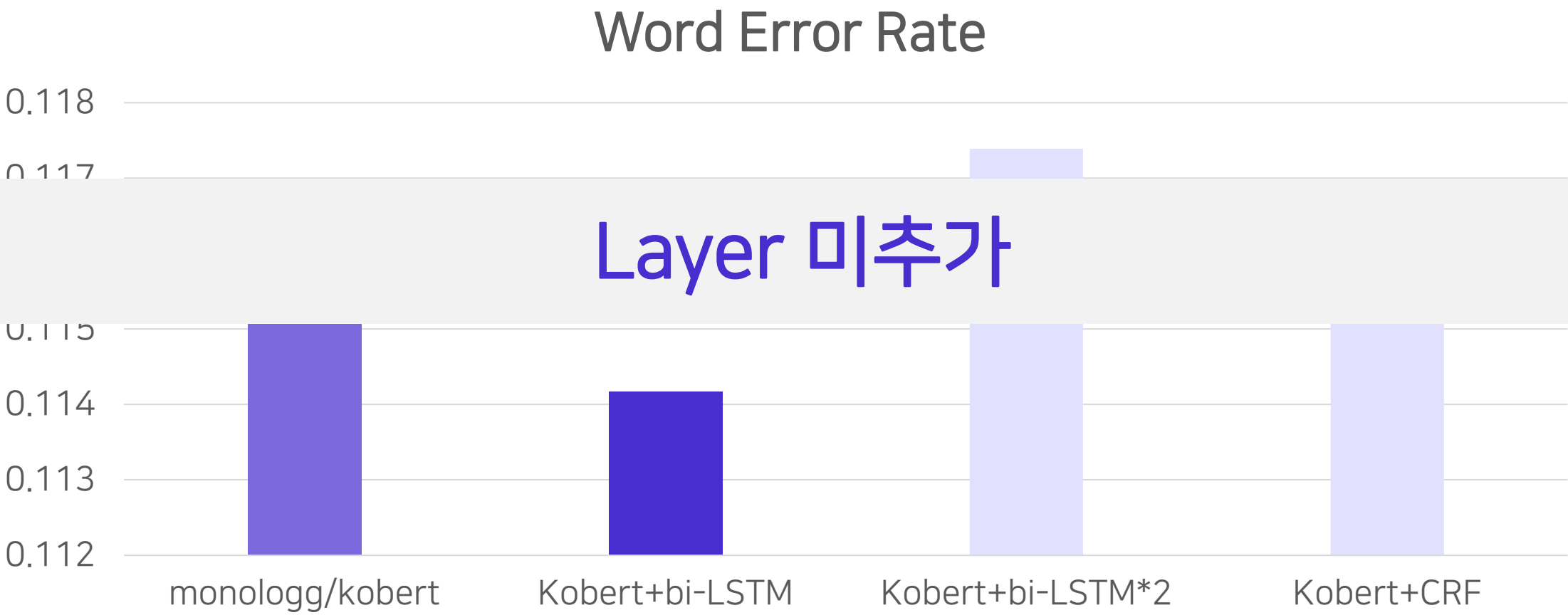
Layer 추가


F1_Score
0.9685
0.968
0.9675
0.967
0.9665
0.966
0.9655
0.965
monologg/kobert
Kobert+bi-LSTM
Kobert+bi-LSTM*2
Kobert+CRF

Layer 추가


Word Error Rate
0.118
0.117
0.116
0.115
0.114
0.113
0.112
monologg/kobert
Kobert+bi-LSTM
Kobert+bi-LSTM*2
Kobert+CRF

Layer 추가



Layer 미추가

문장이 길어지면?

- 문화재 사범들은 공소시효가 만료되기를 기다렸다가 경매업자를 통해 처분·유통하려고 했다. '숭례문 목판'은 1827년 경 양녕대군 후손들이 중각(重刻)한 것으로 **국보제1호** 숭례문의 편액 대자(大字)인 '숭례문'을 판각한 현존하는 유일의 목판본으로서 문화재적 가치가 매우 뛰어나다.' 후적벽부 목판'도 19세기 중반양녕대군의 **유묵으로 서인식되고** 판각됐던 자료라는 **점에서당 시**의 역사상을 살 필 수 있는 **중요한자료**이다.

-> 문장이 길어질수록, 특히 뒤에 위치한 글자들의 정확도 낮아짐

문장이 길어지면?

->Layer 추가, 글자 수에 따른 커리큘럼 러닝 효과 X

문장이 길어지면?

-> 그럼 문장을 쪼개보자!

쉼표 단위로 Split - 전

- 부산웹툰으로 채워지는 메인전시는 최근 2년간 발표된 히트작과 기대작을 전시하는 '세상 밖으로 나온 웹툰전(展)', 영화·드라마로 제작되었거나 제작 중인 작품들을 모아놓은 '영화(드라마)가 된 웹툰전(展)', 해외로 수출되거나 해외전시회 등에 소개된 '물건너(해외로) 간 웹툰전(展),**향 후선보일부산 웹툰의기 대작' 세상밖으로나 올** 웹툰전(展) '과 **웹툰 속캐릭터로만들어진피규어전시까지다채롭게꾸며진다.**

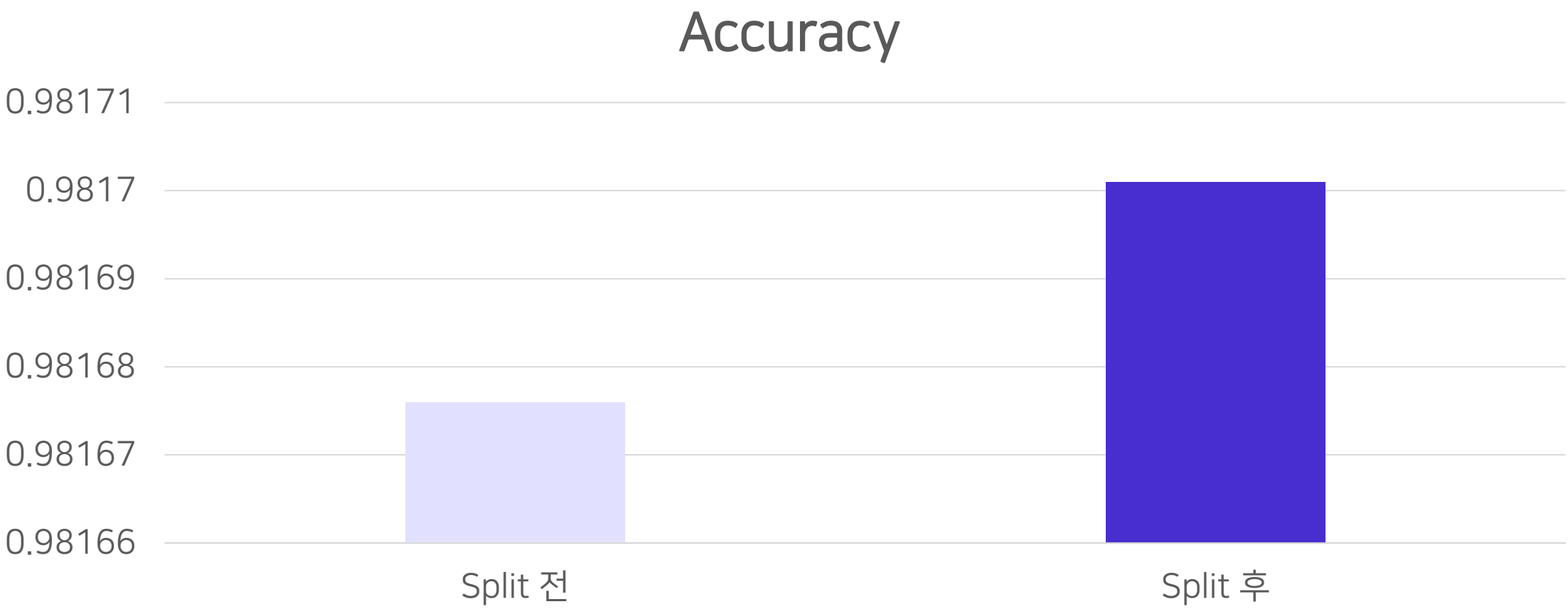
- Inference time : 0.0208 s

쉼표 단위로 Split - 후

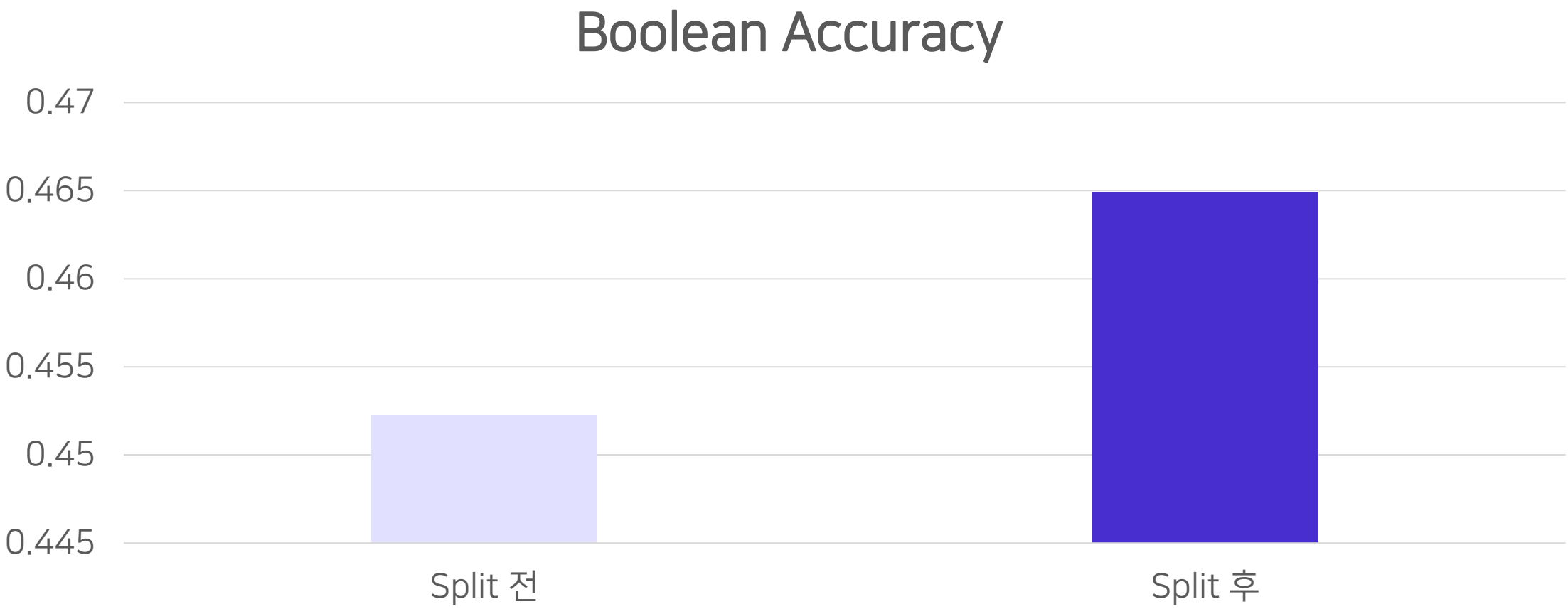
- 부산 웹툰으로 채워지는 메인전시는 최근 2년간 발표된 히트작과 기대작을 전시하는 '세상 밖으로 나온 웹툰전(展)', 영화·드라마로 제작되었거나 제작 중인 작품들을 모아놓은 '영화(드라마)가 된 웹툰전(展)', 해외로 수출되거나 해외전시회 등에 소개된 '물 건너(해외로) 간 웹툰전(展), 향후 선보일 부산 웹툰의 기대작 '세상 밖으로 나올 웹툰전(展)'과 웹툰 속 캐릭터로 만들어진 피규어 전시까지 다채롭게 꾸며진다.

- Inference time : 0.0528 s

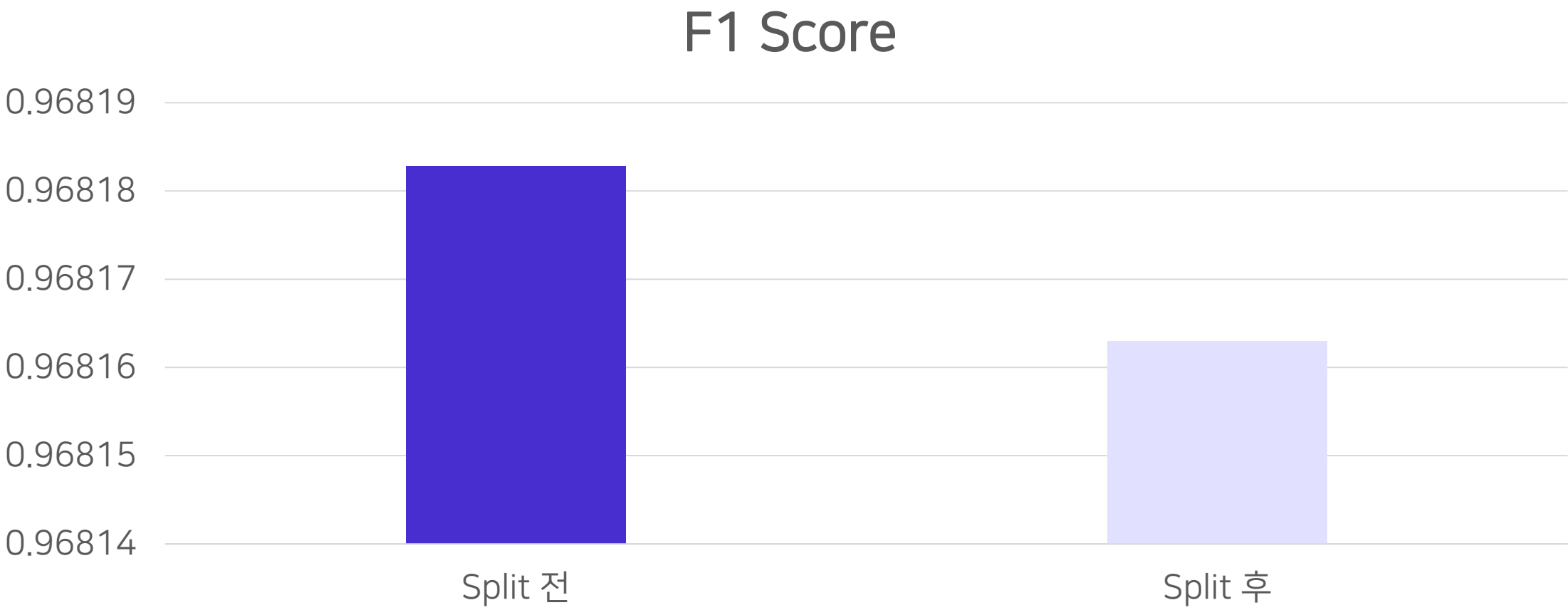
쉼표 단위로 Split – 성능 향상



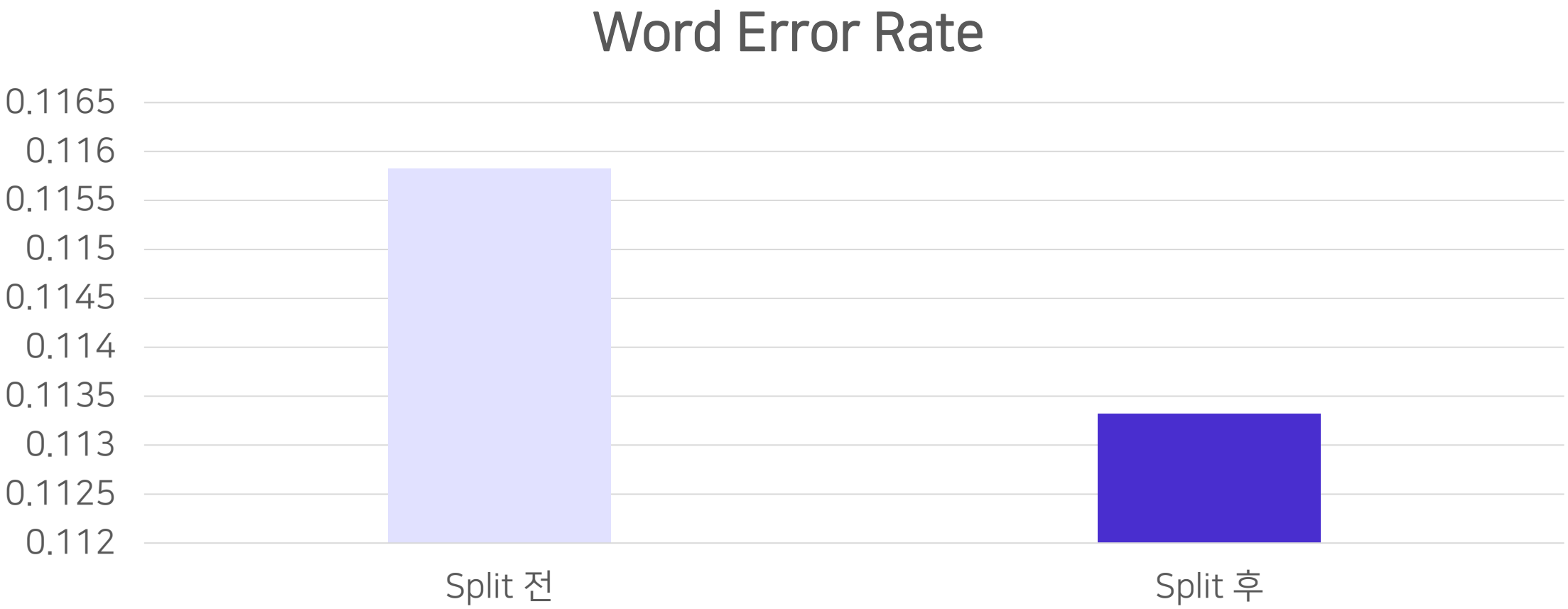
쉼표 단위로 Split – 성능 향상



쉼표 단위로 Split – 성능 향상



쉼표 단위로 Split – 성능 향상



Problem

- 쉼표 단위로 문장을 끊기 때문에, 쉼표의 개수에 따라 Inference 시간 차이가 큼

- 쉼표가 없는 문장의 Inference Time 변화
  - "제주도는 제주도립미술관 건립을 추진하면서 이 같은 인연이 있는 장 화백에게 작품 기증을 요청했다."
  - Inference Time: 0.0302 s -> 0.0297 s

- 쉼표가 많은 문장의 Inference Time 변화
  - 철수와 영희는 감자, 당근, 양파, 고춧가루, 닭다리살, 설탕과 소금을 사서 집으로 갔다.
  - Inference Time: 0.0263 s -> 0.0896 s

약 4배 증가

Solution

- 쉼표로 split 하되, 한 split 당 100자 이상으로 되도록 조정

- 쉼표가 많은 문장의 Inference Time 변화
  - 철수와 영희는 감자, 당근, 양파, 고춧가루, 닭다리살, 설탕과 소금을 사서 집으로 갔다.
  - Inference Time: 0.0263 s -> 0.0289 s

큰 차이 없음

Result

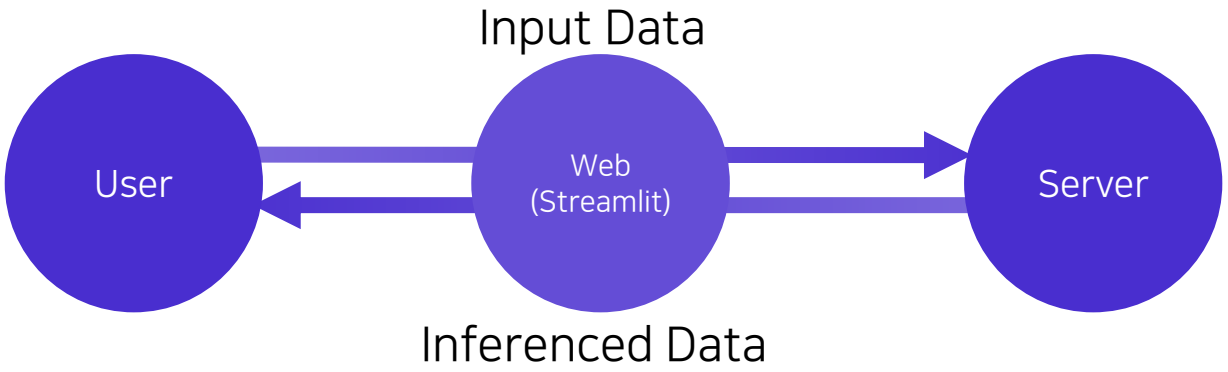
- Test set metric
  - Accuracy: 0.9844
  - Boolean Accuracy: 0.4676
  - F1 score: 0.9715
  - Word Error Rate: 0.1107

달성 기준 baseline

[√] 데이터 측면: 10만 쌍 이상의 train 데이터 사용
[√] 모델 측면: Accuracy 95%
[√] 테스트 측면: 100글자당 Inference 0.1s 이내
[ ] 서빙 측면: 웹으로 구축, streamlit 사용

3. Serving

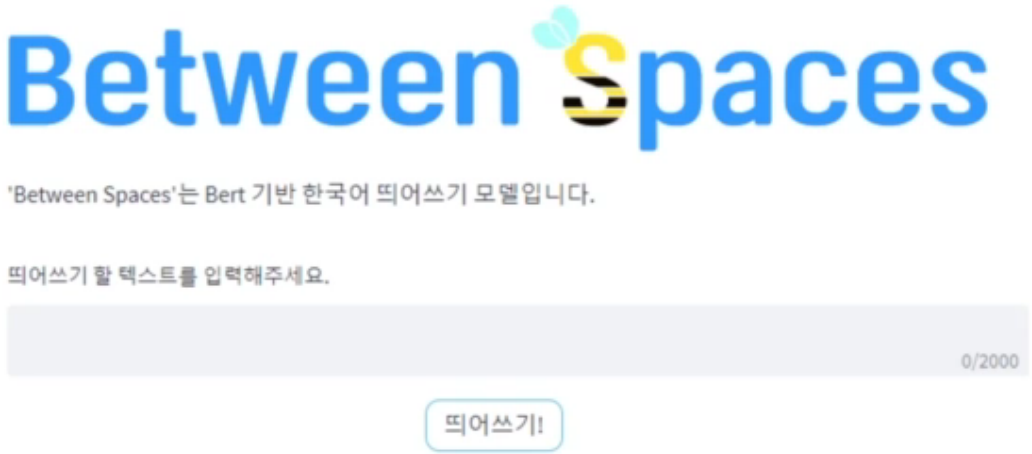
Web Page



달성 기준 baseline

[√] 데이터 측면: 10만 쌍 이상의 train 데이터 사용
[√] 모델 측면: Accuracy 95%
[√] 테스트 측면: 100글자당 Inference 0.1s 이내
[√] 서빙 측면: 웹으로 구축, streamlit 사용

시연 영상


Between Spaces
'Between Spaces'는 Bert 기반 한국어 띄어쓰기 모델입니다.
띄어쓰기 할 텍스트를 입력해주세요.
0/2000
띄어쓰기!

5. Q&A

6. Appendix

역할 및 소개

Data

Web, Data 전처리 및 제작
#풀스택 #디자인을곁들인

Data 전처리 및 제작
#아이디어뱅크 #분위기메이커

Data 전처리 및 제작
#은둔고수 #속전속결코딩

Modeling

Modeling 총괄, Fine tuning
#코딩왕 #모델링의신 #갓다인

PM, Modeling
#발표왕 #인간스케줄러 #서기

Modeling
#공유왕 #팀장님 #추진력

한 줄 소감

프로젝트 End-to-End를 훌륭한 팀원들과 함께 할 수 있어서 감사했습니다, Thanks to BumbleBee!

강진선

김다인

많은 것을 배울 수 있었던 기회였습니다. 범블비 팀원 분들 감사합니다.

직접 선택한 주제로 프로젝트를 진행하게 되어 새로운 경험을 쌓을 수 있었습니다

김민지

송이현

기획, 모델링, 서빙까지 프로젝트 하나를 처음부터 끝까지 끝마치게 되어 뿌듯하고, 그동안 동고동락한 팀원들에게 감사합니다.

그동안 학습한 내용을 바탕으로 프로젝트를 진행하며 많이 배울 수 있었고, 팀원분들께 감사합니다

이나영

신원지

부스트캠프를 통해서 많은 것을 배울 수 있었습니다. 팀원분들께 감사합니다.

References

- KoSpacing: http://freesearch.pe.kr/archives/4759
- Konlpy: https://konlpy-ko.readthedocs.io/ko/v0.4.3/morph/
- Komoran: https://docs.komoran.kr/
- 한국어 띄어쓰기 모듈: https://blog.naver.com/roootwoo/221590316102
- Images
  - https://www.researchgate.net/figure/tagtog-showing-entity-annotations-for-a-full-text-document-from-PubMed-Central_fig3_283835929

감사 합니다

감사합니다!