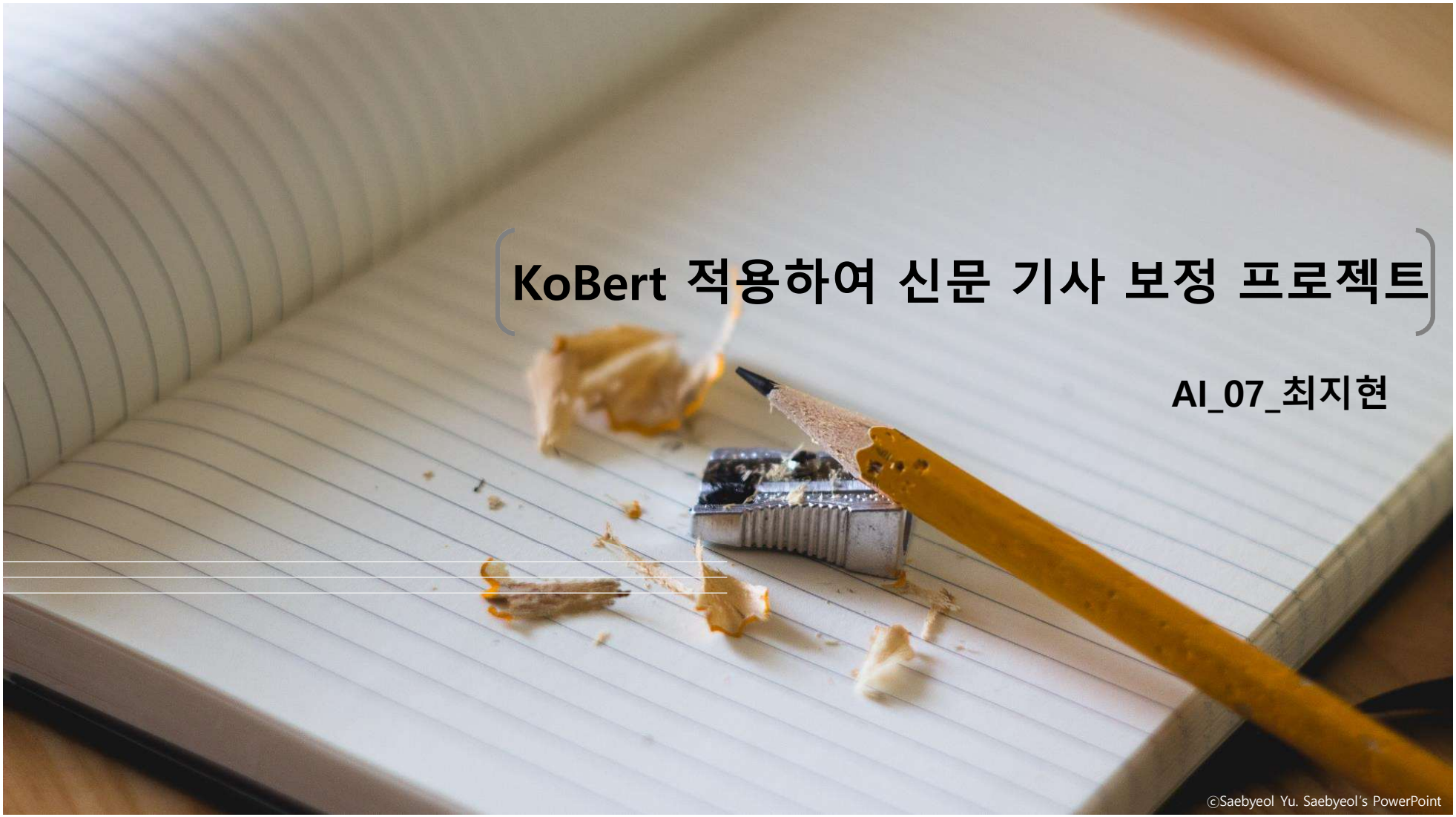
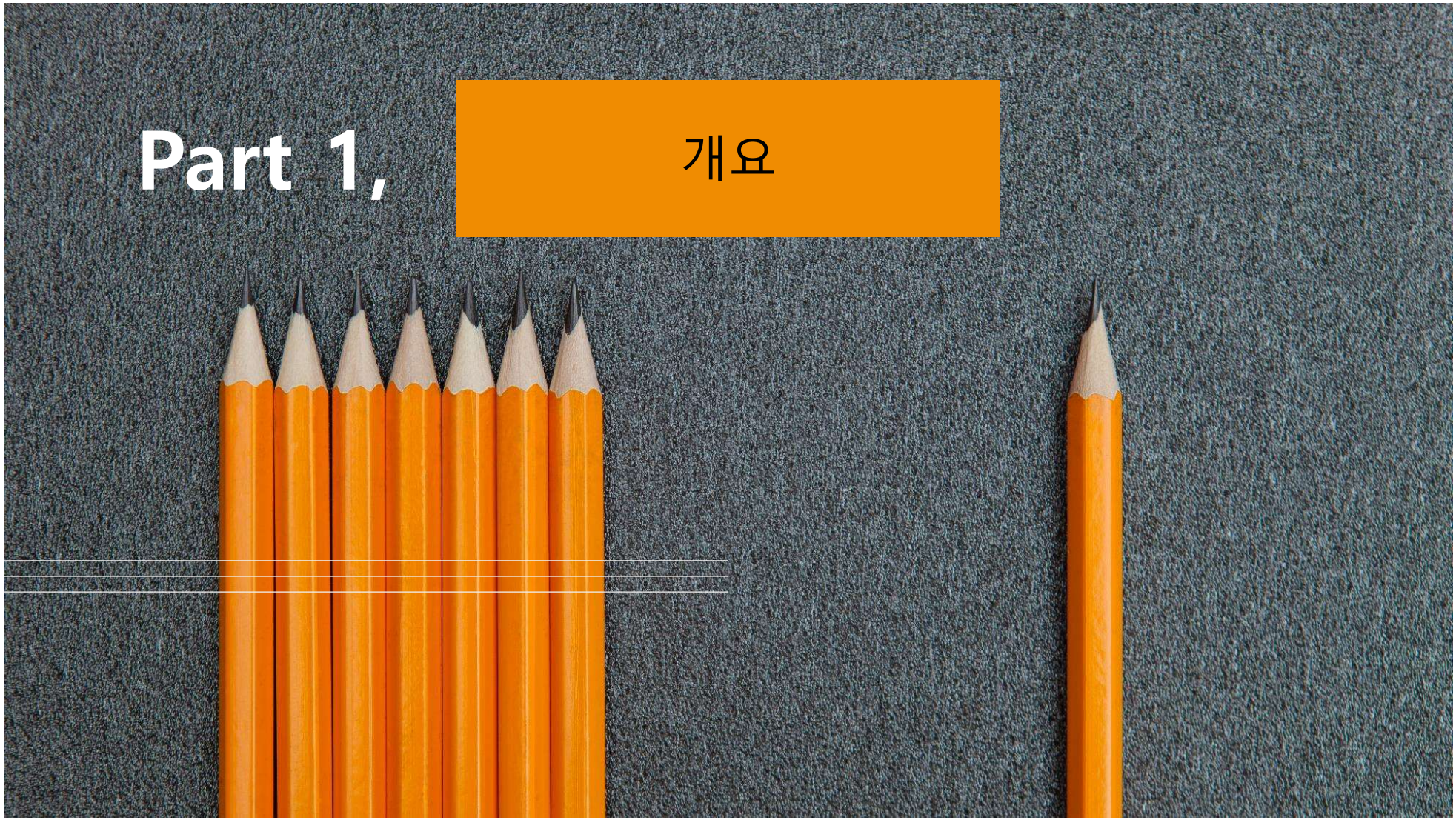

KoBert 적용하여 신문 기사 보정 프로젝트
AI_07_최지현
©Saebyeol Yu. Saebyeol's PowerPoint

목차


Part 1,
개요

개요

- 중소 신문사에서는 인터넷 기사 혹은 웹상에서 지면 신문 보기가 구현되지 않은 경우가 많습니다.

  이를 보완하고 적용하고자 합니다.
- 이번 프로젝트에서는 **신문 PDF를 OCR처리한 결과물을 보정**하는 것이 목적입니다.
  - 신문 원문이 자연스러운지 확인
- 이 목적을 달성하기 위해 아래 세가지 과정이 필요합니다.
  - 카테고리 분류
  - 띄어쓰기 검사
  - 단락 연결 여부 검사

workflow

1주차 | 2주차 | 3주차

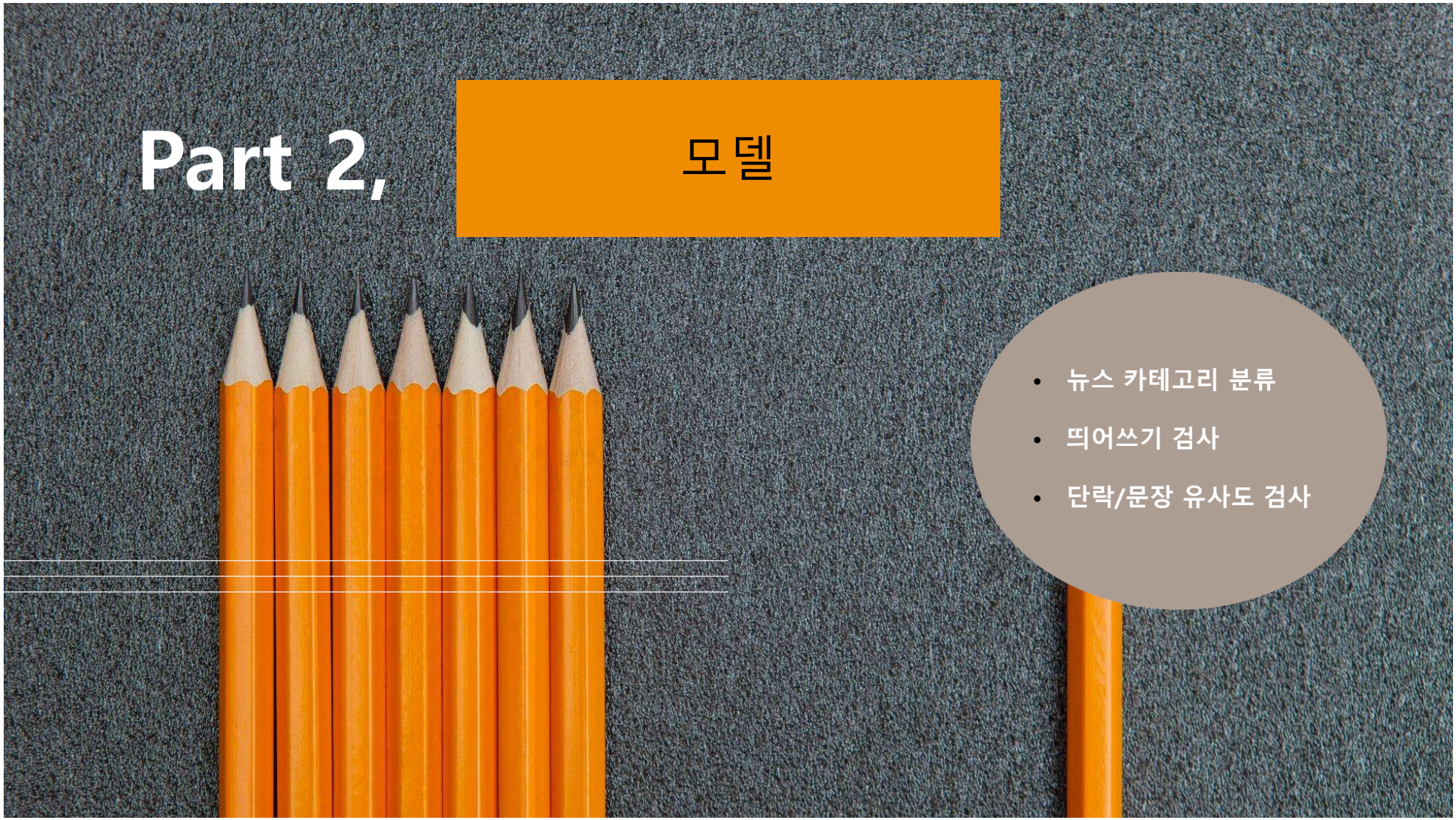
- **뉴스 카테고리 분류**
- SKT의 KoBERT 모델 fine-tuning
  (voca : 8002)
- BertClassifier 클래스 상속

- **한글 띄어쓰기 검사**
- SKT Kobert 경량화 버전,
  DistilKoBERT('monologg/kobert')
  모델 fine-tuning
- BertModel 클래스 상속

- **단락 연결 여부 검사**
- 단락/문장 유사도 파악
  - Kr-bert 모델 fine-tuning
    - snunlp/KR-Medium (voca : 16,424)
    - klue/bert-base'
  - BertNextSentencePredict
- 단락/문장 연결성 파악
  - Mask에 들어갈 단어 예측

Part 2,

모델



- 뉴스 카테고리 분류
- 띄어쓰기 검사
- 단락/문장 유사도 검사

카테고리 분류

- 데이터셋 : 한 문장(개행문장으로 나눔)의 길이가 7 이하, 특수 문자로 시작하거나 "쪽, -, >,]"으로 끝난다면 제거한 뒤 정해진 크기 (256자)까지 문장을 이어 붙여 한 데이터를 만든다.

- ㅇ, 『, (, ┌, |, ㄴ, ㄴ, ┌, ㅏ, ◎, [, ■, ㄱ, -, ., <, vs, 만, 해!로 시작하면 제거 // 쪽 -, >,]로 끝난다면 제거

- 데이터셋 크기

label	content
ITscience	16362
culture	15139
economy	16302
entertainment	15334
health	16004
life	14764
politic	13766
social	14396
sport	14956

- 데이터셋 형태 (content, encoded_label)

content	label	encoded_label
된장국도 끓이고 배추 겉저리도 담궜다. 두루 둘러앉아 식사를 하는데 큰 형님 게오르...	life	5
그리고 마무리는 게오르기 형님이 낮에 준비한 커다란 수박을 잘라먹었다.수박을 커다란...	life	5
그리고는 내게 이렇게 오셔서 덕분에 우리가 고려 말을 많이 합니다.그리고 앞으로도 ...	life	5

카테고리 분류

Dataset 개수

- Train data : 87,694
- Validation data : 21,924
- Test data : 27,405

Result

Validation accuracy : 0.8631

- **max_len = 256**
- **batch_size = 32**

• Best 모델 :
max_len이 512
하지만, 대부분의 기사 길이가
200~300이었기 때문에
max_len을 256으로 정했다.

Hyperparameter 조정

- Epoch 5일 때 비교
- Max_len : 64, batch_size : 64
  - Validation acc : 0.8131
- Max_len : 128, batch_size : 32
  - Validation acc : 0.8468
  - 20 epoch, 과적합 현상이 나타났다.

    validation acc : 0.7701
- Max_len : 256, batch_size : 32
  - Validation acc : 0.8631
- Max_len : 512, batch_size : 8
  - Validation acc : 0.8753

띄어쓰기 검사

- 띄어쓰기 검사이기 때문에 여기서 128자는 tokenizer.encode결과가 아닌 문장 길이이다.
- 들어가기 전에 한글에 띄어쓰기, 붙여쓰기 둘 다 허용하는 단어가 많기 때문에 원문의 띄어쓰기와 예측 값이 다를 수 있다.
- 데이터셋 : 한 문장(개행문장으로 나눔)의 길이가 7 이하, 특수 문자로 시작하거나 "쪽, -, >,]"으로 끝난다면 제거한 뒤 정해진 크기 (128자)까지 문장을 이어 붙여 한 데이터를 만든다.
- ㅇ, 『, (, ┌, |, └, ㄴ, ┌, ├, ◎, [, ■, ㄱ, -, ., <, vs, 만, 해!로 시작하면 제거 // 쪽 -, >,]로 끝난다면 제거

- 데이터셋 크기
- Train : 183,360
- Val : 45,841
- Test : 57,301

- 모델링 결과 : pred_slot_label, B앞에 띄어쓰기가 들어가야한다.

input_data	pred_slot_label	pred_content
차기대통령 첫 덕목은 '소통과 통합'	[B, I, B, I, I, B, B, I, I, B, I, I, I, B, I, I]	차기 대통령 첫 덕목은 '소통과 통합'
'청렴·도덕성' '강력 리더십' 順 반기문·문재인 2강 이재명 1중	[B, I, I, I, I, I, I, I, B, I, I, I, I, I, I, ...	'청렴·도덕성' '강력리더십' 順 반기문·문재인 2강 이재명 1중
조기 대선이 가시화된 가운데 차기대통령이 갖춰야 할 덕목으로 국민 3명 중 1명은 ...	[B, I, B, I, I, B, I, I, I, B, I, I, B, I, B, ...	조기 대선이 가시화된 가운데 차기 대통령이 갖춰야 할 덕목으로 국민 3명 중 1명은...
차기 대선후보 선호도에서는반기문 전 유엔사무총장(21.7%)과 문재인 전 더불어민...	[B, I, B, I, I, I, B, I, I, I, I, I, B, I, I, ...	차기 대선후보 선호도에서는 반기문 전 유엔사무총장(21.7%)과 문재인 전 더불어민...
1일 서울신문이 새해를 맞아 에이스리서치에 의뢰해 지난달 28~29일 19세이상 남...	[B, I, B, I, I, I, I, B, I, I, B, I, I, B, I, I, ...	1일 서울신문이 새해를 맞아 에이스리서치에 의뢰해 지난달 28~29일 19세 이상 ...

띄어쓰기 검사

Dataset 개수

- Train data : 183,360
- Validation data : 45,841
- Test1 data : 57,301
  (Corpus data)
- Test2 data : 2048
  (실제 기사 데이터, xml 추출)

- **'Monologg/kobert'**
- **13 epoch**
- **max_len = 128**
- **batch_size = 64**

Hyperparameter 조정

- Epoch 5일 때 비교
- 'monologg/kobert'
  - 10 epoch : **best model (저장 x)**
    - validation (loss : 0.0308, f1 : 0.9300)
  - 13 epoch
    - validation (loss : 0.0060, f1 : 0.9811)
    - test(기사 원문의 xml 파일) : (loss : 0.0354, f1 : 0.9293)

띄어쓰기 검사

실제 데이터 적용 : 세계 기업인과 학자 -> 세계 기업인 과학자 (오류 발생), 그래서 확인할 부분만 교체하여 넣었다.

1 뉴스 원문

```
['독일 뮌헨에서 북쪽으로 200㎞ 떨어짐',
 '진 암베르크는 인구 4만 4000명의 작음',
 '은 도시다. 하지만 앙겔라 메르켈 독일음',
 '총리는 물론 전 세계 기업인과 학자들음',
 '이 4차 산업혁명을 공부하기 위해 찾음',
 '는 곳이다. 세계적인 전기전자 기업 지음',
 '멘스가 자랑하는 스마트 공장이 이곳음',
 '에 있기 때문이다. 이 공장은 어떤 비밀음',
 '이 있기에 4차 산업혁명의 메카로 불음',
 '리는 걸까.음',
 '암베르크 공장은 커다란 병원 수술음',
 '실 같았다. 안내자가 "건초 더미에서음',
 '바늘을 찾는 게 더 쉬울 것"이라고 말음',
 '할 정도로 먼지 한점 없이 깨끗했다. 축음',
 '구장 1.5배인 1만㎡의 널찍한 공간에음',
 '서 컨베이어벨트는 쉴 새 없이 돌아갔음',
 '지만 '사람'은 거의 보이지 않았다. 의음',
 '사 수술복과 비슷한 파란색 유니폼을음',
 '입은 직원들이 드문드문 눈에 띄었는음',
 '데, 모니터만 들여다보고 있을 뿐 뭔가음',
```

2 확인할 부분 전처리

한 라인(마침표로 끝나는 문장)에서
띄어쓰기 확인할 단어는 아래와 같다.

```
([[('독일', ' 떨어', 20), ('진', ' 작', 22),
  [('SRT', '독일', 21), ('총리는', ' 학자들
  [('SRT', '지', 21), ('멘스가', ' 이곳', 1
  [('SRT', '비밀', 22),
   ('이', ' 불', 20),
   ('리는', ' 걸까.', 6),
   ('암베르크', ' 수술', 18),
   ('실', 'FIN', 20)],
  [('SRT', '더미에서', 20), ('바늘을', ' 말
  [('SRT', '축', 22), ('구장', ' 공간에', 2
  [('SRT', '의', 22),
   ('사', ' 유니폼을', 19),
   ('입은', ' 띄었는', 19),
   ('데,', ' 뭔가', 21),
   ('를', 'FIN', 20)],
  [('SRT', '암베', 20), ('르크', ' 인', 21),
  [('SRT', '프로', 22), ('그램', ' 장치' (PL
```

3 결과 확인

```
0 line :
    found : [ 떨어진] 독일 뮌헨에서 북쪽으로 200㎞ 떨어진 암베르크는 인구 4만4000명의 작은 도시다.
    found : [ 작은] 독일 뮌헨에서 북쪽으로 200㎞ 떨어진 암베르크는 인구 4만4000명의 작은 도시다.
1 line :
    not found : [독일총리는] 하지만 앙겔라 메르켈 독일 총리는 물론 전 세계 기업인 과 학자들이 4차 신
    found : [ 학자들이] 하지만 앙겔라 메르켈 독일 총리는 물론 전 세계 기업인 과 학자들이 4차 산업혁
    found : [ 찾는] 하지만 앙겔라 메르켈 독일 총리는 물론 전 세계 기업인 과 학자들이 4차 산업혁명을
2 line :
    found : [지멘스가] 세계적인 전기전자 기업 지멘스가 자랑하는 스마트 공장이 이곳에 있기 때문이다.
    found : [ 이곳에] 세계적인 전기전자 기업 지멘스가 자랑하는 스마트 공장이 이곳에 있기 때문이다.
3 line :
    found : [비밀이] 이 공장은 어떤 비밀이 있기에 4차 산업혁명의 메카로 불리는 걸까. 암베르크 공장은
    found : [ 불리는] 이 공장은 어떤 비밀이 있기에 4차 산업혁명의 메카로 불리는 걸까. 암베르크 공장
    not found : [ 걸까.암베르크] 이 공장은 어떤 비밀이 있기에 4차 산업혁명의 메카로 불리는 걸까. 암
    found : [ 수술실] 이 공장은 어떤 비밀이 있기에 4차 산업혁명의 메카로 불리는 걸까. 암베르크 공장
```

found : 붙여쓰기

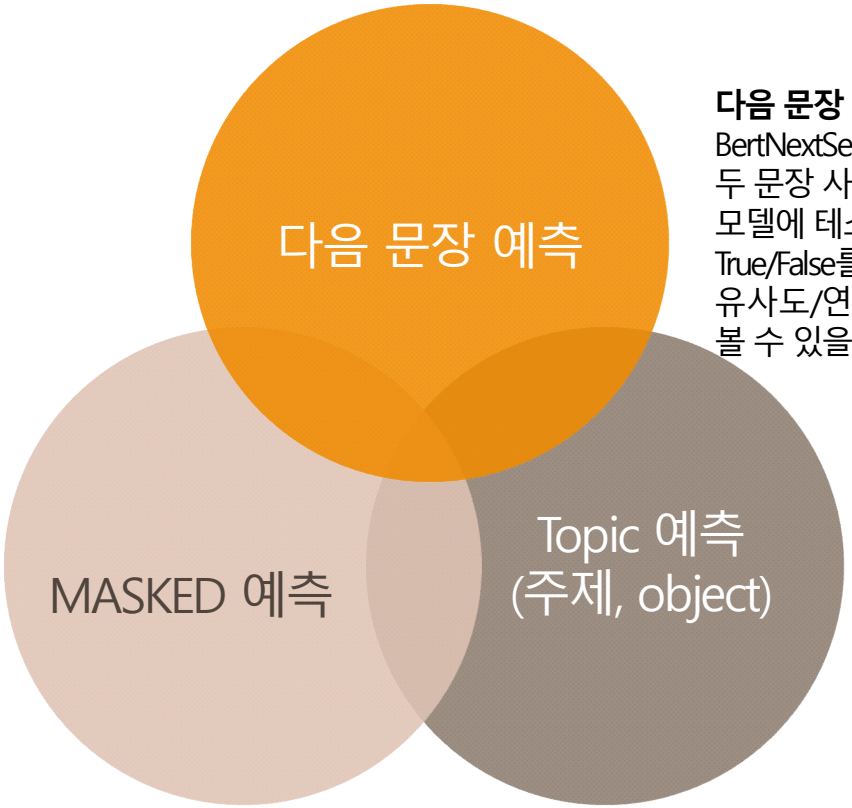
Not Found : 띄어쓰기

4 예측 결과

join을 통해서 원문에서 필요한 단어만 바꿔서 문장 이어붙이기

'독일 뮌헨에서 북쪽으로 200㎞ 떨어진 암베르크는 인구 4만 4000명의 작은 도시다. 하지만 앙겔라 메르켈 독일 총리는 물론 전 세계 기업인과 학자들이 4차 산업혁명을 공부하기 위해 찾는 곳이다. 세계적인 전기전자 기업 지멘스가 자랑하는 스마트 공장이 이곳에 있기 때문이다. 이 공장은 어떤 비밀이 있기에 4차 산업혁명의 메카로 불리는 걸까. 암베르크 공장은 커다란 병원 수술실 같았다. 안내자가 "건초 더미에서 바늘을 찾는 게 더 쉬울 것"이라고 말할 정도로 먼지 한점 없이 깨끗했다. 축구장 1.5배인 1만㎡의 널찍한 공간에서 컨베이어벨트는 쉴 새 없이 돌아갔지만 '사람'은 거의 보이지 않았다. 의사 수술복과 비슷한 파란색 유니폼을 입은 직원들이 드문드문 눈에 띄었는데, 모니터만 들여다보고 있을 뿐 뭔가를 만들거나 조립하지는 않았다. 암베르크 공장에선 전체 공정의 75%가 인간의 손이 필요 없는 자동화로 진행된다. 이 공장이 만드는 건 '시마틱 프로그램 가능 논리 제어 장치'(PLCs)로 불리는 일종의 칩이다. 기계나 로봇을 조종하고 움직이는 '두뇌'라고 생각하면 된다. 암베르크 · 뮌헨(독일) 임주형기자 hermes@seoul.co.kr ▶관련기사 8면'

문맥 유사도/연결성 파악



다음 문장 예측
BertNextSentencePredict
두 문장 사이에 유사도 측정
모델에 테스트할 앞뒤 단락을 넣었을 때
True/False를 잘 구별해낸다면 두 단락의
유사도/연결성이 있는지, 한 문장으로
볼 수 있을지 확인이 가능하다고 가정

MASKED 예측

Topic 예측

단락/문장 유사도/연결성 파악

다음문장 예측 (유사도 파악) → MASKED 예측 (연결성 파악) → TOPIC 예측 (X)

- **다음 문장 예측**
- BertNextSentencePredict
- 두 단락/문장 사이에 유사도 파악
- 구현 완성

- **MASKED 예측**
- BertForMaskedLM
- 두 단락 사이에 연결성 파악
- 미완성 (진행중)

- **Topic 예측 - 구현 x**
- 다음문장 예측과 중복되기 때문에 민감성이 더 높다고 판단되는 다음문장 예측 모델을 적용하였다.

다음 문장 예측 (데이터셋 만들기)

1 하나의 기사를 특정 길이 (128자)로 잘라서 순서대로 나열

```
[['특정한 풍경을 통하여 개인적인 감정을 시각화한 사진전박진호는 1990년대에 자신의 벗은 몸을 복사기로 복제하여 실험적인 결과물을 생산한 작가로서 유명하다.그런데 이번에는 새벽녘의 달과 하늘을 카메라 앵글에 담아서 감상적인 이미지를',
  0],
 [' 전시하였다.작가가 이번에 인사동에 있는 나무 갤러리에서 전시한 작품들은 달빛과 구름 낀 새벽하늘을 감성적인 느낌이 드는 결과물로 재구성하여 보여주고 있다.전시 작품마다 외형적으로 컬러가 자극적이고 전체적으로 톤도 어두워서 보는 ',
  1],
 ['이들을 감성적으로 동화시키는 표현전략을 구사하고 있다는 것을 알 수 있다.작가가 관심을 갖고 카메라 앵글에 담은 대상을 구체적으로 살펴보면 구름과 산봉우리 그리고 강변 풍경이다.그런데 왠지 표현대상과 소재가 낯설게 느껴지지 않고 ',
  2],
 ['너무나도 익숙하다.왜 그런 것 일까?그것은 작가가 표현대상으로 선택한 소재와 표현방식이  ‘일요 사진가’ 라고 일컬어지는 아마추어 작가들이 흔히 찍는 탐미적인 사진과 별다른 차이점이 없기 때문이다.사진은 시각예술이다.그러므로 작가가 ',
  3],
 ['표현하고자하는 주제와 관계없이 외형적으로 보여 지고 느껴지는 것이 무엇보다도 중요하다.그런데도 아쉽게도 이번에 박진호가 발표한 풍경사진들은 외형적으로 아마추어 작가들의 유미주의적이고 감상적인 작품과의 차별화에 실패하였다.풍경사진이',
  4],
 ['나 정물사진은 절제된 프레이밍과 세련된 컬러와 톤이 작품의 완성도를 보장하는데 있어서 중요한 요소로 작용한다.그러므로 주제선택과 더불어서 외형적으로는 그것을 좀 더 염두에 두고 작업을 진행을 했다면 최종 결과물의 완성도가 달라졌을',
  5],
```

2 데이터 : (전문장, 이어지는 다음 문장, label)

label이 0이면 이어짐, label이 1이면 안이어짐

```
('무슨 책을 읽어야 할지 모른겠다면 이 책을 보라 가끔 사람들이 내게 묻는 말이 있다. 살아가면서 즐거운 일이 있냐고? 그때 난 말한다. 즐거운 일이 뭐가 있겠는가. 고달프고 부대끼면서 사는 거지. 그러면서 한 가지 덧붙이는 게 있',
 ' 휴업기간은 4월 3일부터 5월 31일까지다. 폐업 방침을 발표했던 지난 2월 26일 당시 진주의료원 입원 환자는 200명이 넘었는데, 현재 3명만 남아 있다. 홍 지사가 이번 주 안에 진주의료원 폐업 발표를 할 것으로 보인다. ',
 1)
```

다음 문장 예측

- 데이터셋 형태 : 균형 데이터 (이어지는 문장을 수만큼 랜덤으로 이어지지 않는 문장을 추가하였다.)
- 토큰화한 결과를 바탕으로 전문장, 뒷문장 각각 모두 128의 길이를 가지고 있다. (넘으면 연결 부위에서 먼 곳을 자른다.)

- 데이터셋 : 총 12,600개의 기사에서 455,188개의 data 추출
- Train data : 291,320개, dataset의 64%
- Validation data : 72,830개, dataset의 20%
- Test data : 91,038개, dataset의 20%
- Test2 data : OCR 결과 파일 내 기사 텍스트를 이용해 모델 평가
- Test3 data : 최근 기사 80개

- 모델링 결과 : 각 문장(batch_size) 연결여부[True, False]에 대한 확률값.

  pred_label : [[True일 확률, False일 확률] * batch_size]
- Pred_label에서 두 값중 max값을 갖는 index를 구하면 True인지 False인지 알 수 있다. (0 : T, 1 : F)

다음 문장 예측

Dataset 개수

- Train data : 291,320개,
- Validation data : 72,830개
- Test1 data : 91,038개
  (Corpus data)
- Test2 data : 2740개 (T,F 포함)
  (실제 기사 데이터, xml 추출)
- Test3 data : 205개
  (최근 기사 데이터 80개 대상)

- **'snunlp/KR-Medium'**
- **max_len = 256 (두 문장의 합)**
- **batch_size = 32**

Hyperparameter 조정

- 'klue/bert-base' : 3 epoch
  o validation (loss : 0.3855, accuracy : 0.9246)
  o test (loss : 0.3870, accuracy : 0.9230)

- 'snunlp/KR-Medium'
  2 epoch : best model
  o validation (accuracy : 0.981)
  o test (loss : 0.3316, accuracy : 0.9807)
  3 epoch :
  o validation (loss : 0.3333, accuracy : 0.9793)
  o Test (loss : 0.3329, accuracy : 0.9796)
  o Test2 (기사 원문의 xml 파일) : (loss : 0.3654, accuracy : 0.9459)
  o Test3 (최근 기사 데이터)
  max_len : 256, accuracy : 0.9024
  max_len : 512, accuracy : 0.8780

다음 문장 예측 (틀리게 분류한 경우)

001 >> 중복되는 단어가 이전 문장과 다음 문장에 둘 다 나타남(주제 동일)

False인데 True로 분류한 경우
원인 : "정부/정권, 산업자원부/산업통상자원부", "공약/정책, 범죄수사,수사권,처벌/수사"처럼 비슷한 단어가 겹쳤기 때문에 이어지는 문맥이라고 판단한 것으로 여겨진다.
같은 페이지에 있는 기사는 주제가 동일한 경우가 많다. 특히 위의 경우에는 경기도 선거에 출마할 후보의 공약 내용이었기에 겹치는 부분이 더욱 많았다.

002 >> 중복되는 단어의 부족, '그런데'로 시작하면서 반전되는 내용이 나타남

True인데 False로 분류한 경우
원인 : 전체 문장이 아니라 이어지는 문장 부근으로 잘라서 보니 겹치는 단어가 부족하다. 다음 문장이 "그런데"로 시작하는 경우 내용이 반전되고 새로운 이야기가 시작되었다.

개선점

1 카테고리 분류 모델

- ✓ 성능 향상
- ✓ 길이를 넉넉하게 잘라서, 토큰화할 때 길이 맞추기 (padding 줄이기)
- ✓ 실제 데이터에 대한 성능 평가 필요

2 Masked 모델 완성

- ✓ Masked 모델 에러 해결
- ✓ 완성 시 단락 연결성(자연스러움) 확인을 위해 실제 데이터에 적용

3 띄어쓰기 검사 결과물

- ✓ 띄어쓰기 검사 결과물 깔끔하게 출력하기
- ✓ PyKoSpacing과 비교, 장점 및 단점 확인

3주 동안 고생하셨습니다!

감사합니다.