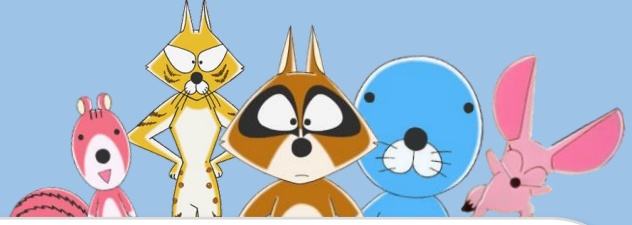
국내 관광 활성화를 위한 카테고리 분류

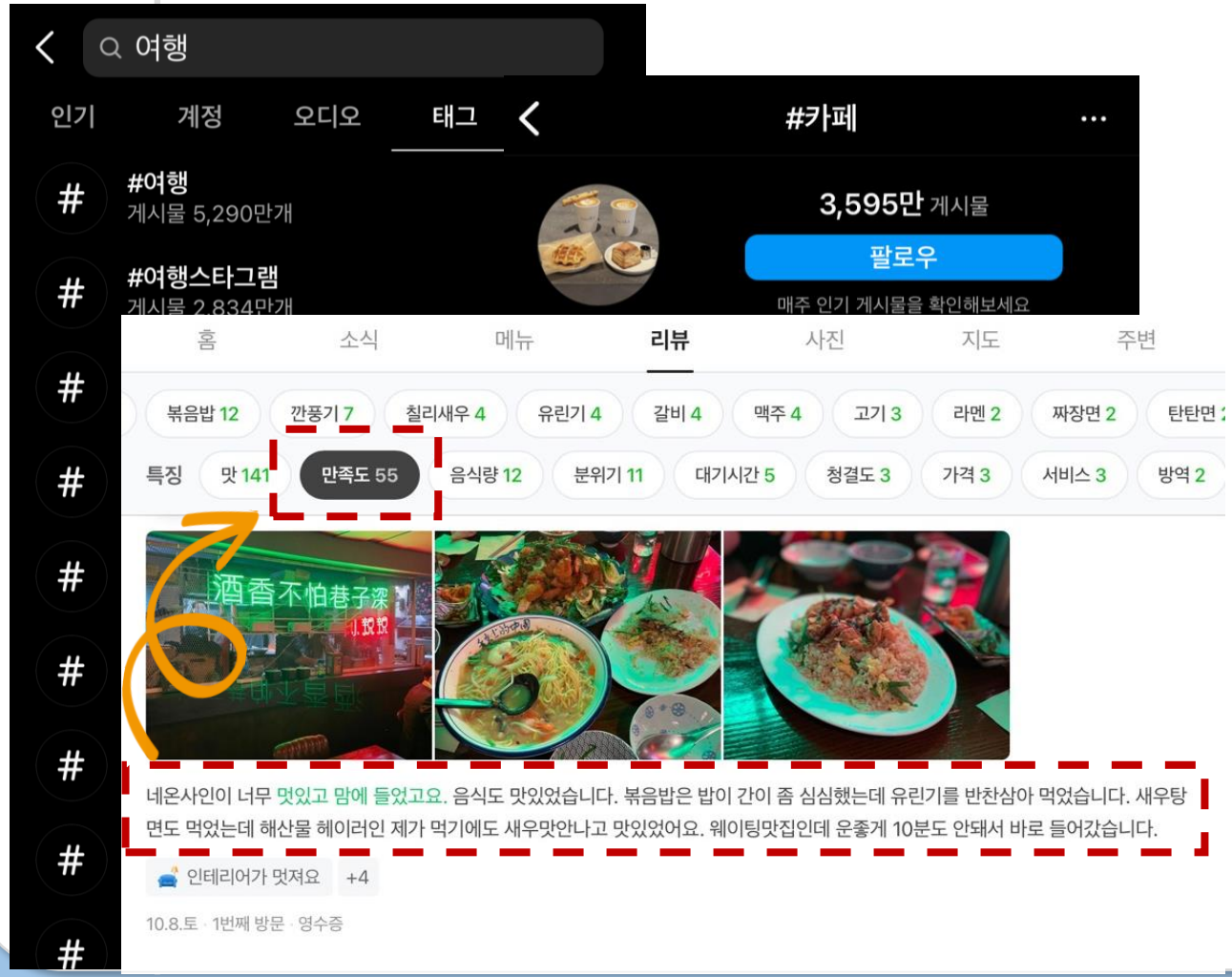
김예찬 / 박시언 / 박윤아 / 정승민 / 김민

주제선정 및 배경

주제선정 및 배경



1. 분석 배경

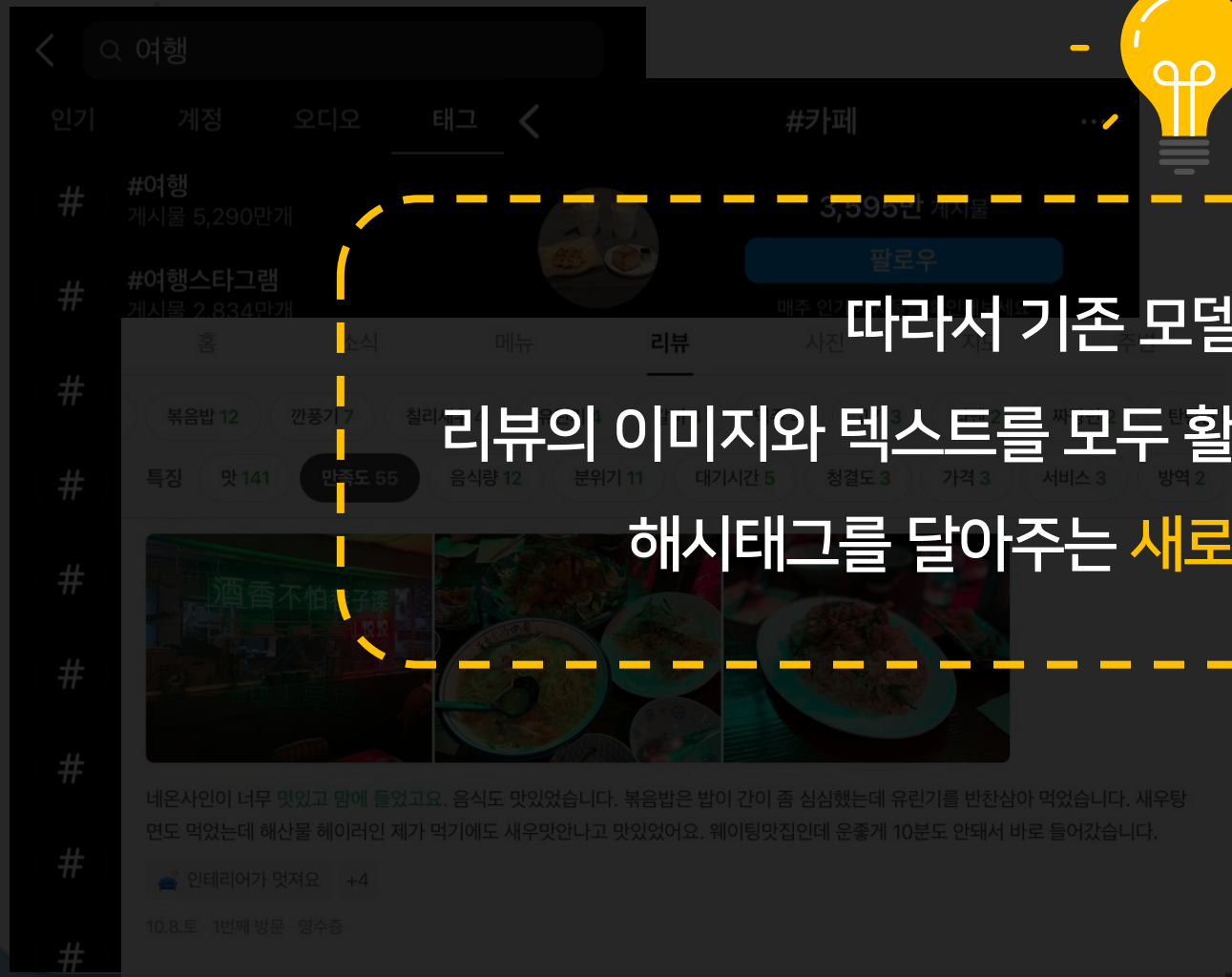


코로나 19의 감소와 함께 관광업계가 살아나면서
최근 # 해시태그를 통한 장소의 검색과 홍보가
활발하게 이루어지고 있음

그러나 기존모델은 리뷰 텍스트 중
키워드만 추출하여 해시태그 반영하기 때문에
정보가 한정되어 있음

주제선정 및 배경

1. 분석 배경



따라서 기존 모델을 개선하여

리뷰의 이미지와 텍스트를 모두 활용해서 **카테고리를 분류** 한 후

해시태그를 달아주는 **새로운 모델**을 만들어보자

최근 # 해시태그를 통한 장소의 검색과 홍보가

활발하게 이루어지고 있음

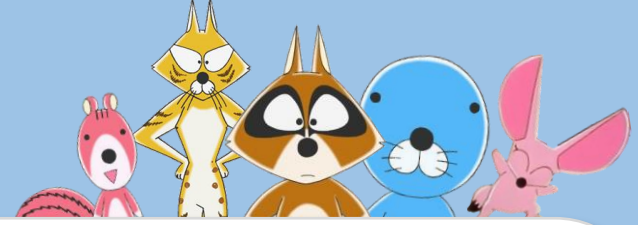
그러나 기존모델은 리뷰 텍스트 중

키워드만 추출하여 해시태그 반영하기 때문에

정보가 한정되어 있음

EDA 및 전처리

EDA 및 전처리



1. 데이터 소개

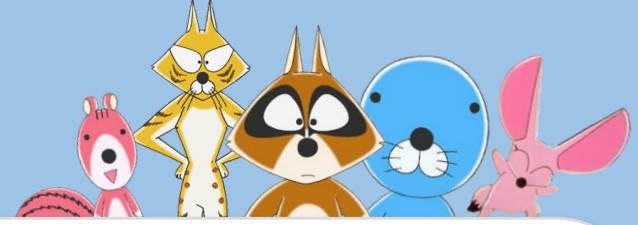
한국관광공사가 제공하는 국문관광 데이터 train. csv

16987 X 6

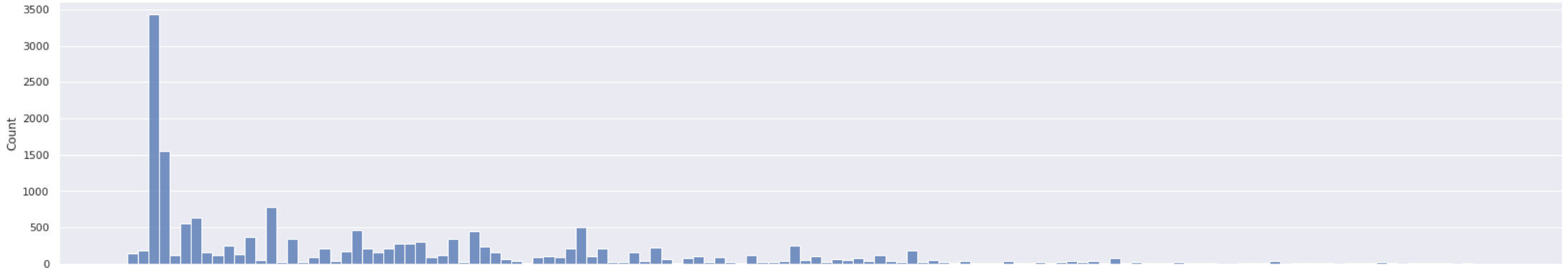
id	Img_path	Overview	Cat1	Cat2	cat3
TRAIN_00000	./image/train/TRAIN_00000.jpg	소안항은 조용한 섬으로 인근해안이 청정해역으로 일찍이 김 양식을 …	자연	자연관광지	항구/포구
TRAIN_00001	./image/train/TRAIN_0001.jpg	기도 이천시 모가면에 있는 골프장으로 대중제 18홀이다. 회원제로 개장을 …	레포츠	육상레포츠	골프
TRAIN_00003	./image/train/TRAIN_00002.jpg	금오산성숯불갈비는 한우고기만을 전문적으로 취급하고 사용하는 부식 자재 또한 유기농법으로 …	음식	음식점	한식
…	…		…	…	…

이미지 데이터와 텍스트 데이터를 이용해 자동으로 카테고리를 분류하는 것이 목표

EDA 및 전처리



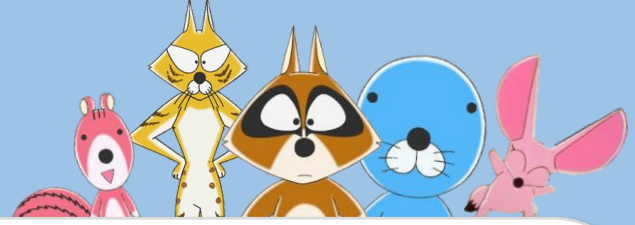
2. DATA 증강 전 분포



Category 3

총 분류는 128개, 매우 불균형한 분포를 보임

EDA 및 전처리



3. 라벨링 수정

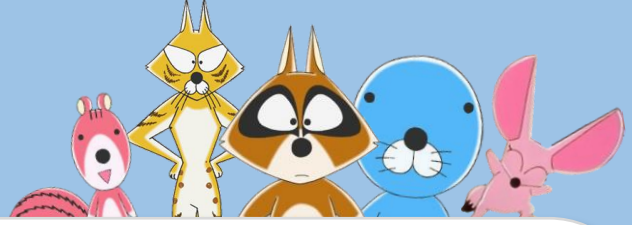
'한식'이 '서양식'으로 잘못 표기됨

id	Img_path	Overview	Cat1	Cat2	Cat3
TRAIN_00000	./image/train/TRAIN_00000.jpg	한국 미식 프로그램에 자주 소개된 매장이다 대표메뉴는 꽃게찜이다. 서울시 성동구에 있는 한식전문점이다.	음식	음식점	서양식
TRAIN_00001	./image/train/TRAIN_0001.jpg	기도 이천시 모가면에 있는 골프장으로 대중제 18홀이다. 회원제로 개장을 …	레포츠	육상레포츠	골프
TRAIN_00003	./image/train/TRAIN_00002.jpg	금오산성숯불갈비는 한우고기만을 전문적으로 취급하고 사용하는 부식 자재 또한 유기농법으로 …	음식	음식점	한식
…	…		…	…	…

잘못 라벨링된 경우가 존재하는 것을 발견하고

직접 모든 행을 확인하여 라벨을 수정함

EDA 및 전처리



3. 라벨링 수정

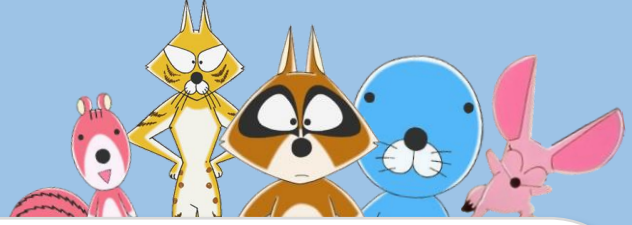
한국관광공사에서 제공된 데이터의 분류 기준은 전반적으로 너무 모호함

= 분류 모델이 학습에 어려움을 겪을 수 있음

따라서 구분이 모호한 카테고리에 대해

일관된 기준을 설정하여 라벨 수정을 진행함!

EDA 및 전처리

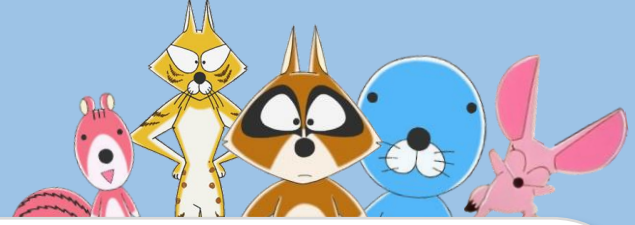


3. 라벨링 수정

Cat3 카테고리 재분류 기준

1. ‘컨벤션’과 ‘전시회’를 모두 ‘전시회’로 통일
2. '해수욕장’ 카테고리 중 ‘섬’에 해당하는 데이터 재분류
3. ‘상설시장’ 카테고리 중 '5일장’에 해당하는 데이터 재분류
4. '호텔’ 카테고리 추가
5. ‘한옥 스테이’ 및 '모텔’ 카테고리를 ‘호텔’로 재분류
6. ‘홈스테이’ 카테고리 전부 재분류 후 삭제

EDA 및 전처리



4. 데이터 증강

Easy Data Augmentation

현재 보유하고 있는 데이터를 변형하여 증강하는 방법

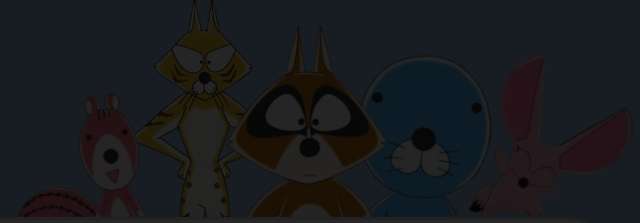
SR — Synonym Replacement,
특정 단어를 유의어로 교체

RI — Random Insertion,
임의의 단어를 삽입

RS — Random Swap,
문장 내 단어 위치를 임의로 바꿈

RD — Random Deletion,
문장 내 단어를 임의로 삭제

SR, RI 방법을 적용할 경우 문장의 의미가 변형될 가능성이 높기 때문에,
RS, RD 방법을 통해서만 증강을 하기로 결정



4. 데이터 증강

Easy Data Augmentation

현재 보유하고 있는 데이터를 변형하여 증강하는 방법

데이터 역전 현상

고유값 개수 기준으로 증강할 경우 특정 값보다
조금 큰 데이터는 증강되고, 특정 값보다 조금 작은
데이터는 증강되지 않는 문제 발생

SR Synonym Replacement,
특정 단어를 유의어로 교체

RI Random Insertion,
임의의 단어를 삽입

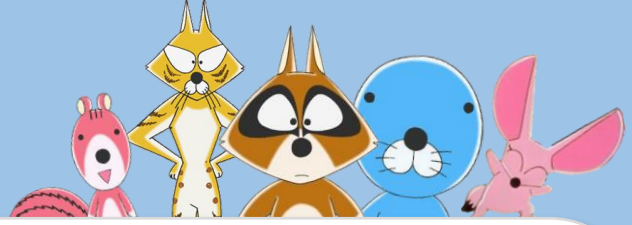
RS Random Swap,
문장 내 단어 위치를 임의로 바꿈

RD Random Deletion,
문장 내 단어를 임의로 삭제

확률적인 방법으로 데이터 증강

SR, RI 방법을 적용할 경우 문장의 의미가 변형될 가능성이 높기 때문에,
RS, RD 방법을 통해서만 증강을 하기로 결정

EDA 및 전처리



4. 데이터 증강 - 확률적인 접근

Step 1. $\frac{\text{전체 데이터의 개수}}{\text{해당 소분류(cat3)에 속하는 데이터 개수}}$ 구하기

Step 2. 분산을 줄이기 위해 루트를 씌어 줌

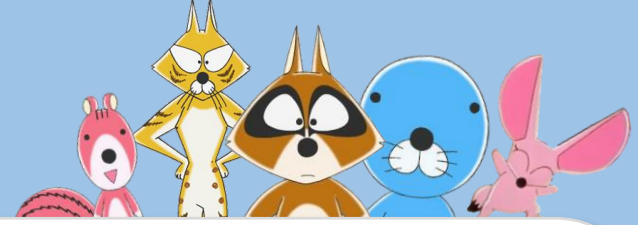
Step 3. 해당 값을 0~1 사이의 확률 값으로 만들기 위해 **min-max scailing 수행**

Step 4. 각 관찰 값에 대해 binomial distribution을 통해 0, 1 값 추출

Step 5. 0이면 데이터 증강을 하지 않고, 1이면 데이터를 증강

소분류(cat3)에 속하는 데이터 수가 적을 수록 증강될 확률이 높아지기 때문에 전체 데이터의 분포를 해치지 않고 증강할 수 있다는 장점이 존재함

EDA 및 전처리



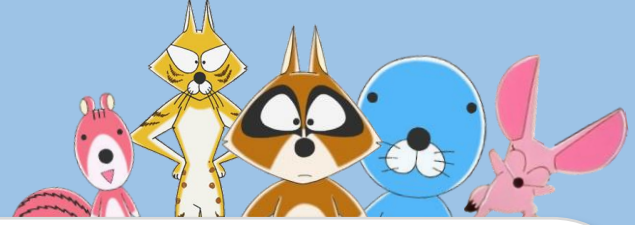
5. DATA 증강 후

Category 3

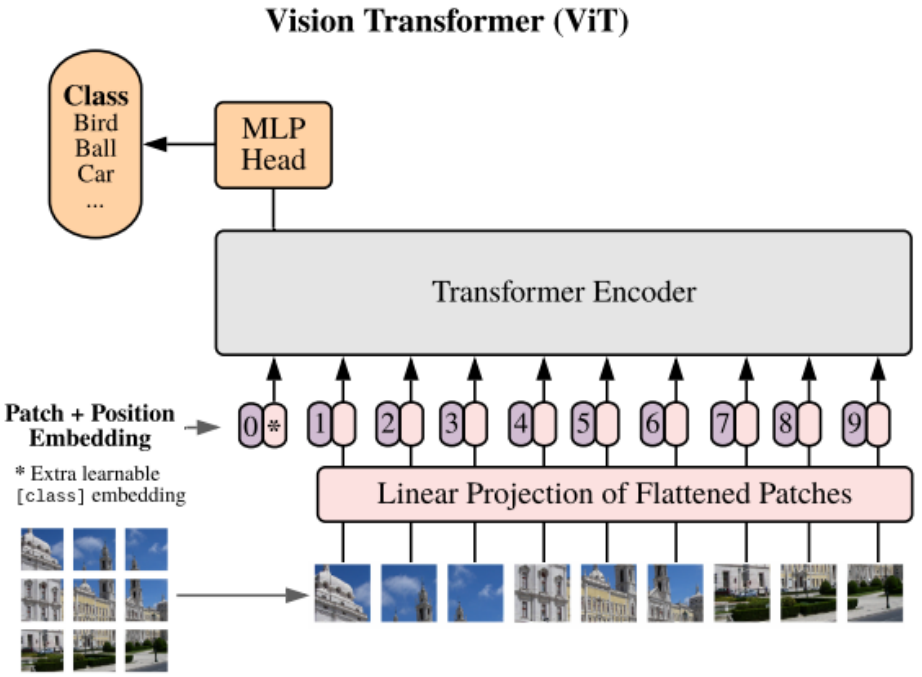
데이터 수가 적은 카테고리를 시각화한 결과, 상당히 증강된 것을 알 수 있음

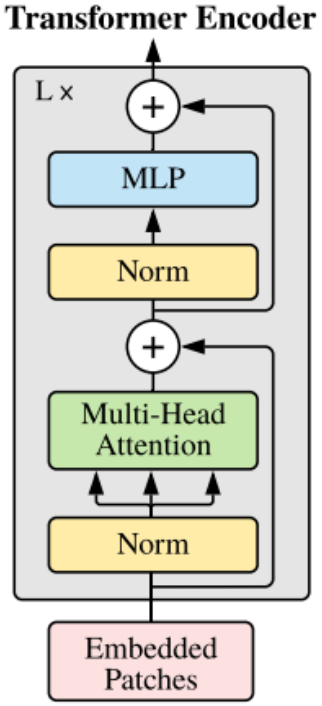
카테고리 분류 모델

카테고리 분류 모델



ViT





이미지를 패치로 분할

▼

Positional embedding vector와

CLS token을 concat

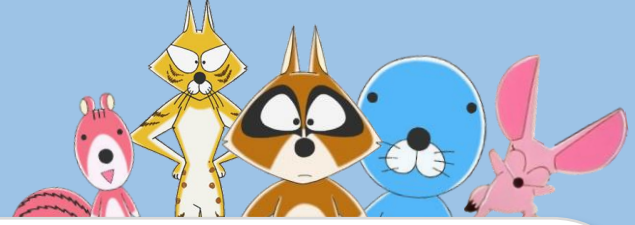
▼

위 1차원 embedding vector를 input

▼

인코더, FC layer를 거쳐 최종적으로 분류

카테고리 분류 모델



BERT

BERT

Bidirectional Encoder Representations from Transformers

NLP 분야에서 기록적인 성능을 보이는 transformer 기반 모델

레이블이 없는 방대한 양의 데이터를 사전 학습한 모델

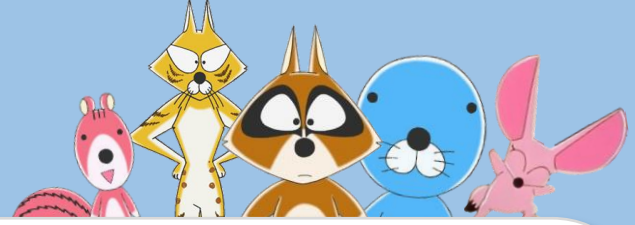
사전 학습한 모델을 레이블이 있는 데이터에 대해 파인 튜닝하여 사용

기존의 모델처럼 문장을 왼쪽에서 오른쪽으로만 읽어 문맥을 파악하지 않고,

양방향에서 전체 맥락을 파악하는 것이 특징

→ 의미 파악 성능 향상

카테고리 분류 모델



Multimodal

Concat

[Image Processing]

- ViT
- Custom CNN layer

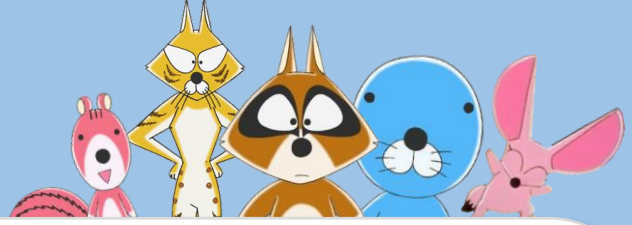
[NLP]

- RoBERTa
- KoBERT
- KcELECTRA

"편백골 관광농원 캠핑장은 글램핑과
야영장이 함께 운영되는 캠핑장이다.
캠핑장 옆으로 계곡물이…"

최종 모델

최종 모델



모델시도1

< 가정 1 >

텍스트가 긴 경우 가운데보다 초반과 후반부에 분류에 핵심적인 내용이 많을 것이다.

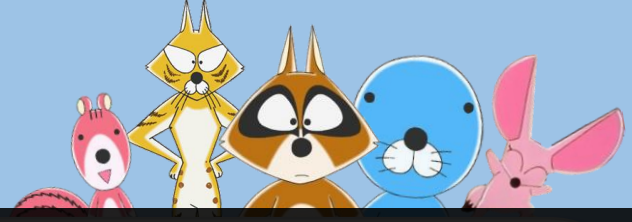
▼

문장의 개수가 5개 이상일 경우

앞 문장 3개와 뒷 문장 2개만 추출하여 학습시켜보자 !

" 부여안방마님은 충남 부여군의 유일한 한옥 체험 숙소다. 안채 상량을 기준으로 1896년 지어진 한옥을 복원했다. 안채와 별채, 사랑채, 행랑채 등으로 이뤄진 한옥이다.
…
가족 또는 동호회 모임, 작은 음악회 등의 장소로도 대관한다. 한옥 카페에서는 직접 만든 쌍화차, 대추차, 한방차 등을 판매한다. "

최종 모델



2. 모델시도1

" 부여안방마님은 충남 부여군의 유일한 한옥 [illegible]안채 상량을 기준으로 1896년 지어진 한옥을 복원했다. 안채와 별채, 사랑채, 행랑채 등으[illegible]다.
…
가족 또는 동호회 모임, 작은 음악회 등의 장소로도 대관한다. 한옥 카페에서는 직접 만든 쌍화차, 대추차, 한방차 등을 판매한다. "

텍스트 길이 조정 전 데이터로 학습시켰을 때보다 성능 하락

< 가정 >

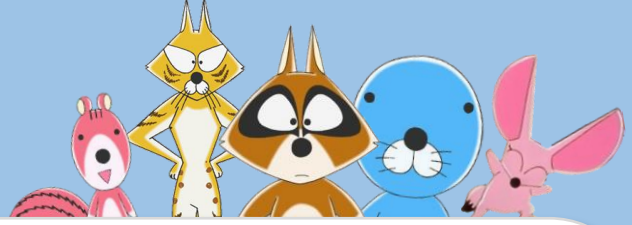
텍스트가 긴 경우 가운데보다 초반과 후반부에 분류에 핵심적인 내용이 많을 것이다.

텍스트 중반부의 내용도 모델의 분류 성능에
유의미한 역할을 했을 것…

문장의 개수가 5개 이상일 경우

앞 문장 3개와 뒷 문장 2개만 추출하여 학습시켜보자 !

최종 모델



모델시도2

< 가정 2 >

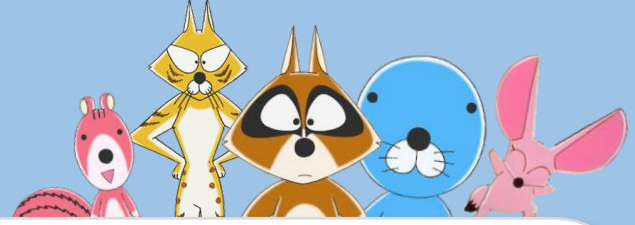
핵심 내용만을 추출하여 모델을 학습시키면 분류 성능이 좋아질 것이다.

▼

KeyBERT를 활용하여 키워드를 추출한 후

모델을 학습시켜 카테고리 예측을 진행해보자 !

최종 모델



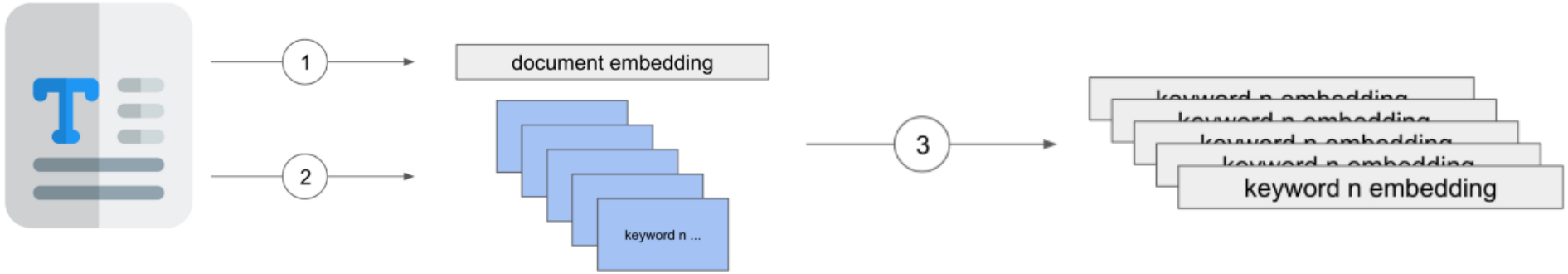
모델시도2

KeyBERT

BERT 기반 키워드 추출 모델로,

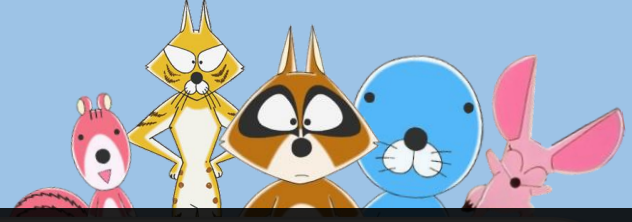
BERT를 통해 문서의 주제를 파악하고, bag-of-words 기법으로 n-gram 임베딩

이후 코사인 유사도를 계산하여 키워드 추출



총 10개의 키워드를 추출하고 이를 통해 학습을 진행!!

최종 모델



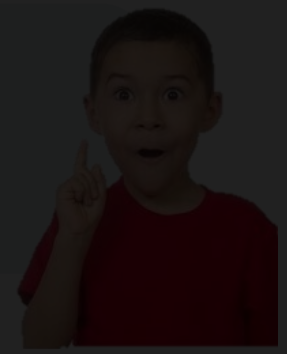
3. 모델시도2

< 가정 2 >

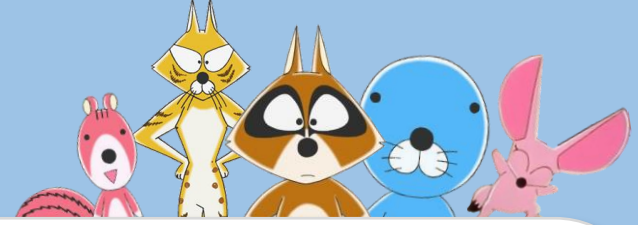
원본 텍스트 데이터로 학습시켰을 때보다 **성능 하락**

문장의 문맥적인 의미가 모델 학습에 반영되지

않은 점이 성능에 부정적으로 작용했을 것 …



최종 모델

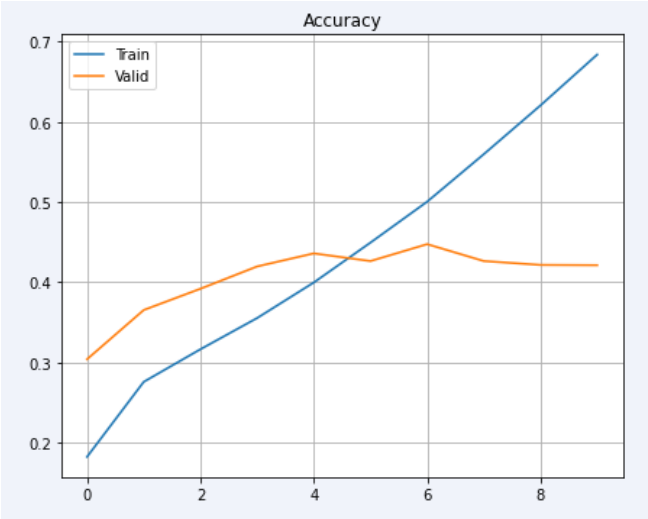


모델 성능 비교

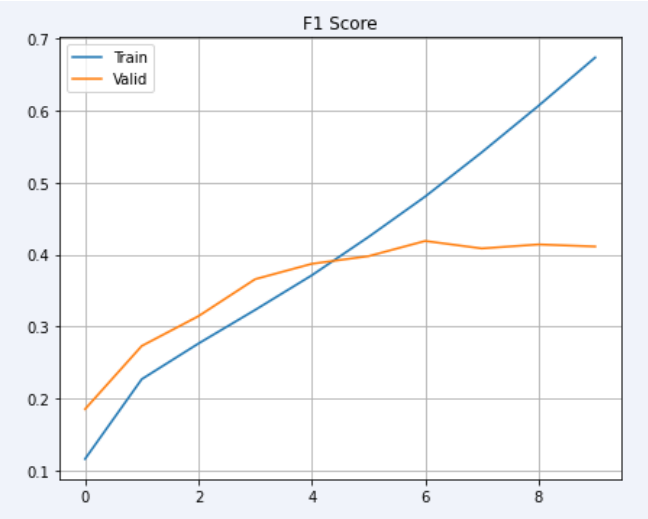
KoBERT

Accuracy = 0.3909

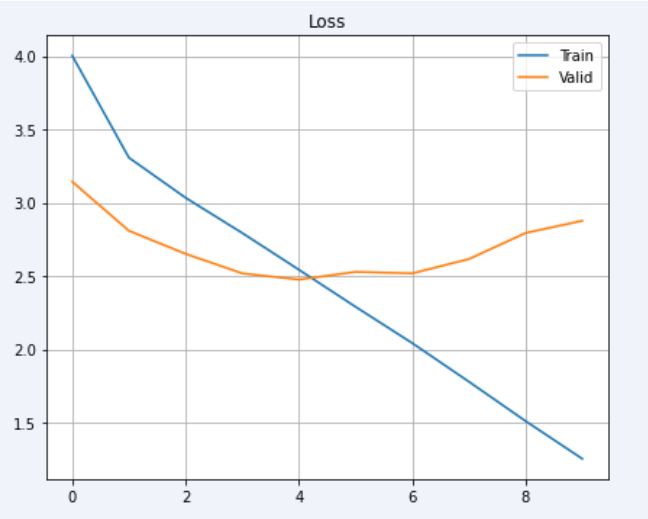
F1-score = 0.3858



Accuracy

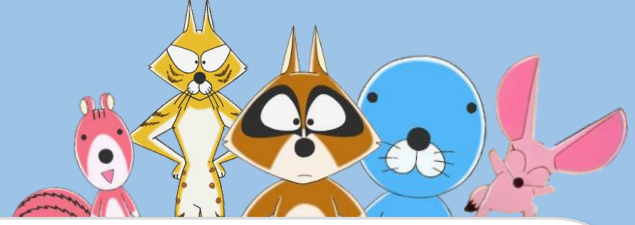


F1 - score



Loss

최종 모델

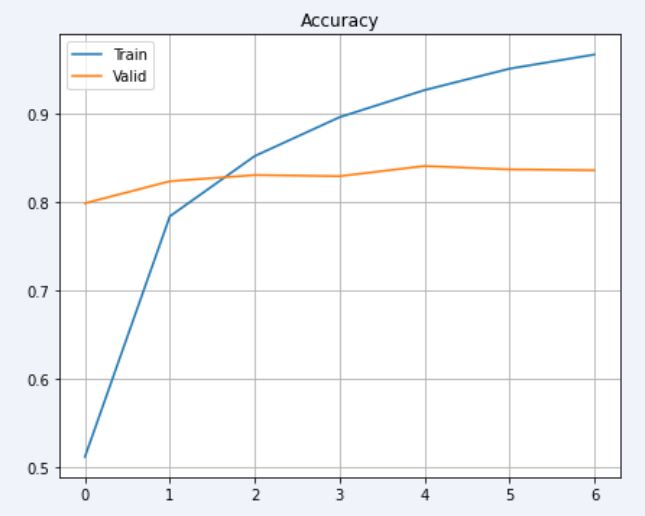


모델 성능 비교

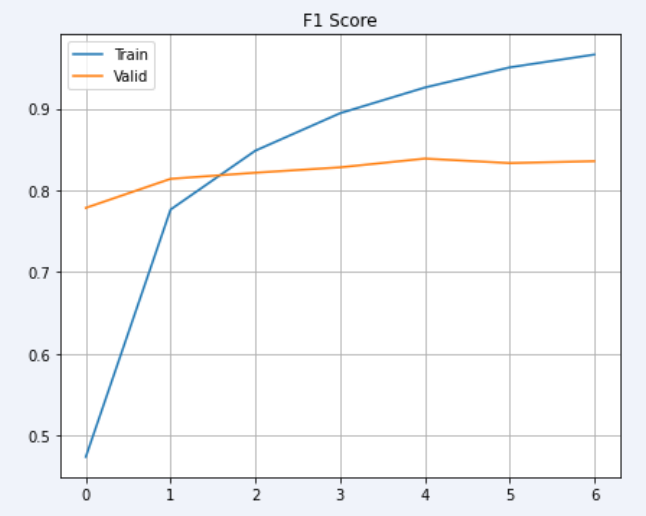
RoBERTa

Accuracy = 0.8733

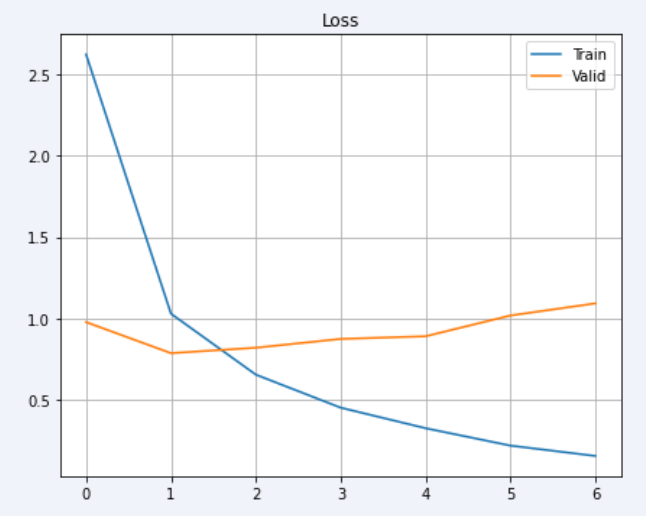
F1-score = 0.8740



Accuracy

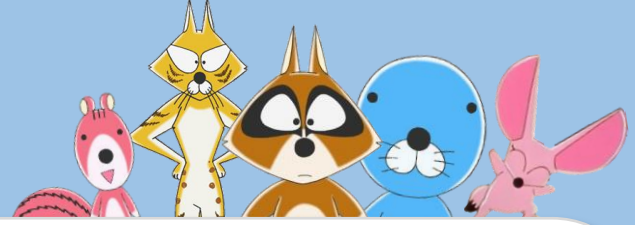


F1 - score



Loss

최종 모델

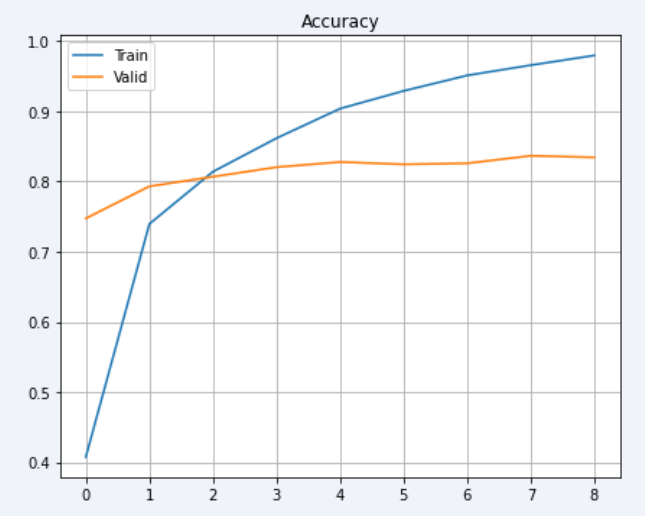


모델 성능 비교

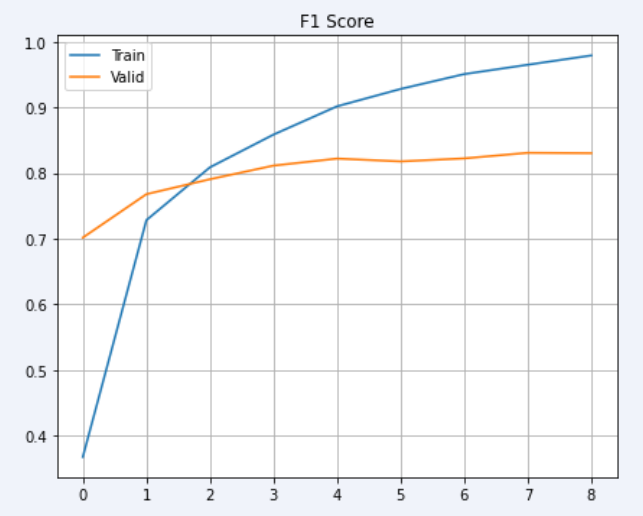
ViT + RoBERTa

Accuracy = 0.8744

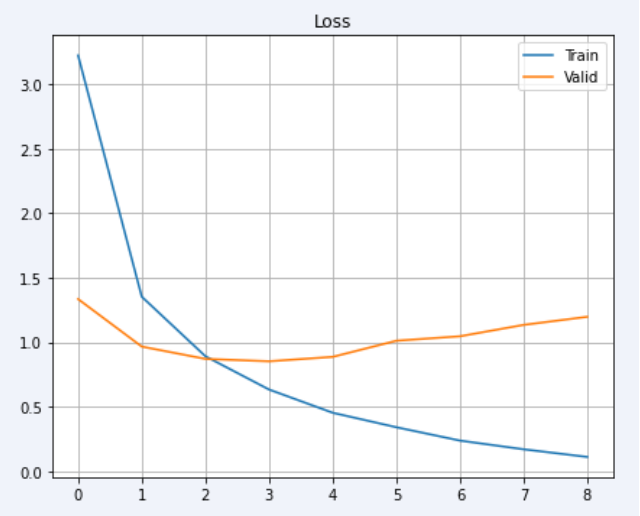
F1-score = 0.8728



Accuracy

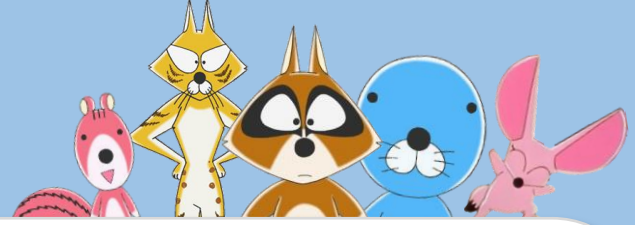


F1 - score



Loss

최종 모델



모델 성능 비교

Multimodal

	true
한식	1471
야영장,오토캠핑장	650
바/까페	335
유적지/사적지	236
일반축제	222
...	...
터널	2
헬스투어	1
이색체험	1
백화점	1
요트	1

112 rows × 1 columns

RoBERTa

	true
한식	1443
야영장,오토캠핑장	649
바/까페	332
일반축제	227
유적지/사적지	222
...	...
문화관광축제	1
헬스투어	1
영화관	1
백화점	1
요트	1

112 rows × 1 columns

데이터의 개수가 적은 클래스에 대해서 얼마나 분류가 잘 이뤄졌을까?

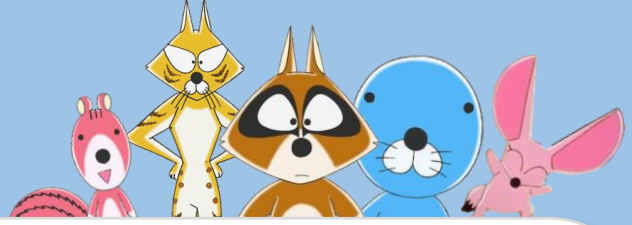
Multimodal VS RoBERTa

Test set에 존재하는 cat3: 총 123개

Multimodal과 RoBERTa 모두 112개의
Cat3에 대한 예측 결과가 존재

F1-score는 RoBERTa가 더 높았지만
예측한 소분류의 개수는 동일함

최종 모델



모델 성능 비교

Multimodal

이미지 데이터에 일관성이 생긴다면
유의미한 성능 향상을 기대할 수 있을수도..?

Multimodal을 택한 이유

✓ 성능 지표는 RoBERTa가
멀티 모달에 비해 약간 좋음

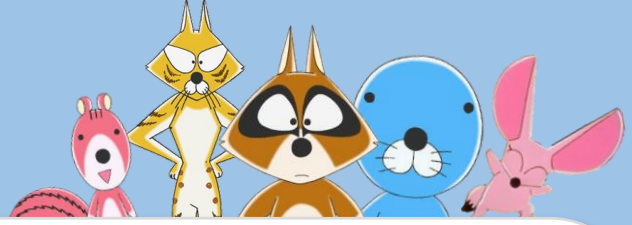
BUT,

✓ 이미지 데이터가 일관성이 없다는
점을 고려해야함

✓ 성능 차이가 그렇게 크지 않기에,
SNS에서 활용하기 좋은 **멀티모달**을
최종적으로 채택!!

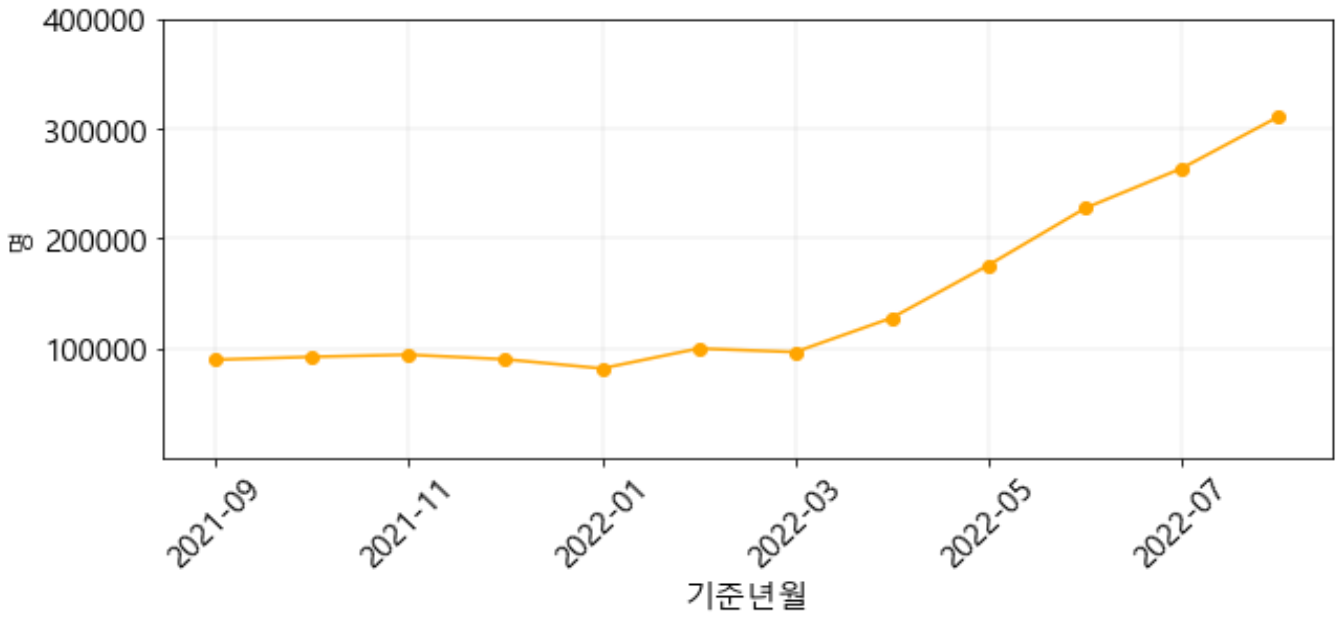
토이 프로젝트

토이 프로젝트



토이 프로젝트 선정 배경

방한 외래관광객 추이

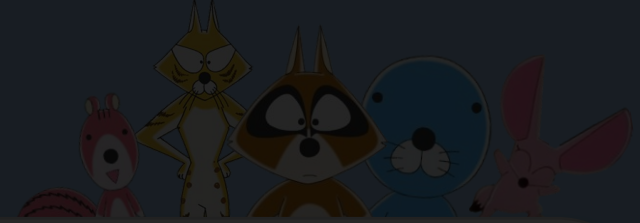


외래 관광객 수

276% 증가

작년 9월 대비 올해 9월 방한 외래 관광객의 수는 276% 증가한 337,638명이 입국했으며

방한 외래 관광객 수는 계속 증가할 것으로 예상됨



토이 프로젝트 선정 배경

현재 SNS상에서 외국인 관광객을 위한 **영문 관광자료**의 수요가 높음

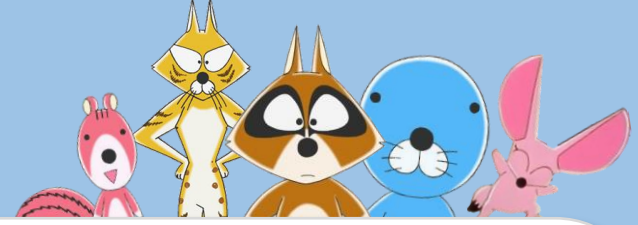
구분	2021년
전 생애 한국여행 경험	19.4%
향후 3년 내(~2024년) 한국여행 의향	47.0%
한국 방문 예상 시기	2024년(35.7%)
	2022년(32.0%)
	2023년(28.9%)
	2021년 7~12월(3.4%)

외국인 관광객을 위한
해시태그 생성 모델을 추가로 만들어 보자!

기계번역 → 카테고리 분류 → 해시태그 생성 모델

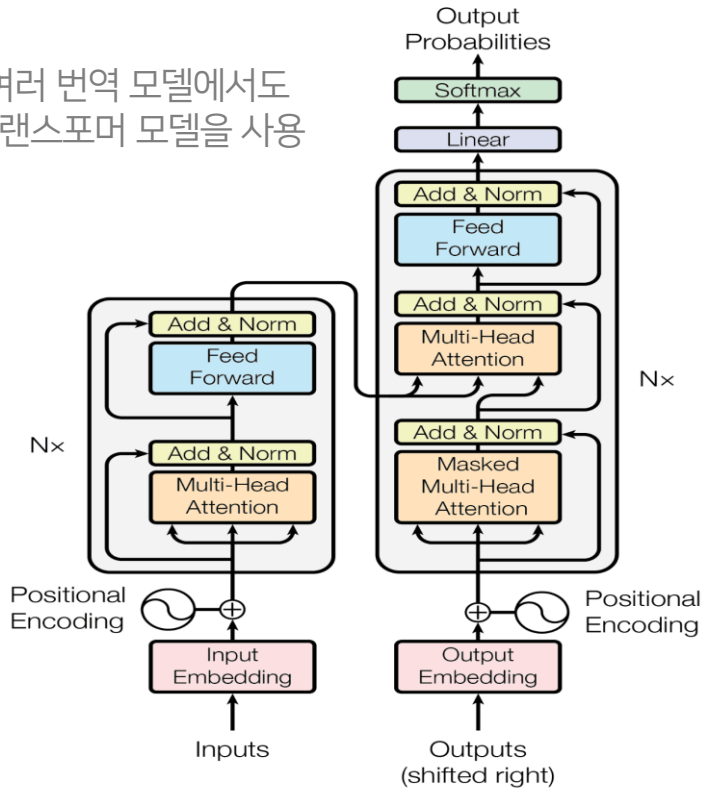
2021년 잠재 방한여행객 조사 결과 향후 3년 내에 한국여행 의향이 47.0%로 나타났으며,
앞으로 많은 외래 관광객이 한국을 방문할 것으로 예상됨

토이 프로젝트



기계번역

여러 번역 모델에서도
트랜스포머 모델을 사용

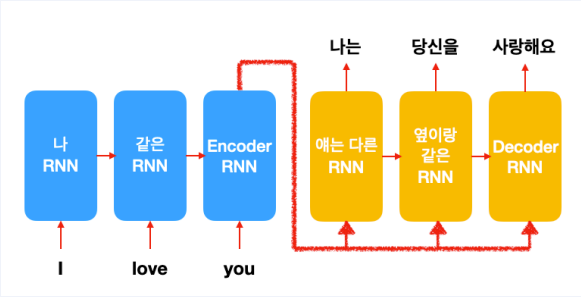


자세한 내용은 딥러닝 1주차 주제분석 참고

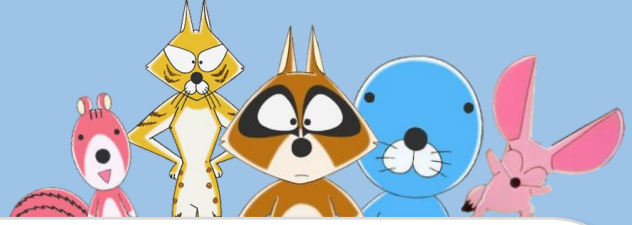
NMT

✓ 신경망기반 기계번역

✓ 두 언어의 말뭉치를 학습시켜 번역하는 모델을 구현



토이 프로젝트



기계번역

Step 1. 말뭉치 토크나이즈

Source: 도깨비에 대한 현대적 활용은 앞으로도 과제가 될 것이다.

Target: The modern use of goblin will be a challenge in the future.

▼

<sos>도깨비/에/대한/현대적/활용은/앞으로도/과제가/될/것/이다.<eos>

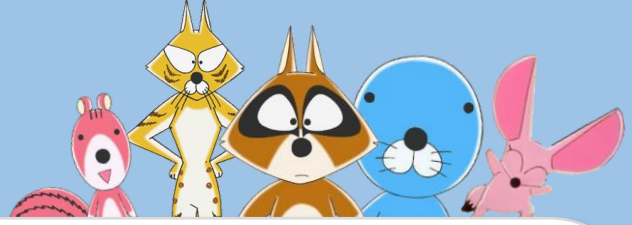
<sos>The/modern/use/of/goblin/will/be/a/challenge/in/the/future.<eos>

▼

Step 2. 임베딩 벡터

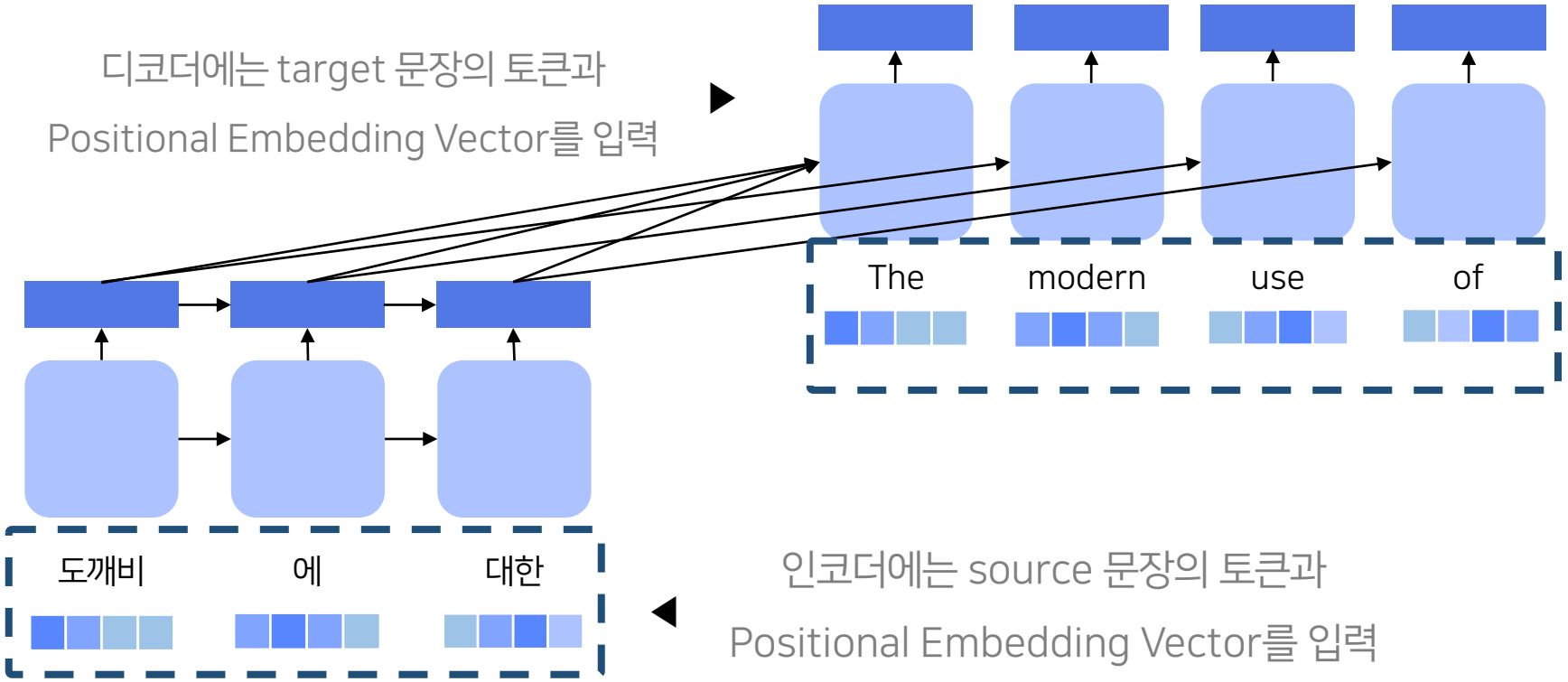
[0, 234, 34, ···, 34, 451, 203]

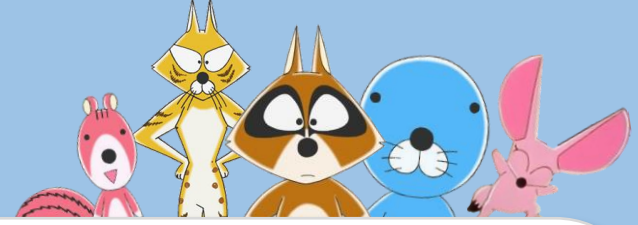
[0, 1234, 3, ···, 34, 456, 234]



기계번역

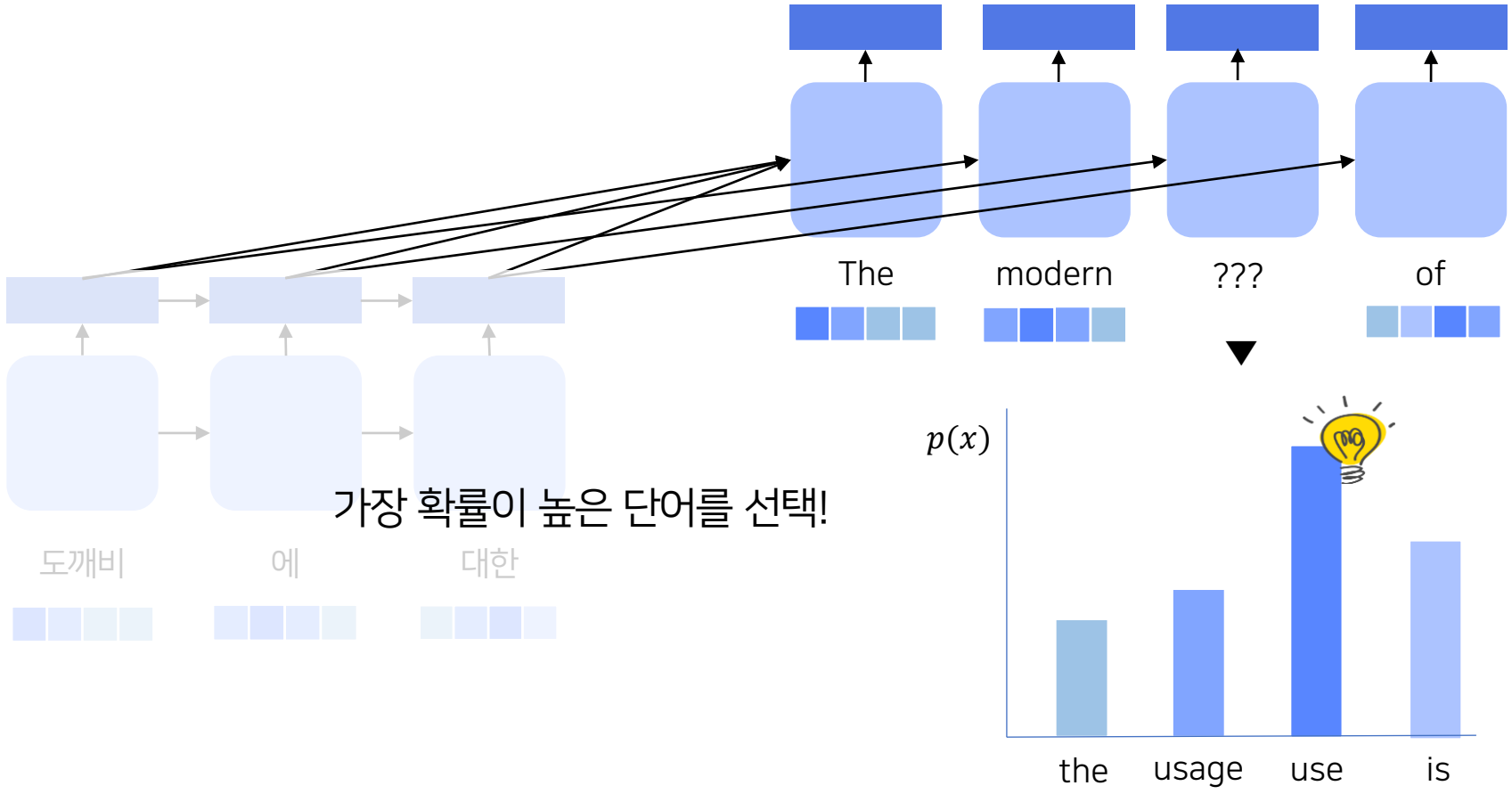
Step 3. Transformer 모델을 이용한 학습

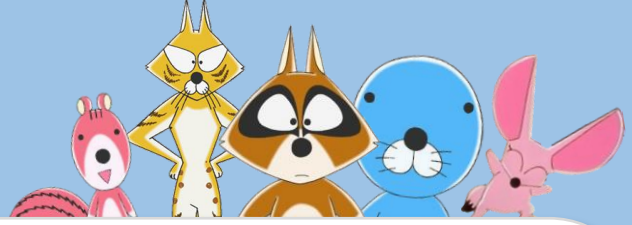




기계번역

Step 4. 학습된 모델을 통해 단어를 한 단어 씩 예측





BLEU(Bilingual Evaluation Understudy)

BLEU란?

기계 번역의 성능이 얼마나 뛰어난가를 측정하기 위해 사용되는 대표적인 방법

측정 기준은 n-gram에 기반

n- gram 이란?

"N개의 연속적인 단어 나열"

n-gram 언어 모델은 카운트에 기반한 통계적 접근을 사용

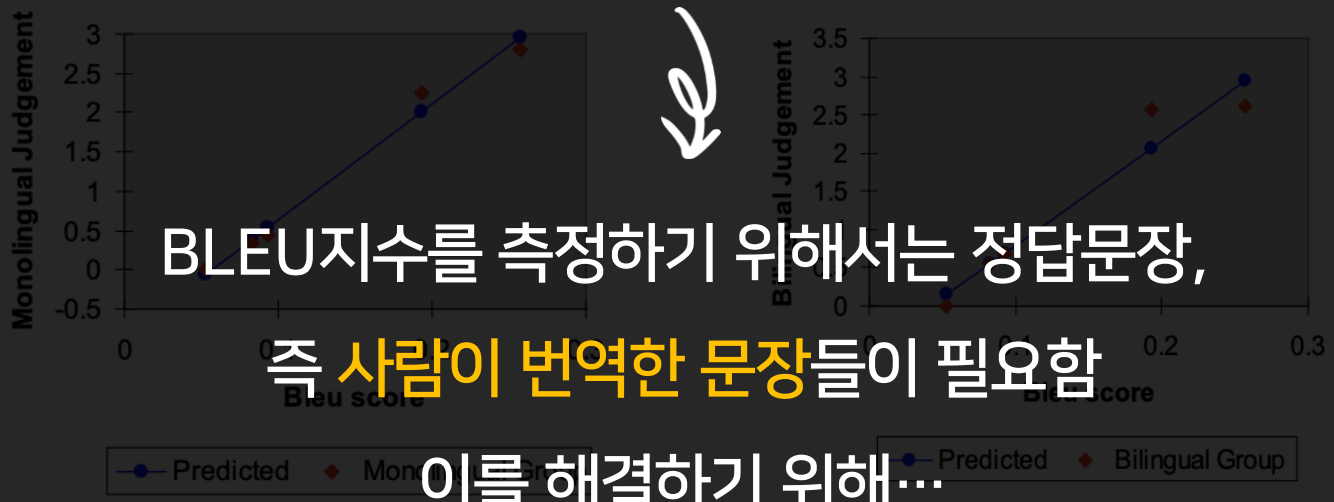
이전에 등장한 모든 단어를 고려하는 것이 아니라 일부 단어만 고려

*이때 일부단어의 개수가 n을 의미함

3. BLEU(Bilingual Evaluation Understudy)

사람이 직접 번역한 평가한 결과 VS BLEU Score 비교

하지만 여기서 한가지 문제가 발생



BLEU지수를 측정하기 위해서는 정답문장,

즉 사람이 번역한 문장들이 필요함

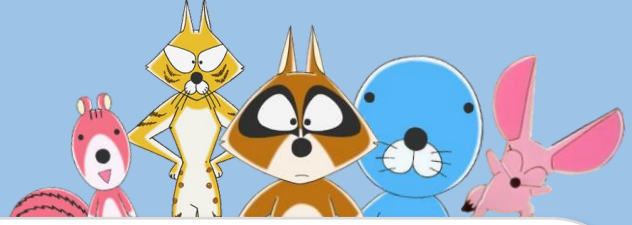
이를 해결하기 위해…

▲ monolingual인 사람의 평가 결과

▲ bilingual인 사람의 평가 결과

실제 사람이 평가한 결과와 BLEU SCORE가 거의 유사

토이 프로젝트



기계번역의 성능 평가

P-SAT의 영어 능력자분들께 도움을 받았습니다

고분회귀 징글팀장

선대 Fox

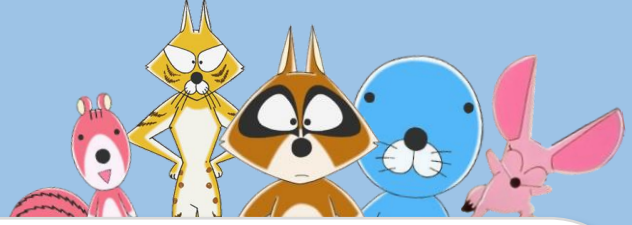
열정데마 썬샤인

딥러닝 다니엘

시계열 금수저

범주 실세

토이 프로젝트



기계번역의 성능 평가

Source	40년에 걸쳐 형성된 부평해물탕거리에서는 인천 앞바다에서 공수해온 제철의 싱싱한 해산물…
Reference Feat.(P-SAT)	Formed over the course of 40 years Bupyeong Haemultang Street offers Haemultang cooked with fresh seasonal seafood …
Transformer	the fukujuen was a full service country in which was added over 60 years old and it can be said that the instant street in the middle.

BLEU score -> 5.5026e-155

Cumulative 1-gram: 0.481481

Cumulative 2-gram: 0.136083

Cumulative 3-gram: 0.000000

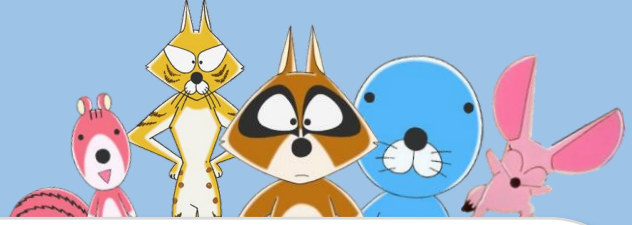
Cumulative 4-gram: 0.000000

1-gram과 2-gram에서는 어느정도 점수가 나왔지만,

3-gram부터는 점수가 0점이 나와

최종적인 BLEU score가 매우 낮음

토이 프로젝트



기계번역의 성능 평가

문맥에 맞지 않은 단어,
고유명사의 부족,
적은 데이터 셋의 개수

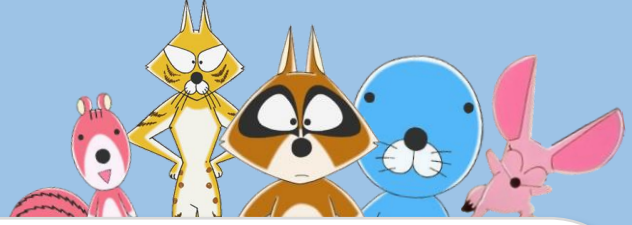
→

낮은 N-Gram으로 인한
낮은 BLEU Score

→

낮은 기계번역 성능 때문에
파파고를 이용하여
overview 번역

토이 프로젝트



기계번역의 성능 평가

Source	40년에 걸쳐 형성된 부평해물탕거리에서는 인천 앞바다에서 공수해온 제철의 싱싱한 해산물…
Reference Feat.(P-SAT)	Formed over the course of 40 years Bupyeong Haemultang Street offers Haemultang cooked with fresh seasonal seafood …
Papago	At Bupyeong Seafood Soup Street which has been formed over 40 years you can t aste seafood soup boiled with seasonal fresh seafood …

BLEU score -> 0.764851

Cumulative 1-gram: 0.962963

Cumulative 2-gram: 0.902671

Cumulative 3-gram: 0.838628

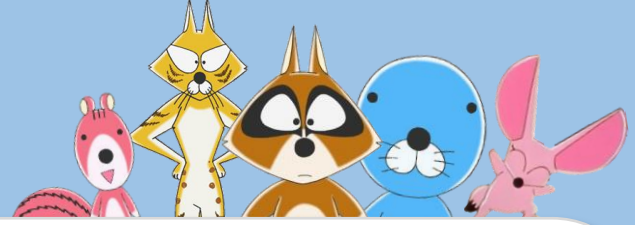
Cumulative 4-gram: 0.764851

1-gram에서 높은 값을 보여주고,

n이 커질수록 점점 줄어들지만

감소폭이 크지 않아 최종 BLEU score는 준수함

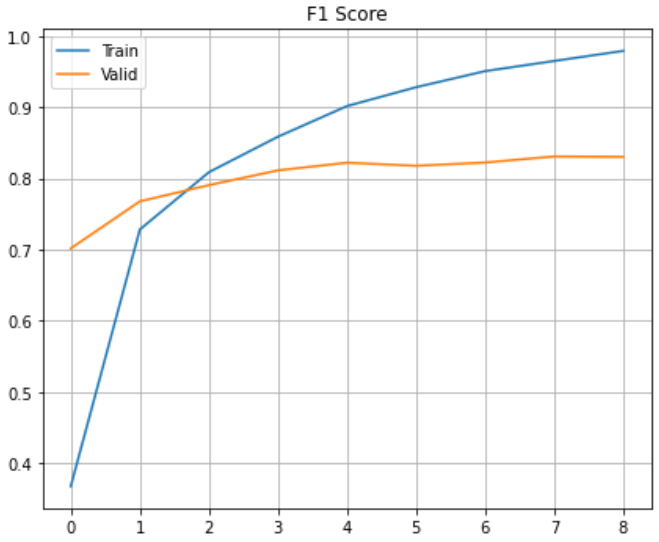
토이 프로젝트 결과



한국어 데이터와 한-영 번역데이터 성능 비교

F1- SCORE 비교

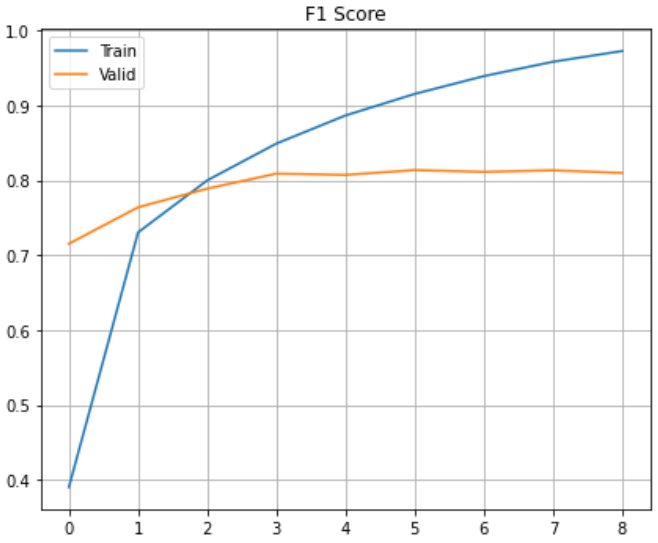
한국어 데이터 분류한 모델



▲ Multi Modal

Test F1 : 0.8728

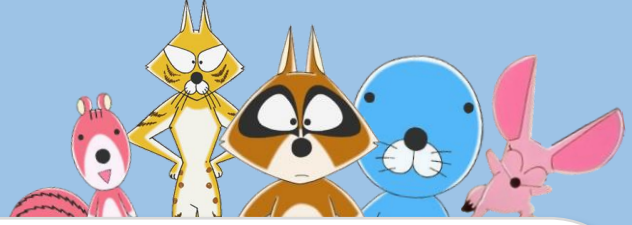
영어 데이터 분류한 모델



▲ 한영 번역 모델

Test F1 : 0.8459

토이 프로젝트 결과



한국어 데이터와 한-영 번역데이터 성능 비교

한-영 번역 모델이 성능이 낮은 이유는?

다음의 이유일 것으로 추측

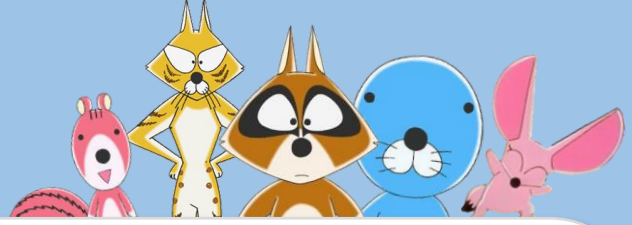
① 번역 성능의 불완전성

② 문맥을 충분히 반영하지 못함

"혜화수산" 이라는 상호명을

"Hyehwa Seafood Restaurant"가 아니라 "HyehwaSUSAN"으로 반영

토이 프로젝트 결과



KeyBERT 결과 비교

한국어 기반

수도권에서 가까운 위치, 문산천을 따라 걷는 산책코스, 한여름 더위를 날려버릴 시원한 물놀이장 등 가족이 함께 즐기기 좋은 캠핑장이다…

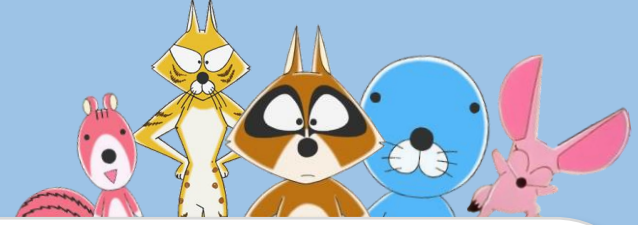
위치 # 눈썰매장 # 캠핑장
한여름 # 문산천

번역 모델 기반

Located close to the metropolitan area, it is a good camping site for families to enjoy together, including a walking course along Munsancheon Stream, and a cool water playground to blow away the midsummer heat...

camping # outdoor # summer
playground # pool

토이 프로젝트 결과



최종 결과 출력

Psat_deep
Hyewha, Seoul

♥ 610 Likes

Psat_deep

Strawberry cream cake is a famous bakery.
The representative menu is bread. It is a cafe located in Songpa-gu, Seoul.
#food #restaurant #bar/café #bakery #strawberry #cake #cafe #seoul

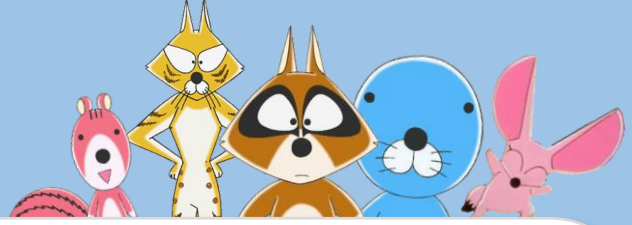
Psat_deep
Hyewha, Seoul

♥ 610 Likes

Psat_deep

딸기생크림 케이크가 유명한 베이커리이다.
대표메뉴는 빵이다. 서울특별시 송파구에 있는 카페다.
#음식 #음식점 #바/까페 #카페 #생크림 #송파구 #베이커리 #딸기

성과 및 한계



의의

의의, 기대효과

① 관광정보의 생산을 인공지능의 힘으로 자동화

더 적은 공공의 예산으로 더 많은 POI 데이터 만들 수 있음

② 이를 해시태그로 활용하여

SNS상에서 국내 관광정보의 접근성을 향상할 수 있음

→ 국내 관광 활성화에 도움