데이터 캡스톤 디자인 15주차 과제 진행내역

2014103929 김찬규

1) 작은 길이의 문장 데이터셋으로 축소

- 장문의 데이터셋인 CNN/Dailymail에서 Inshorts NEws Dataset (https://www.kaggle.com/shashichander009/inshorts-news-data) 로 변경을 하였습니다.

2) 텍스트 전처리

- i'm, you'll 등의 축약문을 펼쳐주는 contractions 단어사전을 이전과 같이 사용하였습니다.
- nltk stopword library를 사용한 stopword 삭제는 득보다 실이 많은 듯하여 수행하지 않았습니다. (having 등의 문자가 통째로 사라지는 문제가 발생하는 것을 발견하였습니다.)
- lemmatization, 표제어 추출은 decoder에서 사용할 시퀀스에만 적용하기로 하였습니다. 생각하여 보니, 만들어진 요약모델을 일반적인 환경에서 사용할 때는 lemmatize 된 텍스트가 입력되는 것이 아니기 때문에 encoder 시퀀스에까지 표제어 추출을 적용하게 된다면 정제된 입력데이터와 날것의 텍스트입력데이터 간에 요약문장이 차이가 나게 될 것입니다.
- 그동안 시퀀스의 최대 길이를 정의하고 그에 미치지 못하는 시퀀스에 대하여 padding만 진행하였었는데 최대길이를 넘어가는 시퀀스에 대하여 잘라주는 작업을 하지 않았다는 것을 깨달았습니다. 아마 이 때문에 제대로 된 결과가 안 나온것이 아닌가 생각이 듭니다.

3) Rouge metric

- 만들어진 요약모델을 평가하기 위한 metric입니다.
- Rouge에서 recall은 다음과 같이 정의합니다.

$$\frac{(\text{기계요약문과겹치는단어의수})}{(\text{정답요약문의전체단어})}$$

- Rouge에서 precision은 다음과 같이 정의합니다.

$$\frac{(\text{그중에정답요약문과겹치는단어수})}{(\text{기계가생성한요약문의단어수})}$$

- Rouge-1은 시스템 요약본과 참조 요약본 간 겹치는 unigram의 수를 나타냅니다.
- Rouge-2는 시스템 요약본과 참조 요약본 간 겹치는 bigram의 수를 나타냅니다.
- Rouge-N은 trigram, 4th-gram...
- Rouge-L은 최장 길이로 매칭되는 문자열을 측정합니다. Rouge-N과는 다르게 단어들의 연속적 매칭을 요구하지 않고, 어떻게든 문자열 내에서 발생하는 매칭을 측정하기 때문에 보다 유연한 성능 비교가 가능합니다.
- Rouge-S는 특정 Window size가 주어졌을 때, Window size 내에 위치하는 단어쌍들을 묶어 해당 단어쌍들이 얼마나 중복되게 나타나는 지를 측정합니다.(skip-gram)

4) 단편적인 결과

- 어느 정도 연관이 있는 단어들이 생성되는 것을 확인하였습니다.
- 생성한 요약문의 Rouge score는 약 0.1748(Rouge-1), 0.02813(Rouge-2), 0.1455(Rouge-L)입니다.
- 굉장히 낮은 점수이지만 아예 0점이 나오는 문장들을 제하면 평균점이 꽤 올라갑니다.

```
for i in range (len(article)):
    _hypothesis=summarize(article[i])
    _reference=summary[i]

    rouge=Rouge()
    scores=rouge.get_scores(_hypothesis,_reference)
    print(scores)
```

```
[{'rouge-1': {'f': 0.16666666166666683, 'p': 0.16666666666666666, 'r': 0.16666666666666666}, 'rouge-2': {'f': 0.0, 'p': 0.0, 'r': 0.0}, 'rouge-l': {'f': 0.08333332833333364, 'p': 0.08333333333333333, 'r': 0.0833333333333333
[{'rouge-1': {'f': 0.19047618557823143, 'p': 0.2222222222222222, 'r': 0.16666666666666666}, 'rouge-2': {'f': 0.0, 'p': 0.0, 'r': 0.0}, 'rouge-l': {'f': 0.19047618557823143, 'p': 0.2222222222222222, 'r': 0.1666666666666666
[{'rouge-1': {'f': 0.43478260415879016, 'p': 0.625, 'r': 0.3333333333333333}, 'rouge-2': {'f': 0.19047618603174613, 'p': 0.2857142857142857, 'r': 0.14285714285714285}, 'rouge-l': {'f': 0.399999954500001, 'p': 0.571428571
[{'rouge-1': {'f': 0.0, 'p': 0.0, 'r': 0.0}, 'rouge-2': {'f': 0.0, 'p': 0.0, 'r': 0.0}, 'rouge-l': {'f': 0.0, 'p': 0.0, 'r': 0.0}}]
[{'rouge-1': {'f': 0.0, 'p': 0.0, 'r': 0.0}, 'rouge-2': {'f': 0.0, 'p': 0.0, 'r': 0.0}, 'rouge-l': {'f': 0.0, 'p': 0.0, 'r': 0.0}}]
[{'rouge-1': {'f': 0.41666666222222226, 'p': 0.625, 'r': 0.3125}, 'rouge-2': {'f': 0.09090908657024814, 'p': 0.1428571428571428, 'r': 0.06666666666666667}, 'rouge-l': {'f': 0.4210526269252078, 'p': 0.5714285714285714, 'r'
[{'rouge-1': {'f': 0.16666666170138905, 'p': 0.18181818181818182, 'r': 0.15384615384615385}, 'rouge-2': {'f': 0.0, 'p': 0.0, 'r': 0.0}, 'rouge-l': {'f': 0.08333332836805586, 'p': 0.09090909090909091, 'r': 0.0769230769230769
[{'rouge-1': {'f': 0.0, 'p': 0.0, 'r': 0.0}, 'rouge-2': {'f': 0.0, 'p': 0.0, 'r': 0.0}, 'rouge-l': {'f': 0.0, 'p': 0.0, 'r': 0.0}}]
[{'rouge-1': {'f': 0.18181817698347122, 'p': 0.2222222222222222, 'r': 0.15384615384615385}, 'rouge-2': {'f': 0.0, 'p': 0.0, 'r': 0.0}, 'rouge-l': {'f': 0.18181817698347122, 'p': 0.2222222222222222, 'r': 0.153846153846153
[{'rouge-1': {'f': 0.19047618575963732, 'p': 0.25, 'r': 0.15384615384615385}, 'rouge-2': {'f': 0.0, 'p': 0.0, 'r': 0.0}, 'rouge-l': {'f': 0.0952380905215422, 'p': 0.125, 'r': 0.07692307692307693}}]
[{'rouge-1': {'f': 0.3157894688088643, 'p': 0.375, 'r': 0.2727272727272727}, 'rouge-2': {'f': 0.2352941120276827, 'p': 0.2857142857142857, 'r': 0.2}, 'rouge-l': {'f': 0.3157894688088643, 'p': 0.375, 'r': 0.272727272727272
[{'rouge-1': {'f': 0.22857142426122457, 'p': 0.36363636363636365, 'r': 0.16666666666666666}, 'rouge-2': {'f': 0.0, 'p': 0.0, 'r': 0.0}, 'rouge-l': {'f': 0.23076922603550304, 'p': 0.3, 'r': 0.1875}}]
```

그림 1 match되는 단어가 하나도 없어 0점이 나오는 문장이 꽤 많은 것을 볼 수 있습니다.