기업과제4_3팀_신현지_개인보고서

1. Abstract

본 프로젝트는 Half-Longformer 모델을 사용해 뉴스 데이터의 생성요약(Abstractive Summarization)을 수행하였다. Half-Longformer 모델은 KoBART에 input token 허용 길이를 반으로 줄인 Longformer를 결합한 모델로, 데이터의 특성을 반영한 모델이다.

학습데이터는 제공된 스포츠 뉴스 기사 9077건 중 7000건에 AI-hub 문서요약 데이터 중 스포츠 뉴스 5566건을 추가하였다. kakao pororo API로 생성한 Ground Truth로 학습 및 튜닝한 모델 테스트 결과, Rouge-L F1 score 기준 0.51을 기록했다.

2. Data

2.1 True Summary

1. **Ground truth : TITLE**
    - 기존 데이터의 ‘TITLE’ 을 Ground truth 로 생성요약을 진행 결과, TITLE 의 길이가 짧아 CONTENT 의 내용을 온전하게 요약하지 못했다.
    - ‘TITLE’에서 ‘GOAL LIVE’, ’GOAL 리뷰’, ’현장목소리’ 등 CONTENT 의 내용과 관계가 없는 내용을 제거하고, 생성요약을 진행했지만 지나치게 함축적이어 CONTENT를 요약한 문장이라고 볼 수 없다.
    - 따라서 생성요약 모델의 성능 향상을 목적으로 Ground truth를 재정의 하는 것이 필요하다고 판단하였다.

2. **Ground truth 재 정의 : Pororo**

- Pororo는 카카오브레인에서 제공하는 NLP Platform로 Text Summarization에서 input 문장 제약이 비교적 자유롭고, sample 확인 결과 본문의 핵심적인 내용을 포함해 요약함을 확인하였다.
- 이후 선정한 **Best model 기준** TITLE 을 label로 학습 한 결과보다 **0.06 더 높은 Rouge L** 기록했다.

2.2 Data Pre-processing

1. **9077건의 뉴스 기사 데이터 (sports_news_data.csv)**
   - Content 가 누락된 2건은 drop 하였다.
   - 데이터 수집 과정에서 발생한 원문과 관계없는 아래 내용을 정규식으로 제거하였다.

항목	예시
사진관련(게티이미지코리아, '사진='등)	사진=OOO 인스타
Twitter-Box	blockquote class="twitter-tweet"
광고 문구 및 특수문자	☆★○●◎◇
줄바꿈 및 중괄호	[스포탈코리아]
기자이름	OOO기자
이모티콘	👍
html tag	...

2. **AI hub에서 수집한 5500건 스포츠 뉴스 기사 추가 데이터**
   - 기존의 데이터의 특성을 고려해 스포츠 카테고리에 속한 신문기사만 추가 데이터로 수집하였다.
   - 생성 요약문이 제공되어 있고, 불필요한 내용에 대한 제거 방식은 위와 동일하다.

3. Model

1. 모델 선택에 관한 근거

a. **KoBART**

- SKT-AI에서 제공하는 한국어로 사전학습한 BART 모델
- 한국어 위키 백과,뉴스,책 모두의 말뭉치,청와대 국민 청원 등 다양한 한국어 데이터로 모델 학습
- BART 모델은 기본적으로 noising이 된 input 값을 넣어, fully content를 만드는 방식으로 학습하는 Encoder-Decoder 모델
- 원래 문장(fully content)를 예측하는 방식으로 학습하기 때문에 생성요약에 강점을 보인다.
- reference
    - https://github.com/SKT-AI/KoBART

b. **Longformer Attention Layer + KoBART**

- KoBART보다 긴 길이의 문장을 학습이 가능하다.

  1) *Sliding window Attention*

  2) *Dilated sliding window*

  3) *Global attention*

  방법을 통해 전체에서 부분의 정보를 학습한다.

- input으로 받을 수 있는 길이는 길어지게 학습하지만, 유실되는 데이터가 존재함에 따라 입력 가능한 길이와 데이터 학습은 부분적으로 trade-off 관계를 가진다.
- KoBART 모델은 수용가능한 최대 token 길이(Max Length)가 1024이다.
- 제공된 뉴스 기사 데이터에서 Max Length가 1024를 넘는 데이터가 90건이 존재하는 것을 확인하였다.
- 데이터의 유실을 최소화하기 위해 token의 최대길이를 더 많이 수용가능한 모델이 필요해 Longformer Attention Layer + KoBART모델을 사용하였다.
- reference
    - https://github.com/Taeksu-Kim/longformer_kobart
    - https://github.com/allenai/longformer

c. **Half-Longformer Attention Layer + KoBART**

- 모델학습에 사용하는 데이터 특성상 token의 길이가 1024를 넘는 데이터들이 대부분 2048미만의 길이를 가진다.
- 기존 Longformer Attention Layer의 Max Length를 반으로 줄인 2048로 설정했다.
- KoBART,Longformer Attention Layer + KoBART, Half-Longformer Attention Layer + KoBART를 통해 학습한 결과 가장 높은 성능을 기록하고 , 학습 시간이 절반으로 단축된 Half-Longformer Attention Layer + KoBART를 모델로 선정하였다.

[Input sequence Max Length]

Model	Max Length
KoBART	1024
Longformer Attention	4096
Half-Longformer Attention	2048

4. Metrics

1. Rouge Score

- 보편적으로 텍스트 요약모델에 대해 평가지표로 사용되는 Rouge Score를 평가지표로 사용하였다.
- Label과 모델이 생성한 Summary를 비교해 성능을 확인하는 Rouge Score 계산을 위하여 Recall, Precision, F1-Score을 모두 확인하였다.
- Recall은 참조 요약 중 생성 요약과 얼마나 많은 단어가 겹치는지를, Precision은 생성 요약 중 얼마나 많은 참조 요약과 얼마나 많은 단어가 겹치는지 확인하기 때문에, 한쪽에 치우치지 않고 성능을 평가하기 위해 Recall과 Precision의 조화 평균값인 F1-score을 주요 평가지표로 삼았다.

2. Metric : Rouge-L, F1-Score

- Rouge-L은 LCS를 사용해 단어의 연속일치가 아닌 문장 수준의 단어 순서를 반영하여, 가장 긴 일치 단어 시퀀스를 측정할 수 있다.
- 시퀀스 내 가장 긴 공통 n그램이 자동으로 포함되며, 보다 유연한 성능 비교가 가능하여 해당 metric을 최종 선정하였다.

Rouge1	Label 과 Summary 중 겹치는 unigram 수
Rouge2	Label 과 Summary 중 겹치는 bigram 수
Rouge3	Label 과 Summary 중 겹치는 trigram 수
RougeL	최장 길이로 매칭되는 문자열을 측정 n-gram과 달리 순서나 위치관계를 고려한 알고리즘

Recall	겹친 단어 수 / 참조 요약 내 전체 단어 수
Precision	겹친 단어 수 / 모델 생성 요약 내 전체 단어 수
F1-score	F1 score는 precision 과 recall의 조화평균

5. Parameter Tuning

- 아래 모델에 대해 평가지표의 성능 향상을 위해 파라미터 튜닝 진행
  - Kobart
  - H-Kobart : Half-Longformer Attention + KoBART

Model	Y	Data	Max_length	Max_pos	Batch Size	Epoch	Optimizer	Learning Rate
`KoBART`	Pororo	9077	1024	256	8	2	Adam	5e-5
`H-KoBART`	TITLE	9077 + 5500	2048	2052	4	1	AdamW	1e-5
`H-KoBART`	Pororo	9077 + 5500	2048	2052	4	1	AdamW	1e-5
`H-KoBART`	Pororo	9077 + 5500	2048	2052	4	4	AdamW	1e-5

6. Result

KoBART (Pororo)	Recall	Precision	F1-Score
Rouge1	0.68	0.43	0.52
Rouge2	0.50	0.33	0.39
Rouge3	0.42	0.28	0.32
RougeL	0.59	0.38	0.45

Best Model (TITLE)	Recall	Precision	F1-Score
Rouge1	0.41	0.45	0.42
Rouge2	0.21	0.23	0.21
Rouge3	0.12	0.13	0.12
RougeL	0.59	0.38	0.45

Best Model (Pororo)	Recall	Precision	F1-Score
Rouge1	0.59	0.57	**0.56**
Rouge2	0.45	0.45	**0.44**
Rouge3	0.38	0.38	**0.37**
RougeL	0.54	0.52	**0.51**

- **Groud Truth(y)**를 **TITLE**로 설정한 Half-long model 보다 **Pororo Summary**를 설정한 Half-long model 의 Rouge F1-Score이 모두 큰 차이로 높다

  → `Ground Truth 재 정의가 성공적인 시도로 판단된다`

- F1- Score를 기준 **Best Model(Half-long)** 이 KoBART 보다 **0.4~0.6 더 높은 점수** 기록

7. 한계점 및 보완점

- 시간과 비용의 문제로 인해 모델 학습을 더 다양한 방법으로 시도해 보지 못해 아쉬웠다.
- Rouge Score에 대해 단점이 많다는 견해를 접해, 새로운 평가지표인 RDASS를 통해 학습해 보지 못해 이를 보완해 나갈 예정이다.

8. 자신이 담당한 역할

- Longformer KoBART 모델 파인 튜닝
- LongformerAttention 구조에 대한 논문 및 자료 리서치
- 추가 데이터(AIhub) 정제 및 공유
-