BertSUM

BERT를 활용한 한국어 문서 추출요약 봇 (raqoon886)

https://velog.io/@raqoon886/KorBertSum-SummaryBot

- 배경
  - 생성요약 (abstractive summary): 토큰 간의 관계를 계산, 새로운 문장으로 요약
  - 추출요약 (extractive summary): 토큰 간의 관계를 계산, 문서에 포함된 문장으로 요약
    - 추출요약 ⇒ 많은 정보 누락 but 학습 시간, 컴퓨터 리소스가 적게 든다
  - ROUGE Metric: 모델의 요약 능력을 나타내는 자동화된 평가지표
    - "사람이 요약한 문장의 단어들이 얼마나 기계 요약에 많이 등장하는지" 측정

Text Summarization with Pretrained Encoder 세미나

https://drive.google.com/file/d/1wpEChNwPf8O2pRmG5IwX59kRuNjq9Ekj/view

- 자동 문서 요약(Automatic Text Summarization) ⇒ 크게 추출 요약 OR 생성 요약
  - 2015년 이전 대부분의 자동 문서 요약은 추출 요약 모델
  - 2015년 이후부터 신경망 모델을 사용하는 요약(Neural Text Summarization) 모델들이 발표
  - 특히 신경망 기반 언어 생성(Neural Language Generation)이 발전, 문서에 존재하지 않는 단어/표현을 사용하여 요약문을 생성하는 생성 요약 방식이 등장
- 다양한 모델 소개
  - 신경망을 이용한 대표적인 추출요약 모델 SummaRuNNer(2016)와 NeuSum(2018)
    - SummaRuNNer: RNN을 기반, 추출 요약을 sequence의 구성 요소 각각에 classification을 수행하는 sequence tagging 문제로 설정
    - NeuSum: Sentence scoring & Sentence selection을 동시에 하는 추출 요약 모델

(scoring: 각 문장에 중요도 점수를 매김, selection: 추출 요약에 포함시킬 문장을 선택)

- 생성요약 모델인 Pointer-Generator(2017)와 Bottom-up Summarization(2018)
  - Pointer-Generator: Attention seq2seq 방식으로 요약문 생성
    - Bottom-up Attention: Content Selection (단어를 masking) → Bottom-up Attention
- 2019년 발표된 논문 "Text Summarization with Pretrained Encoder"
  - pretrain이 진행된 BERT에 Transformer 구조를 활용해 추출 요약과 생성 요약을 하는 두 가지 모델을 제안한 paper
  - BERT 언어모델 pretraining → downstream task fine-tuning
  - 기존 BERT와 BERTSum의 input 차이: [Sep]로 구분된 문장의 시작마다 [CLS]를 넣음
  - BERTSumExt: Pretrained BERT 위에 Transformer model을 추가해 추출 요약 수행
    1. BERT를 거쳐 나온 output 중 [CLS] 토큰에 해당하는 vector만 선택
    2. 선택한 [CLS] 토큰들은 두 개의 layer로 이루어진 transformer encoder의 input이 됨
    3. 마지막 output을 이용해 binary classification
  - BERTSumAbs: Pretrained BERT 위에 Transformer decoder model을 추가, 생성 요약 수행
    1. 문장 시작을 뜻하는 <BOS>를 받아 요약문의 첫 단어를 예측
    2. 첫 단어를 input으로 다음 단어를 예측, 반복 ...
    3. ... 문장 종료를 뜻하는 <EOS>를 모델이 예측하면 문장 생성 종료
  - BERTSumExtAbs: 추출요약 → 생성요약 순서로 학습
    (Bottom-up Summarization이 Content selection → 생성요약 인 것에서 영감)
    1. 추출요약(BERTSumExt) 학습
    2. 동일한 Encoder를 사용해 생성 요약(BERTSumAbs) 학습

    생성요약만 사용한 BERTSumAbs보다 좋은 성능

텍스트 요약 분야의 주요 연구 주제, Must-read Papers, 이용 가능한 model 및 data 등을 추천 자료와 함께 정리한 저장소

https://github.com/uoneway/Text-Summarization-Repo

Understanding Abstractive Text SummarizatiZon from Scratch

https://pub.towardsai.net/understanding-abstractive-text-summarization-from-scratch-baaf83d446b3