제 2회 삼성카드 데이터 분석 & 아이디어 공모전

https://yerimoh.github.io/
DON'T COPY IT

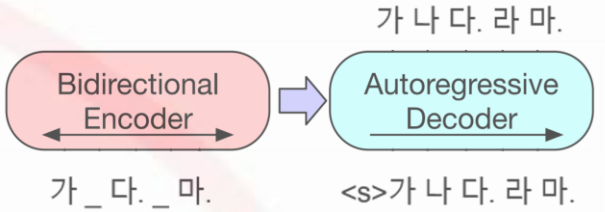
Track1: 고객 피드백 분류 모델 개발

TAVE 팀 김영현, 오예림, 박제윤

Part 1

알고리즘 설명, 선택 사유, 네트워크 & 파라미터 설계

한국어 Large-scale model: KoBART(**B**idirectional and **A**uto-**R**egressive **T**ransformers)



[그림 1] Text Infilling

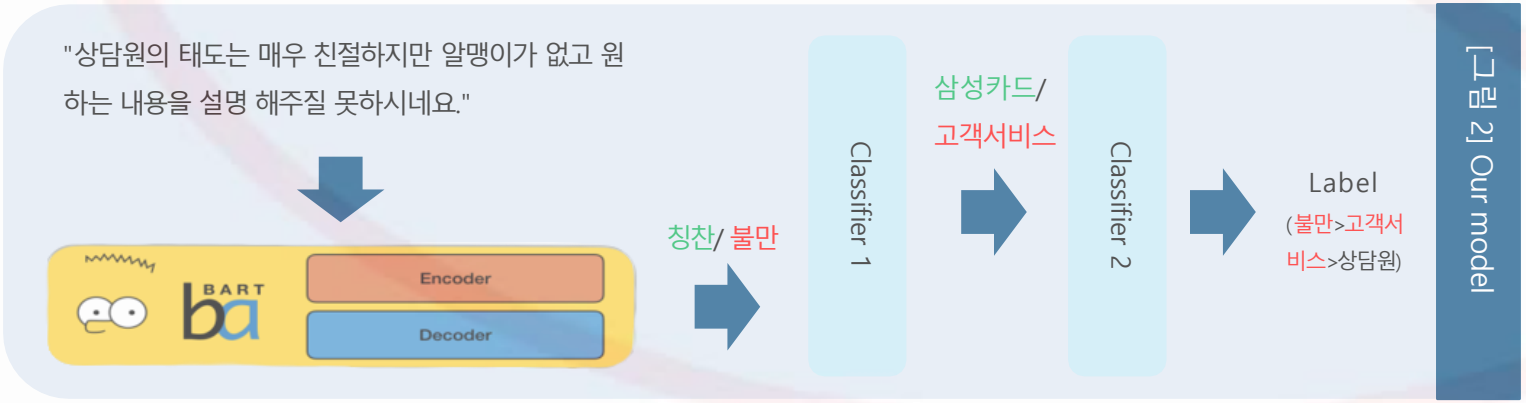
- Bidirectional Encoder와 Auto-regressive Decoder의 Seq2Seq 구조의 모델
- 입력 텍스트 일부에 노이즈를 추가하여 이를 다시 원문으로 복구하는 autoencoder의 형태로 학습이 됨
- [그림1]: Text Infilling 노이즈 기법을 사용한 모델 학습
- Text Infilling: 임의의 길이 만큼 토큰들이 하나의 [MASK] 토큰으로 대체되고, 모델은 얼마나 많은 토큰들이 대체되었는지 예측하는 방식
- 선택사유: 기존 Masked language model들과 다르게 task 응용력이 좋고, 특히 classification 성능이 매우 우수함
- [표1]: 모델 파라미터 설계 (KoBART)

Hyperparameter	Value
Encoder layer	12
Encoder attention heads	16
Decoder layer	12
Decoder attention heads	16
Dropout	0.1
Batch size	128
Max sequence length	128
Training Epoch	3
Learning rate	5e-5
Activation	GELU

[표 1] Model hyperparameter

The schematic depiction of our model: [그림 2]

- (1) KoBART를 사용하여 칭찬/불만으로 나누고, (2) 삼성카드/고객 서비스를 분류한 뒤, (3) 최종적으로 레이블을 예측함



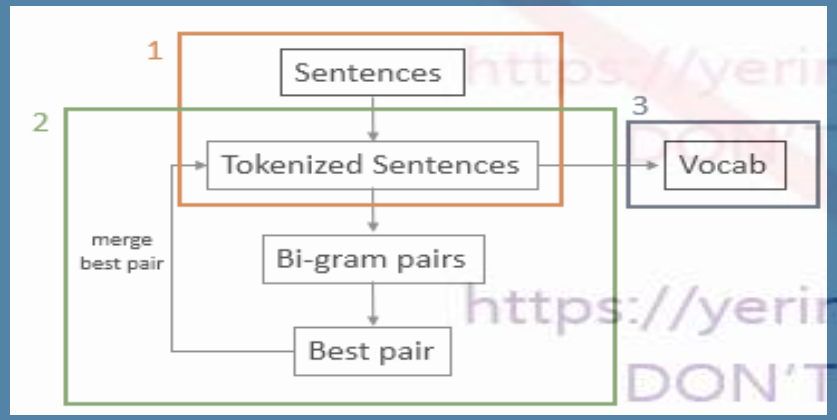
[그림 2] Our model

Part 2

입력 토큰 & Feature 구성 방법

입력 토큰: Character BPE(Byte Pair Encoding)

- BPE(Byte Pair Encoding) : 빈도 수에 기반해 단어를 의미 있는 패턴으로 잘라 토큰화 하는 Sub-word 분리 알고리즘



Ⓒ 입력 문장을 Character 단위로 분리

Ⓒ 가장 많이 등장하는 Bi-gram Pair을 찾아 병합하는 과정을 정해진 반복 수 만큼 수행

③ 수행 완료 후 병합된 토큰들로 Vocab 생성

➪ Word Embedding의 단점인 희귀 단어 및 신조어, OOV(Out-Of-Vocabulary) 문제 완화 가능

Part 3

칭찬/ 불만 레이블링 결과의 후처리의 방법

1 [칭찬/불만]의 1차 분류 후, 입력 문장과 하위 분류 단어(삼성카드/고객서비스/기타)와 유사도 비교

- 입력 문장을 BPE 방식의 서브 워드로 나눈 후, 하위 분류 단어와 서브 워드 마다 코사인유사도를 계산하고 이를 평균 내어 값을 비교
- 가장 높은 유사도의 하위 분류 단어로 2차 분류
- 만약 삼성카드/고객서비스의 유사도의 차이가 일정 값 미만일 경우 기타로 분류

2 [삼성카드/ 고객서비스/ 기타]의 2차 분류 후, 입력 문장과 세부 하위 분류 단어들과 유사도 비교

- 1번과 동일한 방식의 유사도 비교 후, 가장 높은 유사도의 세부 하위 분류로 최종 레이블링됨

예시)

"상담원의 태도는 매우 친절하지만 알맹이가 없고 원하는 내용을 설명해주질 못하시네요."

[불만]

[삼성카드] : 0.08
[고객서비스] : 0.9
[기타] : 0.02

코사인 유사도 비교

[고객서비스]

[상담원] : 0.64
[상담시스템] : 0.34
[혜택] : 0.02
.
.
.

[불만>고객서비스>상담원]

*BPE 과정은 생략

Part 4

복문에 대한 대응전략

1. 실험: 제공된 샘플 데이터 1,000건

- 1만 건의 데이터로 44개 클래스를 분류할 수 있는 모델을 구현하는건 매우 어렵기 때문에 효율적인 접근법이 필요하다고 판단함
- [표 2]: 실험 결과, 한국어 large-scale model은 제공된 샘플 데이터셋에 대해 칭찬/ 불만으로 나누는 task에서 매우 높은 성능을 나타냄
- 특히, 복문 32문장에 대해서도 83%라는 높은 성능을 보여줌

Dataset	Accuracy
SamsungCard	0.887
SamsungCard (Complex sentence)	0.830

[표 2] KoBART를 NSMC dataset에 fine-tuning 시킨 후, 제공된 샘플 데이터셋으로 inference한 결과

2. 대응전략

- 복문을 위해 한 문장을 44개의 클래스로 분류하는 문제로 풀기에는 데이터 구축과 모델 학습에 너무 많은 비용이 소모됨
- [표 2]와 같이, 복문에도 robust한 Pretrained large-scale model을 사용한 모델링을 통해 비용을 정말 많이 줄일 수 있음
- 이후 후처리를 통해 classification을 하여 성능도 우수하며 비용도 크게 절감한 모델을 구현하여 대응하려함
- 제공되는 데이터 외에도 추가적으로 데이터를 수집하여 성능을 끌어올리려고 함