한국어 스포츠 기사의 요약모델 개발

장 수 림 | think.percento@gmail.com

담　　당 | Preprocessing, Define Ground truth & Metric

Contents

Our Task

한국어 스포츠 기사의 주요한 핵심 내용을 요약하는 모델을 개발해야한다.

요약은 `본문의 주제`를 담아야하며, 이용자가 `이해하기 쉬운 형태`(정제된 글자, 간결한 문장 수)와 `의미`(독해가능한 문장 구성)로 도출되어야한다.

데이터셋의 개요

`sports_news_data` : 약 6개월 간의 축구 뉴스 기사 데이터로 한 개 이상의 언론과 여러 명의 기자에 의해 생성

- 컬럼 : {Title: 기사 제목, Content : 기사 내용, Publish_DT : 발행일}
- 총 9,077개의 레코드

Define Ground Truth and Metric

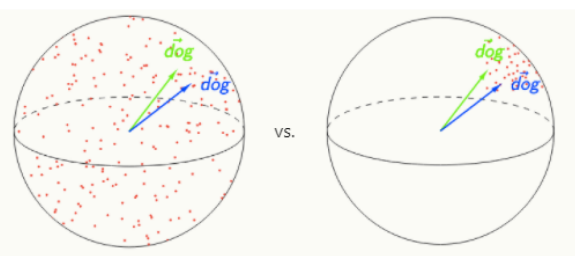
1. Ground Truth - Y를 어떻게 만드는가?

1.1. 몬테카를로 시뮬레이션을 차용한 방법

"`몬테카를로 시뮬레이션`**은** 불확실한 사건의 가능한 결과를 추정하는 데 사용되는 수리통계적 기법이다."

본 과제는 데이터의 내재된 요인도 분포도 아무것도 모르지만 모델을 평가해야하는 현실적인 문제를 다룬다.
이렇게 불확실한 결과를 추정하는 의사결정 시 '**실행을 무수히 반복해 추정된 값 범위를 기반으로 결과 세트를 예측**'하는 몬테카를로 기법을 적용할 수 있다.

만들어진 Ground Truth의 궁극적인 모사점이 인간이 의미적·형태적으로 인지가능한 요약이라 할 때, 그 값은 원데이터 특유의 성질을 잘 보존한 값일 것이다. 따라서 본 방법론은 원데이터의 분포를 기반으로 생성된 Ground Truth를 차용하기 때문에 **데이터의 Uniformity가** 지켜질 것으로 사료된다.



[그림1] 좌 - Uniformity, 우 - Anisotropic

가 설 | 몬테카를로 시행을 통해 구한 (요약)데이터 분포의 대표값은 원데이터의 고유성을 보존할 것이다.

적 용 |

```
예) 3개의 문장으로 구성된 스포츠 기사가 있다.

1. 단어 또는 문장의 순서를 무작위로 섞는다.

  >>> RandomSwap(['손흥민이 헤트트릭에 성공했다', '역사상 가장 짜릿한 역전골이네요'])
  ['헤트트릭에 손흥민이 성공했다', '짜릿한 역전골이네요 역사상 가장']

2. 1을 k번 반복한다.

  k = {50,100,500..}, 분포를 이루기에 적당한 값을 선정

3. 구해진 k개 문서를 선정한 모델을 이용해 요약한다.

4. 구해진 k개 요약된 문서의 모든 pair의 코사인 유사도를 구한다.

5. 코사인 유사도의 mean(혹은 max)에 가장 가까운 값 pair를 고른다.

6. 두 요약된 문서 중 임의의 한 문서를 y_true로 선정한다.
```

1.2. 검증된 기존 모델을 차용한 방법

“kakaobrain의 공개 모델인 `PORORO`로 추출 요약을 진행해 도출된 요약을 Ground Truth로 사용.”

분석을 위해 사용한 토크나이저는 서브워드 기반의 bpe8k이며, 사전학습 모델은 RoBERTa를 다음의 말뭉치를 학습시킨 `brainbert.base.ko.summary` 를 사용하였다.

- dataset: Dacon summarization corpus + AI Hub summarization corpus (1st release)
  - ref: https://dacon.io/competitions/official/235671/data/
  - ref: https://www.aihub.or.kr/node/9176
- metric: Rouge-1 (42.67), Rouge-2 (31.80), Rouge-L (43.12)

이렇게 생성된 Ground Truth는 **분석 모델을 비교대조 평가**할 수 있다는 의의를 갖는다. 단, 이 경우 검증 단계에서 나온 score를 요약 품질의 좋고 나쁨으로 정량화하는 해석력이 떨어질 수 있으므로 결과 평가에 신중해야한다.

가 설 | 기존 모델을 통해 구한 요약 데이터는 요약모델의 정량적 성능을 평가하는 데에 유효할 것이다.

적 용 | PORORO로 추출 요약한 결과를 Ground Truth로 사용한다.

1.3. 기사 제목을 차용한 방법

“기사 제목은 데이터의 생성자가 생성한 Ground Truth라고도 비유할 수 있다.”

가 설 | 기사 제목은 전문가의 요약 데이터이므로 요약모델의 정성적 성능을 평가하는 데에 유효할 것이다.

적 용 | 기사 제목을 (분석과 동일한 방법으로)전처리하여 Ground Truth로 사용한다.

단, 이 방법은 분명한 한계를 지니고 있다. 기자마다 헤드라인 선정 요인이 각기 다를 것이라 쉽게 추측할 수 있기 때문이다. (A는 주제를 중점으로, B는 클릭베이츠 헤드라인을 정할 수 있다.)

요인 통제가 불가하기 때문에 요약의 품질 또한 신뢰하기 어려워 분석 비교를 위한 잠정적 비교군으로만 남겨두고 이 방법은 실제 채택하지 않는다.

1.4. Ground Truth 선정

각 방법론은 각기 다른 평가 기준을 지닌다.

데이터의 inherent를 지켜 **요약의 품질을 평가**할 것인지(1), 모델 간 성능 비교를 통해 **도메인 데이터에 적합한 모델인지 평가**할 것인지(2)에 따라 Ground Truth를 선정하는 논의 과정이 수반되었다.

가장 큰 선정 요인은 과제 수행을 위해 주어진 시간이 3일뿐이란 현실적 제약과 타협하는 것이었다.
팀은 ‘Lean한 분석’을 모토삼아 먼저 구현한 방법론을 채택하는 것으로 합의했고, **검증된 기존 모델인 PORORO의 요약값을 Ground Truth로 사용**하기로 최종 결정하였다.

---

2. Metric - 어떻게 평가하는가?

요약 모델을 Co-selection(동시 선택), Ground Truth를 Content-based(내용 기반)으로 각각 평가한다.

2.1. Model : Rouge Score

추출 요약 기법은 단어의 빈도, 위치, 유사성과 같은 통계적·언어적 특성을 이용해 요약 결과를 도출한다.

`Rouge-n Score` 은 n-gram을 통해 정확도(Precision)과 재현률(Recall)을 산출한 뒤 최종적으로 F-Score 결과를 측정하는 방식으로 추출 요약의 주요한 정량적 평가지표로 사용된다.

따라서 실제 요약문과 구성된 요약문의 문장이 얼마나 일치하는지를 확인하는 메트릭인 Rouge-n Score를 사용하면 추출 요약 모델의 성능을 평가할 수 있다.

적 용 | 단어 간 긴밀한 관계를 관찰하기 위해 uni-gram인 **Rouge-1**을 채택한다.

2.2. Ground Truth : LSA 기반

'PORORO가 다른 모델로 생성한 요약에 비해 Ground Truth로서 더 좋은 모델인지 검증한다.'

요약문이 Ground Truth로서의 지위를 가지려면 요약 문서와의 완벽한 일치가 아닌, 얼마나 유사한지 여부를 측정하는 것에 초점을 맞춘 평가지표가 필요하다.

이러한 문제의식에 따라, **요약 모델이 실제로 본문의 주제를 잘 감지하고 있는지**를 검증하기 위한 머신러닝 방법론을 적용한 자체 지표를 구성하였다. LSA 또한 빈도에 영향을 받으므로 추출 요약에 적합한 지표로 보인다.

적 용 | 원데이터를 `Mecab`으로 토크나이징하고 `Tf-Idf`와 `Truncated SVD`를 이용해 학습한다. 모델이 도출한 요약문에 이와 같은 방식을 적용해 서로 비교한다.

- 요약문과 원데이터(원문)의 유사도를 품질의 척도로 사용
- 요약문과 원데이터의 문장 개수와 토큰 수가 상이하므로 차원 축소해 비교
- 축약된 원문과 요약문 사이의 벡터 산출해 비교

Preprocessing

1. 텍스트 전처리

관찰된 패턴에 따른 텍스트 전처리 내역은 다음과 같다.
(전처리 함수는 각 모델 인풋에 따라 수정되어 최종 채택된 전처리 코드와 구성이 상이할 수 있다.)

▼ HTML태그 정제

```
def HTMLCleaner(doc):
  """ 본문에서 불필요한 HTML Tags를 제거하고 문장을 나눕니다. """
  doc = re.sub('\n', '<p>', doc)                # Unify Newline Character
  doc = re.sub('\.', '.<p>', doc)
  soup = BeautifulSoup(doc, "html.parser")      # Handling HTML tags
  blocklist = ['strong']  # Remove specific tags
  for tag in blocklist:
    tagging_soup = soup.findAll(tag)
    for script in tagging_soup:
      script.decompose()
  clean_list = [sentence for sentence in soup.strings]
  return clean_list
```

- **줄바꿈 표지 통일**

  문장을 온점 단위로 분리할 계획이므로, 불필요한 줄바꿈 표지를 제거

- **불필요한 태그 제거**

  본문을 꾸며주는 태그나 본문과 관련없는 외부 문구(트위터 등)를 가져오는 태그를 삭제한다.

  Memo: `blocklist = ['span', 'strong', 'blockquote']` 적용 이후, 전문이 `<span>` 으로 묶이는 레코드가 발견되어 `<strong>`만 적용하고 나머지는 re로 제거하기로 결정.

▼ 텍스트 정제

```
def DocCleaner(text):
  """ 개별 문서에 적용해 문서를 전처리하는 함수입니다. """
  doc = HTMLCleaner(text)
  if doc:
    if doc[0].startswith('['):                # Remove media source
      doc[0] = re.sub('^\[.*?\] ', '', doc[0])
      doc[0] = re.sub(r'^.*?= ', '', doc[0])  # Remove reporter name
  if doc[-1].startswith(('사진', '그래픽')):   # Remove photo source
      doc = doc[:-1]

  doc = ' '.join(doc)
  doc = re.sub('\\\'', '', doc)        # Remove escape code
  doc = re.sub('\<.*?\>', '', doc)     # Remove remained tags
  return doc
```

- **언론 표지 제거**

  '[언론사1], [언론사2]' 등의 언론 표지를 제거

- **기자명 제거**

  '000 기자 =' 와 유사한 패턴의 기자명 제거

- **사진 출처 표지 제거**

  '사진 =, 그래픽 =' 과 유사한 패턴의 출처 표지 제거

- **이스케이프 코드 제거**

  '\'손흥민\'' 과 같이 역슬래시가 남아있을 시 이를 제거

- **HTML 태그 제거**

  남아있을 수 있는 HTML를 후처리

▼ 결측치 처리 및 전처리 수행

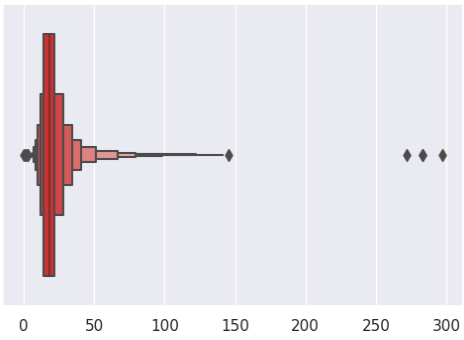
```
def Preprocessing(df, title, col='summary'):
  """ 전체 데이터프레임에 적용해 문서를 전처리하는 함수입니다. """
  # Handling the missing values
  # We'll use the title as the text and its summary. (제곧내!)
  df[col] = df[col].fillna(title)
  # Cleaning the documents
  df['SUMMARY'] = df[col].apply(lambda x : DocCleaner(x))

  return df
```

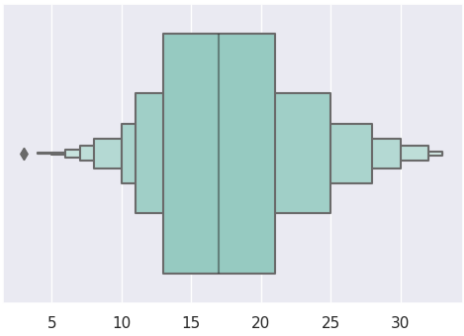
- **결측치 발생 시 제목을 본문으로 대체**

  기획 기사가 적고 단문 형태의 속보(경기 소식 전달)가 잦은 도메인 데이터를 특성을 고려해 결측치가 발생하면 제목을 본문으로 대체했다. 이에 따라 총 2건의 결측 레코드가 대체되었다.

2. 이상치 핸들링



[그림2] 스포츠 기사의 문장 수 분포



[그림3] 이상치 제거 시 (IQR 기준)

본 데이터셋의 문장 수 분포를 살펴보면, 위와 같이 문장이 과도하게 많거나 아예 없거나 적은 문서가 데이터셋 내에 존재함을 알 수 있다.

만일 이상치를 처리한다면 (1)**텍스트 데이터의 특성에 맞는 처리 방법을 모색**해야할 것이고, 그 결과 (2)**모델이 다양성을 지닌 현실의 데이터를 학습**할 수 있도록 해야할 것이다.

그러나 결측치를 제목으로 대체한 것과 다르게, 긴 문서와 짧은 문서를 **특정 텍스트로 대체**하는 것은 추가적인 고려가 필요하다. 데이터를 탐색한 결과, 기사의 내용은 단순히 경기와 선수에 관한 내용뿐만 아니라 부고부터 절반 이상이 광고인 기사까지 다종다색이었다. 그 다양성을 간단한 전처리 기법으로 일반화하는 것은 바람직하지 않다.

또한, 일괄적으로 **이상치를 제거**할 경우에도 학습 데이터의 다양성을 해쳐 결과적으로 모델의 보편성을 잃게 한다.

따라서 이상치를 스포츠 기사가 지닌 inherent의 일부로 보고, 처리하지 않는 것으로 결정한다.

Modeling

1. 모델링 요약

데이터 마다 고유한 내재적 특성이 있을 것이다. 따라서, 그 **특성을 더 효과적으로 파악하는 모델**을 찾기 위해 단일 모델의 탐색보다 다음의 **4가지 모델 방법론** 각각을 적용해가며 최적의 모델을 선정하는 전략을 사용함.

- `koBERTSum` **채택** ✅
- `koBART`
- `PORORO` `(abstractive)`
- `MatchSum`

2. koBERTSum

최종 채택된 `koBERTSum`의 모델 학습 과정은 다음과 같다.

- **Pre-train**

  using `KoBERT` `(with Dacon - Bflysoft datasets)`

- **Fine-tunning**

  using `koELECTRA` `(with Aihub - News Summary datasets)`

  - **hyperparams**
    - batch_size = 5000

        - ext_dropout = 0.1
        - max_pos = 512
        - lr = 2e-3
        - warmup_steps = 100
- **Validation**

  loss = `xent (CrossEntropy basis)`

    - **hyperparams**
        - batch_size = 3000
        - test_batch_size = 100
        - max_pos = 512
        - max_length = 200, min_length = 50
        - alpha = 0.95

Conclusion

1. 모델 성과

ROUGE-1

추출 생성

Model	koBERTSum	koBART	PORORO
PORORO (y true)	**0.625**	0.468	0.423

ROUGE-L

추출 생성

Model	koBERTSum	koBART	PORORO
PORORO (y true)	**0.625**	0.465	0.418

LSA 기반 지표

추출 생성

Ground Truth	PORORO (baseline)	koBERTSum	Title	PORORO (abstractive)	koBART
Score	**1461**	1443	1278	1191	1156

- 일치성에 기반한 '좋은 요약'이 언제나 본질적 의미에서의 '좋은 요약'은 아니다

요약 모델을 검증하며, 추상 요약을 추출 요약에 적합한 Rouge Score로 평가하는 것이 부적절하다는 것을 확인할 수 있었다. 먼저, 추출 요약으로 만든 Ground Truth와 추상 요약을 비교하는 것 자체에 비약이 있었다. LSA 기반 지표로 추상 요약을 평가했을 때는 추상 요약된 모델이 Rouge Score에 비해 비교적 나은 점수를 얻었다.

2. 더 나아가야 할 지점들

현실의 데이터를 마주할 때 견지할 것들

- 데이터를 전처리할 때에 이용자의 언어패턴을 유심히 살펴 가지치기를 진행해야한다.
- 단순히 복잡한 매커니즘을 지닌 딥러닝 모델이 늘 이점을 지닌 것이 아니다. 만약, batch가 아닌 online으로 서빙이 이루어지는데 더 가볍고 빠른 머신러닝 모델이 비슷하지만 조금 낮은 성능을 보인다면 이를 채택해야한다.
- 데이터 전처리의 수준(greedy or not)부터 모델의 속도와 성능 간의 트레이드 오프까지, 이들을 결정하는 하나의 관점을 가져야 분석에 가치를 부여할 수 있다.

이번 과제를 통해 현실의 데이터가 모델에 적용될 때의 난점을 깊이 이해할 수 있는 기회를 얻었다.

실제 현업에서 풀어야할 문제들은 분석의 목표가 모델의 지표뿐 아니라 비즈니스 성과와 같은 특정 지표를 포함하는 관점에서 설계되어야한다는 관점을 체득할 수 있었다.

도메인 데이터의 고유성을 잘 학습하는 모델 추론이 필요하다

각 데이터셋에 내재된 고유성은 모두 다를 텐데, Generic한 모델을 사용하는 것이 늘 바람직할까?

모든 모델링에 사전훈련된 언어 모델을 사용하였지만, 이에 대한 비판적 시각을 늘 담지해야한다. PLM은 일관성을 향상되어 모델의 이해 수준이 높아지지만 문맥을 뭉뚱그려 이해할 수 있다는 단점을 지닌다.

Domain specific한 분석을 진행할 때에는 우선 내 모델이 데이터의 고유성을 잘 학습하는 아키텍쳐인지를 검증할 필요가 있다. 바로 PLM을 사용하거나 Scratch부터 학습하기 전에, 평가 지표를 설정해 해당 모델을 적용하는 것에 이점이 있는지 확인해 볼 수 있으면 좋겠다.

Cites

- **Datasets**

문서요약 텍스트

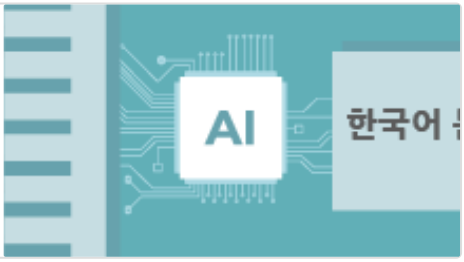
데이터셋명 문서요약 텍스트 데이터 분야 음성/자연어 데이터 유형 텍스트 구축기관 비플라이소프트 데이터 관련 문의처 담당자명 최재웅(비플라이소프트) 가공기관 비플라이소프트, 위고, 테스트웍스, 고려

https://aihub.or.kr/aidata/8054

한국어 문서 추출요약 AI 경진대회

1. train.jsonl - 학습에 사용 할 데이터셋 - media : 기사 미디어 - id : 각 데이터 고유 번호 - article_original : 전체 기사 내용, 문장별로 split되어 있음 - abstractive : 사람이 생성한 요약문 - extractive : 사

https://dacon.io/competitions/official/235671/data/



- **Model Architecture**

https://github.com/uoneway/KoBertSum

```
@misc{park2020koelectra,
  author = {Park, Jangwon},
  title = {KoELECTRA: Pretrained ELECTRA Model for Korean},
  year = {2020},
  publisher = {GitHub},
  journal = {GitHub repository},
  howpublished = {\url{https://github.com/monologg/KoELECTRA}}
}
```

```
@misc{pororo,
  author       = {Heo, Hoon and Ko, Hyunwoong and Kim, Soohwan and
                  Han, Gunsoo and Park, Jiwoo and Park, Kyubyong},
  title        = {PORORO: Platform Of neuRal mOdels for natuRal language prOcessing},
  howpublished = {\url{https://github.com/kakaobrain/pororo}},
  year         = {2021},
}
```