데이터 캡스톤 디자인 10주차 과제 진행내역

2014103929 김찬규

1) CNN/DailyMail 텍스트 데이터 전처리

- 이전 주에 빠트렸던 Lemmatizing을 추가하였습니다. lemmatization은 번역하면 표제어 추출이며, 단어의 원형을 찾는 작업이라고 말할 수 있습니다. 단어 원형을 찾는 일에는 Stemming과 lemmatization 두 종류의 작업이 있는데 표제어 추출이 어근추출보다 단어의 의미를 잘 보존한 원형을 찾아낼 수 있으므로 표제어 추출방식을 적용하였습니다.

2) Sequence to Sequence 모델

- loss가 nan이 나오는 문제는 해결하였습니다. epoch의 수를 충분히 줘야 제대로 된 결과와 성능평가가 가능하지만, 제 컴퓨터에서는 시간이 너무 오래 걸리는 일이기에 일단 다음 모델을 설계해보고 그 뒤에 학교 GPU서버를 활용하여 제대로 진행해보려고 합니다.

3) transformer 모델

- 디코더 부분에만 어텐션 메커니즘을 활용하였던 제 seq2seq모델과는 달리, RNN셀은 전혀 사용하지 않고 오로지 어텐션 메커니즘만 인코더-디코더 구조에 적용한 모델입니다. 일반적으로 LSTM셀을 사용하는 것에 비해 성능향상이 있다고 알려져 있지만, 이번에도 제 노트북 성능의 한계로 10번의 epoch까지만 진행하였습니다. 다음주동안 학교 GPU서버를 활용하여 제대로 진행해보려고 합니다.

4) 사전훈련모델

- transformer 구조를 토대로 사전훈련된 언어모델을 활용하여 번역&요약을 진행하는 방식입니다. 이번 주에는 이 사전훈련모델들을 어떻게 사용하는지에 대해 공부를 하였고, 다음 주에도 공부를 하면서 그와 병행하여 사전훈련모델을 활용하는 코드를 작성해 보려고 합니다.

일단 진행상황을 간단히 요약해서 진행상황 보고서로 제출하겠습니다. 무단 결석하여 죄송합니다 ㅠㅠ