- 본인 아이디 및 닉네임

jinsuc28(블로그 아이디) , 6팀_최진수

- 게시글 URL

https://jinsuc.tistory.com/8

- 게시글 캡처

0. 들어가며

NLP task 중 NLG 그리고 extractive summarization task에 대해서 알아보겠습니다.

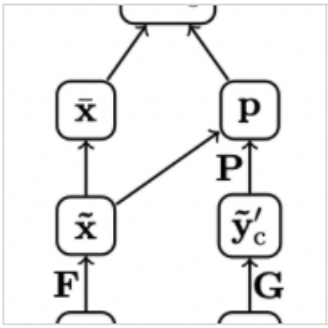
1. [NLG] extractive summarization task

Text Abbreviation의 영역으로 긴 문장에서 핵심만 뽑아 주는 것이 주요 과제입니다.
또한, 하위 카테고리로 **Summarization**(요약) 있으며 이 요약은 **Abstractive(생성 요약)**와 **Extractive(추출 요약)**로 더욱 세분화 됩니다.

둘의 차이 점은 **Abstractive(생성 요약)**은 **input 데이터에 없는 output(요약)을 생성**해냅니다.
하지만 **Extractive(추출 요약)**은 **input 데이터에 있는 output(요약)을 추출**해냅니다.

※ 참고자료

https://supkoon.tistory.com/40



[자연어처리][paper review] A N…

1.Background Sentence Summarization Sentence Summarization은 Input Text의 핵심 의미를 포착하여 …

supkoon.tistory.com

1-2 Summarization Extractive

-Task 설명

긴 문장을 요약해주는 것이 과제이며 문서에 없는 단어들로 문서를 요약해줍니다.

-데이터 세트

대표적으로 supervised summarization된 데이터로 CNN/Daily Mail이 있습니다.
이 데이터는 **CNN과 Daliy Mail** 뉴스 기사들을 크롤링한 데이터입니다.
밑에 그림처럼 **document**와 **summary가 쌍**으로 주어지며 t**rain, valid, test**각각 **286,817개, 13,368개, 11,487개 쌍**으로 주어집니다.

document 데이터는 평균적으로 29,74개 문장과 766개의 단어를 가지고 있습니다.
summary 데이터는 평균적으로 3.72개의 문장과 53개의 단어로 되어있습니다.

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

데이터 예시

위의 **문서(x)**에 대한 실제 사람이 요약한 **정답(y)**을 제공해줍니다.

-모델 평가 방법

summarization 모델 평가 방법은 **ROUGE-1**, **ROUGE-2**,**ROUGE-L** 총 3가지로 평가 됩니다.
자세한 설명은 영상 참고해주시면 좋을 것 같습니다.
(**ROUGE**: **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation)

예시)

✓ ROUGE-1

예시)

✓ ROUGE-1

[Reference Summary: R]
The capital of Korea, Seoul, is one of the biggest cities of the world.

[Model Summary: M1]
Seoul is the biggest city of the world. $\text{ROUGE} - 1 = \frac{7}{14} = 0.5$

[Model Summary: M2]
World is a biggest cities of the Seoul. $\text{ROUGE} - 1 = \frac{7}{14} = 0.5$

ROUGE-1

- **ROUGE**: **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation

✓ ROUGE-2

[Reference Summary: R]
The capital of Korea, Seoul, is one of the biggest cities of the world.

[Model Summary: M1]
Seoul is the biggest city of the world. $\text{ROUGE} - 2 = \frac{4}{13} = 0.3077$

[Model Summary: M2]
World is a biggest cities of the Seoul. $\text{ROUGE} - 2 = \frac{3}{13} = 0.2308$

ROUGE-2

- **ROUGE**: **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation

✓ ROUGE-L

[Reference Summary: R]
The capital of Korea, Seoul, is one of the biggest cities of the world.

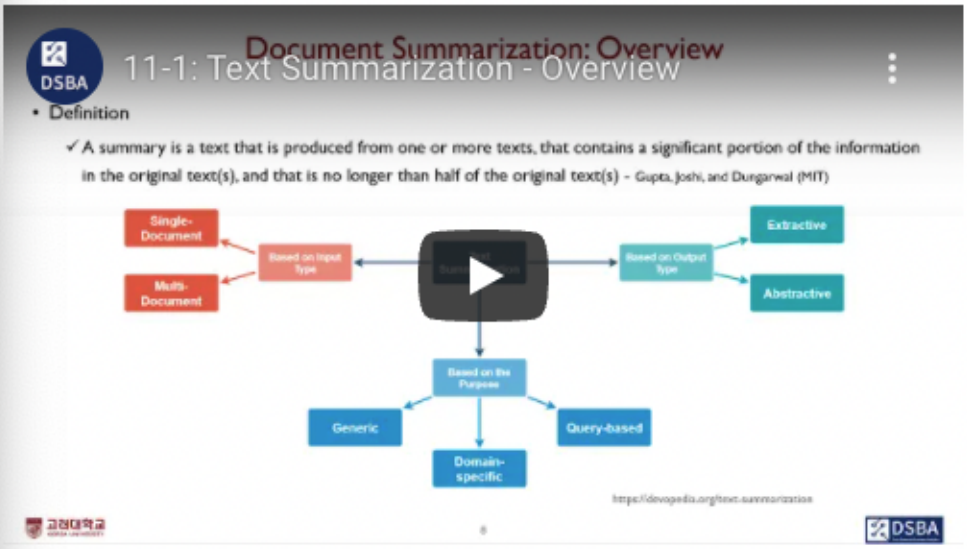
[Model Summary: M1]
Seoul is the biggest city of the world. $\text{ROUGE} - \text{L} = \frac{7}{14} = 0.5$

[Model Summary: M2]
World is a biggest cities of the Seoul. $\text{ROUGE} - \text{L} = \frac{5}{14} = 0.3571$

ROUGE-L

※출처

https://youtu.be/25TEdaQPqQY

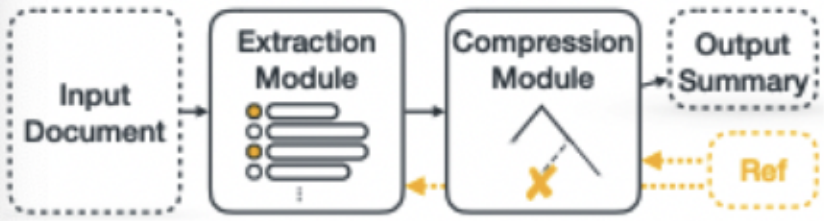


-sota모델 두가지

1. HANSum

문서에서 문장들을 선별하고 이를 압축하여서 압축한 문장을 output으로 내는 모델을 제안합니다.

2022년 2월 기준 가장 상위 sota로 등록되어있습니다.



https://paperswithcode.com/paper/neural-extractive-text-summarization-with

Papers with Code - Neural Extr…

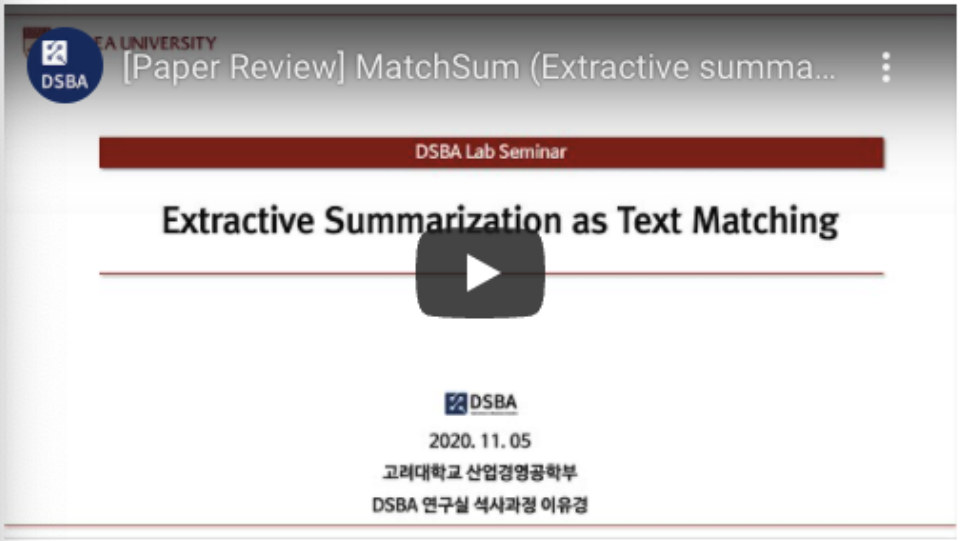
🏆 SOTA for Extractive Text Summarization on CNN / Daily Mail (ROUGE-1 metric)

paperswithcode.com

2. MatchSum

기존 extractive summarization 모델들은 상위 랭크된 문장간 관계로 output을 생성했습니다.
이때 문제점은 **문장**들간의 **의미적으로 연관성**을 고려하지 못했다는 것입니다. MatchSum이를 보완하기 위해
한 **모델과** 새로운 평가 지표인 **pearl-summary**와 **best-summary**를 제안합니다.

※ 참고자료
https://youtu.be/8E2Ia4Viu94



손지아

전체적으로 짧고 간결하게 이해할 수 있도록 작성해주신 것 같습니다. NeRoBERTa 같은 경우에는 같은 리더보드에 있어서 궁금했는데 잘 설명해주신 것 같습니다.

정태호

데이터 세트를 설명하실 때 구체적인 파일형식도 설명하시는 점이 좋았습니다. 또, 결론 부분에서 두 sota 모델을 활용하여 성능 개선할 수 있는 방법론 제안해주신 점도 인상깊은 것 같습니다.

현승환

저랑 동일한 모델을 sota로 소개해주셨는데 저보다 더 중요한 내용들을 잘 설명해 주신 것 같았습니다. 개인적으로 공부하는데 어려움이 있었는데 덕분에 HANSum과 MatchSum 모델을 좀더 잘 정리할 수 있었습니다.