Week1-3 과제

1. Paperswithcode에서 NLG extractive summarization task에 대해서 본인 블로그에 정리해보세요. 아래 3가지 항목에 대해서 정리하세요.

Extractive Text Summarization

문제정의

주어진 문서에서 문서의 요약을 가장 잘 표현하는 단어 혹은 문장의 하위 집합을 선택하는 문제

데이터 소개

CNN / Daily Mail

Source Document
(@entity0) wanted : film director , must be eager to shoot footage of golden lassos and invisible jets . <eos> @entity0 confirms that @entity5 is leaving the upcoming " @entity9 " movie (the hollywood reporter first broke the story) . <eos> @entity5 was announced as director of the movie in november . <eos> @entity0 obtained a statement from @entity13 that says , " given creative differences , @entity13 and @entity5 have decided not to move forward with plans to develop and direct ' @entity9 ' together . <eos> " (@entity0 and @entity13 are both owned by @entity16 . <eos>) the movie , starring @entity18 in the title role of the @entity21 princess , is still set for release on june 00 , 0000 . <eos> it 's the first theatrical movie centering around the most popular female superhero . <eos> @entity18 will appear beforehand in " @entity25 v. @entity26 : @entity27 , " due out march 00 , 0000 . <eos> in the meantime , @entity13 will need to find someone new for the director 's chair . <eos>
Ground truth Summary
@entity5 is no longer set to direct the first " @entity9 " theatrical movie <eos> @entity5 left the project over " creative differences " <eos> movie is currently set for 0000

CNN과 Daily Mail 웹사이트의 뉴스 기사를 보고 사람이 요약 라벨을 만든 데이터셋

- 훈련 데이터 : 286,817개
  - 훈련 데이터에서 Source Documnet는 평균 29.74개의 문장과 766개의 단어가 나타나고 Ground truth Summaray는 평균 53개의 단어와 3.72개의 문장으로 표현됨
- 검증 데이터 : 13,368개
- 테스트 데이터 : 11,487개

SOTA 모델 소개

HAHSum

- 선택 기반 추출 혹은 생성 기반 추상화 대신 공동 추출 및 구문 압축을 기반으로 단일 문서를 요약 : 문서 내 문장을 선택해 구성 요소 구문 분석을 기반으로 압축하고 이를 점수 매겨서 요약문을 생성

NeRoBERTa

- 기학습된 BERT 기반 인코더는 문서에서 문장 정보를 표현하도록 훈련되지 않아써 일관성있고 유익한 요약을 구성하지 못함 → 구문 및 담화 트리로 구성되는 RoBERTa에 대한 중첩 트리 기반 추출 요약 모델로 일관성 향상