(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0066828
(43) 공개일자 2023년05월16일

(51) 국제특허분류(Int. Cl.)
G06F 16/34 (2019.01) *G06F 40/205* (2020.01)
G06N 3/04 (2023.01) *G06N 3/08* (2023.01)
(52) CPC특허분류
G06F 16/345 (2019.01)
G06F 40/205 (2020.01)
(21) 출원번호 10-2021-0152256
(22) 출원일자 2021년11월08일
심사청구일자 없음

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
구명완
서울특시 양천구 목동동로 100, 목동아파트. 1312동 1304호
최예린
경기도 파주시 금바위로 47 가람마을8단지 동문굿모닝힐 811동 1201호
(뒷면에 계속)
(74) 대리인
이지연

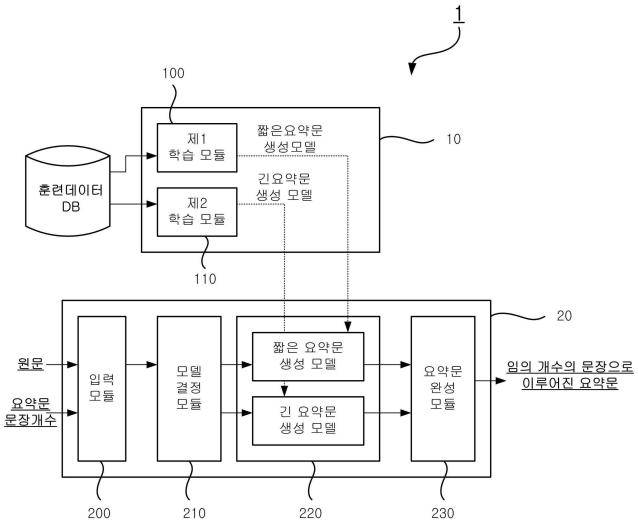
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **문서 자동 요약 시스템**

(57) 요 약

본 발명은 한국어 문서 자동 요약 시스템에 관한 것이다. 상기 문서 자동 요약 시스템은, 훈련 데이터를 이용하여 사전 학습시켜 짧은 요약문 생성 모델 및 긴 요약문 생성 모델을 완성하는 모델 학습 장치; 및 원문에 대하여 임의의 개수의 문장으로 이루어진 요약문을 생성하는 요약문 생성 장치;를 구비한다. 상기 요약문 생성 장치는, 원문 및 요약문 문장 개수를 입력받는 입력 모듈; 상기 입력된 요약문 문장 개수에 따라 짧은 요약문 생성 모델 및 긴 요약문 생성 모델 중 하나를 선택하는 모델 결정 모듈; 상기 모델 결정 모듈에 의해 선택된 모델을 이용하여 원문에 대한 요약문을 생성하는 요약문 생성 모듈; 및 요약문을 구성하는 문장 개수가 상기 입력된 요약문 문장 개수보다 크면 요약문의 앞에서부터 상기 입력된 요약문 문장 개수에 대응되는 문장들까지 추출하여 요약문을 완성하는 요약문 완성 모듈;을 구비한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/045 (2023.01)
G06N 3/08 (2023.01)
(72) 발명자
홍현의
서울특별시 마포구 백범로10길 20 가든힐레지던스 202호

이정필
서울특별시 은평구 역말로9길 17-44 수성렉스빌 501호

이 발명을 지원한 국가연구개발사업

과제고유번호	1711126331
과제번호	2021-0-00016-001
부처명	과학기술정보통신부
과제관리(전문)기관명	iitp
연구사업명	정보통신방송기술개발사업
연구과제명	엣지 컴퓨팅 기반 음성위협 인지 기술 개발
기 여 율	1/1
과제수행기관명	과학기술정보통신부
연구기간	2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

훈련 데이터를 이용하여 사전에 학습시켜 짧은 요약문 생성 모델 및 긴 요약문 생성 모델을 완성하는 모델 학습 장치; 및

원문에 대하여, 임의의 개수의 문장으로 이루어진 요약문을 생성하는 요약문 생성 장치;를 구비하고,

상기 요약문 생성 장치는,

외부로부터 요약할 문서인 원문 및 요약문 문장 개수를 입력받는 입력 모듈;

상기 입력된 요약문 문장 개수에 따라 짧은 요약문 생성 모델 및 긴 요약문 생성 모델 중 하나를 선택하는 모델 결정 모듈;

사전 학습된 상기 짧은 요약문 생성 모델과 상기 긴 요약문 생성 모델을 구비하고, 상기 모델 결정 모듈에 의해 선택된 모델을 이용하여 원문에 대한 요약문을 생성하는 요약문 생성 모듈; 및

상기 요약문 생성 모듈로부터 생성된 요약문을 제공받고, 상기 요약문을 구성하는 문장 개수가 상기 입력된 요약문 문장 개수보다 크면 요약문의 앞에서부터 상기 입력된 요약문 문장 개수에 대응되는 문장들까지 추출하여 요약문을 완성하는 요약문 완성 모듈;

를 구비하는 것을 특징으로 하는 문서 자동 요약 시스템.

청구항 2

제1항에 있어서, 상기 모델 결정 모듈은 모델 선택을 위한 사전 설정된 기준값을 구비하며,

상기 입력된 요약문 문장개수정보가 상기 기준값 이하이면 짧은 요약문 생성 모델을 선택하고, 상기 기준값을 초과하면 긴 요약문 생성 모델을 선택하는 것을 특징으로 하는 문서 자동 요약 시스템.

청구항 3

제1항에 있어서, 상기 짧은 요약문 생성 모델은 원문에 대하여 고정된 개수의 문장들로 이루어진 요약문을 생성하는 신경망 모델이며,

상기 긴 요약문 생성 모델은 원문에 대하여 일정하지 않은 개수의 문장들로 이루어진 요약문을 생성하는 신경망 모델인 것을 특징으로 하는 문서 자동 요약 시스템.

청구항 4

제1항에 있어서, 상기 모델 학습 장치는

훈련 데이터를 이용하여 제1 신경망 모델을 사전에 학습시켜 짧은 요약문 생성 모델을 완성하는 제1 학습 모듈을 구비하고,

상기 제1 신경망 모델은 인코더와 디코더를 구비하는 것을 특징으로 하는 문서 자동 요약 시스템.

청구항 5

제1항에 있어서, 상기 모델 학습 장치는

훈련 데이터를 이용하여 제2 신경망 모델을 사전에 학습시켜 긴 요약문 생성 모델을 완성하는 제2 학습 모듈을 구비하고,

상기 제2 신경망 모델은 양방향 인코더와 자동회귀 디코더를 구비하는 것을 특징으로 하는 문서 자동 요약 시스템.

청구항 6

(a) 훈련 데이터를 이용하여 사전에 학습시켜 짧은 요약문 생성 모델 및 긴 요약문 생성 모델을 완성하는 단계;

(b) 외부로부터 요약할 문서인 원문 및 요약문 문장 개수를 입력받는 단계;

(c) 상기 입력된 요약문 문장 개수에 따라 상기 짧은 요약문 생성 모델 및 상기 긴 요약문 생성 모델 중 하나를 선택하는 단계;

(d) 상기 선택된 모델을 이용하여 요약문을 생성하는 단계;

(e) 상기 생성된 요약문을 구성하는 문장 개수가 상기 입력된 요약문 문장 개수보다 크면, 요약문의 앞에서부터 요약문 문장 개수에 대응되는 문장들까지 추출하여 요약문을 완성하는 단계;

을 구비하는 것을 특징으로 하는 문서 자동 요약 방법.

청구항 7

제6항에 있어서, 상기 (c) 단계는 모델 선택을 위한 기준값을 설정하고,

상기 입력된 요약문 문장 개수 정보가 상기 기준값 이하이면 짧은 요약문 생성 모델을 선택하고, 상기 기준값을 초과하면 긴 요약문 생성 모델을 선택하는 것을 특징으로 하는 문서 자동 요약 방법.

청구항 8

제7항에 있어서, 상기 (c) 단계의 모델 선택을 위하여 설정된 기준값은,

상기 짧은 요약문 생성 모델을 사전 학습시킨 훈련 데이터의 고정된 정답 문장 개수와 동일하게 설정된 것을 특징으로 하는 문서 자동 요약 방법.

청구항 9

제6항에 있어서, 상기 짧은 요약문 생성 모델은 원문에 대하여 고정된 개수의 문장들로 이루어진 요약문을 생성하는 신경망 모델이며,

상기 긴 요약문 생성 모델은 원문에 대하여 일정하지 않은 개수의 문장들로 이루어진 요약문을 생성하는 신경망 모델인 것을 특징으로 하는 문서 자동 요약 방법.

청구항 10

제6항에 있어서, 상기 (a) 단계는,

(a1) 훈련 데이터를 이용하여 제1 신경망 모델을 사전에 학습시켜 짧은 요약문 생성 모델을 완성하는 단계; 및

(a2) 훈련 데이터를 이용하여 제2 신경망 모델을 사전에 학습시켜 긴 요약문 생성 모델을 완성하는 단계;

를 구비하고, 상기 제1 신경망 모델은 인코더와 디코더를 구비하며, 상기 제2 신경망 모델은 양방향 인코더와 자동회귀 디코더를 구비하는 것을 특징으로 하는 문서 자동 요약 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 문서 자동 요약 시스템에 관한 것으로서, 더욱 구체적으로는 사전 학습에 의해 생성된 짧은 요약문 생성 모델 및 긴 요약문 생성 모델을 이용하여, 원문에 대하여 사용자로부터 입력된 문장 개수를 갖는 요약문을 우수한 성능으로 자동 생성할 수 있도록 하는 문서 자동 요약 시스템에 관한 것이다.

배 경 기 술

[0002] 문서 자동 요약 기술은 텍스트로 작성된 문서(즉, 원문)에서 중요한 문장을 추출하여 요약하는 기술로서, 딥러닝 기술을 이용하여 원문의 텍스트를 이해하고 원문을 구성하는 문장들의 중요도를 분석하여 핵심 내용을 자동

으로 요약할 뿐만 아니라, 문서가 작성된 형식에 따라 적절한 요약 결과를 추출하여 요약문을 생성하게 된다.

[0003] 문서 자동 요약 시스템은 인코더와 디코더 구조로 이루어진 변환 모델(transformer medel)로 구성되어, 원문의 텍스트가 입력되면 인코더에 의해 원문의 문장의 문맥을 이해한 후, 디코더에 의해 요약된 문장으로 이루어진 요약문을 출력하게 된다.

[0004] 최근, 문서 자동 요약 기법으로는 순환 신경망 모델을 이용하는 방법이 제안되었는데, 이러한 딥러닝 기술을 이용함으로써 자동 요약문의 품질을 향상시킬 수 있게 된다. 한국등록특허 제 10-1944661호는 “순환신경망을 이용한 문서요약 시스템 및 방법”에 관한 것으로서, 순환신경망 모델을 이용하여 문서의 문장 단위로 인코딩 벡터를 재귀적으로 생성하고, 최종적으로 도출된 문서 임베딩 벡터를 이용하여 문서를 요약하는 것을 특징으로 하는 문서 자동 요약 시스템에 대한 기술이 개시되어 있다.

[0005] 하지만, 전술한 종래의 문서 자동 요약 시스템들은 요약하여야 할 텍스트가 하나의 시퀀스로 입력되기 때문에, 입력되는 원문을 구성하는 문장의 개수를 알 수 없는 상태에서 요약문을 생성하게 된다. 따라서, 종래의 문서 자동 요약 시스템들은 요약의 결과인 요약문에서 생성되는 문장의 개수를 조절하기가 어려운 문제점이 있다.

[0006] 또한, 기존의 문서 요약을 위한 신경망 모델은, 훈련 데이터의 정답 문장 개수에 따라 요약문의 문장 개수가 고정되는 모델이거나 훈련 데이터의 정답 문장 개수와는 관계없이 생성되는 요약문의 문장 개수가 일정하지 않은 모델이다. 따라서, 기존의 문서 자동 요약 시스템들은 사용자가 원하는 임의의 개수의 문장들로 이루어진 요약문을 자동으로 생성할 수 없는 단점이 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국등록특허 제 10-1944661호

(특허문헌 0002) 한국공개특허공보 제 10-2021-0125275호

발명의 내용

해결하려는 과제

[0008] 전술한 문제점을 해결하기 위한 본 발명의 목적은 요약할 문서에 대하여 사용자에 의해 정해진 개수의 문장들로 이루어진 요약문을 우수한 성능으로 자동으로 생성할 수 있도록 한 문서 자동 요약 시스템 및 방법을 제공하는 것이다.

과제의 해결 수단

[0009] 전술한 기술적 과제를 달성하기 위한 본 발명의 제1 특징에 따른 문서 자동 요약 시스템은, 훈련 데이터를 이용하여 제1 신경망 모델을 사전에 학습하여 짧은 요약문 생성 모델을 완성하고, 훈련 데이터를 이용하여 제2 신경망 모델을 사전에 학습하여 긴 요약문 생성 모델을 완성하는 모델 학습 장치; 및 원문에 대하여, 임의의 개수의 문장으로 이루어진 요약문을 생성하는 요약문 생성 장치;를 구비하고,

[0010] 상기 요약문 생성 장치는, 외부로부터 요약할 문서인 원문 및 요약문 문장 개수를 입력받는 입력 모듈; 상기 입력된 요약문 문장 개수에 따라 짧은 요약문 생성 모델 및 긴 요약문 생성 모델 중 하나를 선택하는 모델 결정 모듈; 짧은 요약문 생성 모델과 상기 긴 요약문 생성 모델을 구비하고, 상기 모델 결정 모듈에 의해 선택된 모델을 이용하여 원문에 대한 요약문을 생성하는 요약문 생성 모듈; 및 상기 요약문 생성 모듈로부터 생성된 요약문을 제공받고, 상기 요약문을 구성하는 문장 개수가 상기 입력된 요약문 문장 개수보다 크면 요약문의 앞에서부터 상기 입력된 요약문 문장 개수에 대응되는 문장들까지 추출하여 요약문을 완성하는 요약문 완성 모듈;을 구비한다.

[0011] 전술한 제1 특징에 따른 문서 자동 요약 시스템에 있어서, 상기 모델 결정 모듈은 모델 선택을 위한 사전 설정된 기준값을 구비하며, 상기 입력된 요약문 문장개수정보가 상기 기준값 이하이면 짧은 요약문 생성 모델을 선택하고, 상기 기준값을 초과하면 긴 요약문 생성 모델을 선택하는 것이 바람직하다.

[0012] 전술한 제1 특징에 따른 문서 자동 요약 시스템에 있어서, 상기 짧은 요약문 생성 모델은 원문에 대하여 고정된

개수의 문장들로 이루어진 요약문을 생성하는 신경망 모델이며, 상기 긴 요약문 생성 모델은 원문에 대하여 일정하지 않은 개수의 문장들로 이루어진 요약문을 생성하는 신경망 모델인 것이 바람직하다.

[0013] 전술한 제1 특징에 따른 문서 자동 요약 시스템에 있어서, 상기 모델 학습 장치는, 훈련 데이터를 이용하여 제1 신경망 모델을 사전에 학습시켜 짧은 요약문 생성 모델을 완성하는 제1 학습 모듈; 및 훈련 데이터를 이용하여 제2 신경망 모델을 사전에 학습시켜 긴 요약문 생성 모델을 완성하는 제2 학습 모듈;을 구비하는 것이 바람직하다.

[0014] 본 발명의 제2 특징에 따른 문서 자동 요약 방법은, (a) 훈련 데이터를 이용하여 제1 신경망 모델 및 제2 신경망 모델을 각각 사전에 학습시켜 짧은 요약문 생성 모델 및 긴 요약문 생성 모델을 완성하는 단계; (b) 외부로부터 요약할 문서인 원문 및 요약문 문장 개수를 입력받는 단계; (c) 상기 입력된 요약문 문장 개수에 따라 상기 짧은 요약문 생성 모델 및 상기 긴 요약문 생성 모델 중 하나를 선택하는 단계; (d) 상기 선택된 모델을 이용하여 요약문을 생성하는 단계; (e) 상기 생성된 요약문을 구성하는 문장 개수가 상기 입력된 요약문 문장개수보다 크면, 요약문의 앞에서부터 요약문 문장개수에 대응되는 문장들까지 추출하여 요약문을 완성하는 단계;를 구비한다.

[0015] 전술한 제2 특징에 따른 문서 자동 요약 방법에 있어서, 상기 (c) 단계는 모델 선택을 위한 기준값을 설정하고, 상기 입력된 요약문 문장 개수 정보가 상기 기준값 이하이면 짧은 요약문 생성 모델을 선택하고, 상기 기준값을 초과하면 긴 요약문 생성 모델을 선택하는 것이 바람직하며, 상기 기준값은 상기 짧은 요약문 생성 모델을 사전 학습시킨 훈련 데이터의 고정된 정답 문장 개수와 동일하게 설정된 것이 바람직하다.

[0016] 전술한 제1 특징에 따른 문서 자동 요약 시스템에 있어서, 상기 짧은 요약문 생성 모델은 원문에 대하여 고정된 개수의 문장들로 이루어진 요약문을 생성하는 신경망 모델이며, 상기 긴 요약문 생성 모델은 원문에 대하여 일정하지 않은 개수의 문장들로 이루어진 요약문을 생성하는 신경망 모델인 것이 바람직하다.

[0017] 전술한 제1 특징에 따른 문서 자동 요약 시스템에 있어서, 상기 (a) 단계는, (a1) 훈련 데이터를 이용하여 제1 신경망 모델을 사전에 학습시켜 짧은 요약문 생성 모델을 완성하는 단계; 및 (a2) 훈련 데이터를 이용하여 제2 신경망 모델을 사전에 학습시켜 긴 요약문 생성 모델을 완성하는 단계;를 구비하는 것이 바람직하다.

발명의 효과

[0018] 전술한 구성을 갖는 본 발명에 따른 문서 자동 요약 시스템은 신문기사, 칼럼 등과 같은 텍스트 형태의 각종 글에 대한 요약 서비스를 제공할 수 있게 된다.

[0019] 특히, 본 발명에 따른 문서 자동 요약 시스템은, 사용자 또는 외부로부터 요약문 문장개수정보를 입력받고, 이에 대응되는 문장들로 이루어진 요약문을 자동 생성하여 제공하게 된다. 이를 통해 다양한 층위의 요약을 자동 생성하여 제공하게 된다.

[0020] 본 발명에 따른 시스템에서는 짧은 요약문 생성 모델과 긴 요약문 생성 모델의 성능을 평가하기 위하여 평가 코퍼스를 자체 제작하였는데, AI Hub 신문기사 요약 데이터 중 원문의 문장 개수가 20개 이상인 100개의 문서를 무작위로 선정하였고, 이 문서들에서 중요 문장 10개를 선정하였다.

[0021] 먼저, 정량 평가를 위하여, Rouge 및 문장개수별 성능 평가를 하였다. Rouge는 자동 요약을 평가할 때 가장 많이 사용하는 지표로서, 아래의 수학식들로 표시되는 N-gram별 Precision, recall 점수의 F score를 사용한다.

수학식 1

[0022] $$precision = \frac{number\ of\ overlapping\ n\ -\ grams}{n\text{-}grams\ in\ reference}$$

수학식 2

[0023] $$recall = \frac{number\ of\ overlapping\ n\ -grams}{n\text{-}grams\ in\ hypothesis}$$

수학식 3

[0024] $$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

[0025] 여기서, reference는 평가 코퍼스에 있는 정답 요약문이고, hypothesis는 모델이 생성한 요약문이다. 결과는 rouge 1 기준으로, 짧은 요약문 생성 모델(t5)은 50.23 이며, 긴 요약문 생성 모델(kobart)은 43.49 가 나왔다.

[0026] 다음, 문장 개수별 성능 평가를 위하여, 수학식 3의 정확도(Acc)를 계산하였다.

수학식 4

[0027] $$Acc = \frac{1}{100} \sum_{i=1}^{100} \frac{|overlapped\ sentences|}{n}$$

[0028] 여기서, n은 모델이 생성한 문장의 개수이다. 즉, 모델이 생성한 문장 중 몇 개가 정답 문장과 일치하는지로 정확도를 계산하게 된다. 정답과의 일치 기준은 rouge score = 0.55로 선정하였다. 그 이유는 reference가 사람이 원문에서 추출한 요약문, hypothesis가 사람이 생성한 요약문일 때의 rouge score =0.55이기 때문이다. 도 3은 본 발명의 바람직한 실시예에 따른 문서 자동 요약 시스템에 있어서, 긴 요약문 생성 모델과 짧은 요약문 생성 모델에 대한 문장 개수별 성능 평가 결과를 도시한 그래프이다.

[0029] 다음, 정성 평가를 정량 평가와 마찬가지로 100개의 문서에 대하여 실시하였다. 정량 평가뿐만 아니라 정성 평가가 필요한 이유는 다음과 같다. 요약문의 품질을 평가한다는 것은 근본적으로 글의 품질을 평가하는 것이기 때문에, 정량 평가로는 한계가 있다. 사람이 직접 보고 이것이 잘 짜여진 글인지, 중요한 문장을 잘 채택한 요약문인지 판단하는 과정이 필요하다. 서비스의 대상이 사람이기 때문에 사람이 그 결과를 보고 판단한 기준도 평가 기준에 포함되어야 한다. 정성 평가를 거쳐야 실제 서비스를 제공하였을 때의 사용자의 만족도를 높일 수 있게 된다. 정성 평가의 평가 항목은 (1) 일관성, (2) 사실성, (3) 유창성 및 (4) 연관성이다. 일관성은 요약문 자체가 잘 짜여진 글인지, 하나의 글로서 구조를 갖췄는지에 대하여 평가한다. 사실성은 요약문에 원문과 다른 사실이 있는지 없는지에 대하여 평가한다. 유창성은 각 문장이 얼마나 유창한지에 대하여 평가한다. 연관성은 요약문의 내용이 원문에서 진짜 중요한 내용인지 여부를 평가한다.

[0030] 도 4는 본 발명의 바람직한 실시예에 따른 문서 자동 요약 시스템에 있어서, 긴 요약문 생성 모델과 짧은 요약문 생성 모델에 대하여 정성 평가의 항목별 평가 결과를 도시한 그래프이다. 도 4에서는 4 가지 정성 평가 항목에 대하여 100개의 문서에 대하여 각 항목당 4점 만점으로 평가한 결과를 도시한 그래프이다.

[0031] 본 발명에 따른 시스템에 있어서, 모델이 생성한 요약문보다 적은 문장 개수로 요약하여야 할 경우, 생성된 문장 순서대로 앞에서부터 추출한다. 이러한 문장 추출 방식은 실험 결과, 통계적인 방식으로 중요도를 계산하여 추출한 경우보다 더 좋은 성능의 요약문을 추출해 낼 수 있었다.

도면의 간단한 설명

[0032] 도 1은 본 발명의 바람직한 실시예에 따른 문서 자동 요약 시스템을 전체적으로 도시한 블록도이다.

도 2는 본 발명의 바람직한 실시예에 따른 문서 자동 요약 시스템에 따른 문서 자동 요약 방법을 순차적으로 도시한 흐름도이다.

도 3은 본 발명의 바람직한 실시예에 따른 문서 자동 요약 시스템에 있어서, 긴 요약문 생성 모델과 짧은 요약문 생성 모델에 대한 문장 개수별 성능 평가 결과를 도시한 그래프이다.

도 4는 본 발명의 바람직한 실시예에 따른 문서 자동 요약 시스템에 있어서, 긴 요약문 생성 모델과 짧은 요약문 생성 모델에 대하여 정성 평가의 항목별 평가 결과를 도시한 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0033] 이하 첨부된 도면을 참조하여 본 발명의 바람직한 실시예에 따른 문서 자동 요약 시스템 및 그 방법에 대하여 구체적으로 설명한다. 본 발명에 따른 문서 자동 요약 시스템은 컴퓨터 또는 서버에 설치되어 실행되는 프로그램에 의해 구현될 수 있다.

[0034] 본 발명에 따른 문서 자동 요약 시스템은 사전 학습에 의해 모델링된 짧은 요약문 생성 모델 및 긴 요약문 생성 모델을 구비하고, 사용자 또는 외부로부터 요약문을 구성할 문장 개수 정보와 요약하고자 할 문서(이하, '원문'이라 한다)를 입력받고, 사용자로부터 입력된 문장 개수 정보에 따라 짧은 요약문 생성 모델과 긴 요약문 생성 모델 중 하나를 선택하여 실행하여 최종적으로 요약문을 생성하여 출력하게 된다.

[0035] 도 1은 본 발명의 바람직한 실시예에 따른 문서 자동 요약 시스템을 전체적으로 도시한 블록도이다. 도 1을 참조하면, 본 발명에 따른 문서 자동 요약 시스템(1)은 훈련 데이터를 이용하여 사전에 학습하여 짧은 요약문 생성 모델과 긴 요약문 생성 모델을 완성하는 모델 학습 장치(10) 및 상기 짧은 요약문 생성 모델과 긴 요약문 생성 모델을 이용하여 원문에 대한 요약문을 생성하여 출력하는 요약문 생성 장치(20)를 구비한다. 상기 모델 학습 장치(10)는 짧은 요약문 생성 모델을 완성하는 제1 학습 모듈(100)과 긴 요약문 생성 모델을 완성하는 제2 학습 모듈(110)을 구비한다. 상기 요약문 생성 장치(20)는 입력 모듈(200), 모델 결정 모듈(210), 요약문 생성 모듈(220) 및 요약문 완성 모듈(230)을 구비하여, 인공 지능을 기반으로 하여 원문에 대하여 사용자에 의해 결정된 문장 개수를 갖는 요약문을 생성하여 출력하는 것을 특징으로 한다.

[0036] 상기 모델 학습 장치(10)는 사전 설정된 훈련 데이터를 이용하여 제1 및 제2 신경망 모델들을 각각 사전에 학습하여 짧은 요약문 생성 모델 및 긴 요약문 생성 모델을 완성하여 출력한다. 이하, 모델 학습 장치의 각 구성 요소에 대하여 구체적으로 설명한다.

[0037] 상기 제1 학습 모듈(100)은 훈련 데이터를 이용하여 제1 신경망 모델을 사전에 학습시켜 짧은 요약문 생성 모델을 완성하는 모듈이다. 따라서, 상기 짧은 요약문 생성 모델은 요약문의 문장 개수가 훈련 데이터의 정답 데이터의 문장 개수를 따라가도록 구성된 신경망 모델로서, 원문에 대하여 사전에 정해진 고정된 개수의 문장들로 이루어진 요약문을 생성하도록 학습된다. 짧은 요약문 생성 모델을 이루는 제1 신경망 모델은 인코더-디코더를 구비하는 transformer 모델로 이루어지며, 이 모델의 특징은 다른 transformer 모델보다 많은 파라미터를 사용하여 모델의 크기를 키운 것이다. 인코더로 들어온 원문은 그 sequence에서 가장 중요한 값을 찾아내는 attention module을 거치고, feed forward module을 거쳐 인코더가 문맥을 이해하는 과정을 수행한다. 그리고 디코더에서 인코더에서 이해한 내용을 바탕으로 인코더와 같이 attention module과 feed forward module을 거쳐 어떤 토큰을 생성할지 결정하게 된다.

[0038] 상기 제2 학습 모듈(110)은 훈련 데이터를 이용하여 제2 신경망 모델을 사전에 학습시켜 긴 요약문 생성 모델을 생성하는 모듈이다. 따라서, 상기 긴 요약문 생성 모델은 요약문의 문장 개수가 일정하지 않은 신경망 모델로서, 원문에 대하여 정해지지 않은 개수의 문장들로 이루어진 요약문을 생성하도록 학습된다. 긴 요약문 생성 모델을 이루는 제2 신경망 모델은 양방향 인코더(Bidirectional encoder)와 자동회귀 디코더(Autoregressive decode)의 구조로 이루어질 수 있다. 상기 양방향 인코더가 원문을 받아 내용을 이해하고, 자동 회귀 디코더가 인코더가 이해한 내용을 바탕으로 요약문을 생성한다. 자동회귀 디코더는 학습 시에는 디코더의 결과값을 디코더가 생성한 값이 아닌 실제 정답으로 학습하는 것을 특징으로 한다. 양방향 인코더와 자동회귀 디코더는 가장 중요한 내용을 가려내는 attention module을 거친 후, Feed Forward module을 거쳐 최종 토큰을 선정한다. 이때 활성화 함수는 gelu-function을 사용한다.

[0039] 현재 사용 가능한 훈련 데이터로는 한국지능정보사회진흥원이 운영하는 AI 통합 플랫폼인 AI Hub에서 제공하는 문서 요약 데이터와 국립국어원에서 제공하는 모두의 말뭉치의 문서 요약 데이터가 있으며, AI Hub의 문서 요약 데이터는 추출 요약 데이터의 정답이 3문장으로 구성되며, 모두의 말뭉치의 문서 요약 데이터는 생성 요약의 정답이 3문장으로 구성된다. 따라서, 본 발명에 따른 일실시예에서는, 짧은 요약문 생성 모델은 3개의 정답 문장 데이터를 갖는 훈련 데이터로 학습하게 되고, 3개의 문장을 이루어진 요약문을 생성하게 된다. 이러한 짧은 요약문 생성 모델에 의해 생성되는 문장은 한정적이기는 하지만 생성되는 요약문의 질이 좋다는 장점을 가진다. 실험 결과, 후술되는 긴 요약문 생성 모델이 문장을 추출한 결과 보다 짧은 요약문 생성 모델의 요약문이 더 좋은 성능을 보여주었다.

[0040] 한편, 4개 이상의 문장으로 요약할 경우 생성되는 문장의 개수가 다양한 긴 요약문 생성 모델을 사용한다. 긴 요약문 생성 모델은 문장의 개수와는 상관없이 중요한 내용을 담아낼 수 있으며, 같은 원문에 대해 전술한 짧은 요약문 생성 모델보다 대체로 많은 문장들로 구성된 요약문을 생성하게 된다. 이와 같이, 긴 요약문 생성 모델은 문장의 개수와 무관하게 요약문에 포괄적인 내용 및 중요 내용을 담기 위해 사용되므로, 요약문에 생성되는

문장의 개수가 유연하다.

[0041] 긴 요약문 생성 모델과 짧은 요약문 생성 모델은 신문 기사 요약 데이터들을 훈련 데이터로 하여 학습하였으며, 훈련 데이터는 AI Hub에 공개되어 있는 문서 요약 데이터 중 신문기사 데이터와 국립국어원의 모두의 말뭉치의 문서 요약 데이터를 사용하였다. 도메인으로 신문 기사를 선택한 이유는 일반인 사용자가 일상 생활에서 가장 많이 접하는 글이 신문기사이기 때문이다. 논문이나 판결문 등의 도메인도 존재하지만 이들은 특정 집단으로 사용자가 한정된다. 따라서, 본 발명에 따른 시스템은 보다 많은 사용자에게 공급할 수 있는 시스템을 만들기 위하여, 신문 기사를 훈련 데이터와 평가 데이터의 도메인으로 채택하였다.

[0042] 여러 가지 실험 결과, 짧은 요약문 생성 모델은 t5를 사용하였으며, AI Hub 추출 요약 데이터 25만개, 국립국어원 생성 요약 데이터 약 4400개로 훈련하는 것이 가장 좋은 성능을 내었다. 긴 요약문 생성 모델은 kobart 모델을 사용하였으며 각각 8만개 및 약 4400개로 훈련하였을 때에 가장 좋은 성능을 얻을 수 있었다.

[0043] AI Hub에 의해 제공되는 문서 데이터의 경우, 훈련 데이터로 요약 데이터(abstractive data)가 아닌 추출 데이터(extractive data)를 사용하였다. 이는 생성 결과를 관찰하였을 때 요약 데이터로 훈련시켰을 때의 결과는 추출 데이터로 훈련시켰을 때의 내용을 하나의 문장으로 만든 것에 불과하였다. 문장의 개수는 요약 데이터의 경우 하나의 문장으로만 생성되고, 추출 데이터의 경우 3 문장 이상으로 생성된다. 이 두 경우의 내용은 유사하고 문장 개수만이 차이가 있었기 때문에, 본 발명에서는 문장 개수의 다양화가 용이한 추출 데이터를 사용하였다.

[0044] 한편, 상기 요약문 생성 장치(20)는 사용자로부터 요약할 문서(즉, 원문) 및 요약문 문장 개수를 입력받고, 요약문 문장 개수에 따라 전술한 짧은 요약문 생성 모델과 긴 요약문 생성 모델 중 하나를 선택하여 요약문을 생성한 후, 요약문 문장 개수에 맞게 요약문의 문장을 추출하여 요약문을 완성하여 출력한다. 이하, 요약문 생성 장치의 각 구성 요소들에 대하여 구체적으로 설명한다.

[0045] 상기 입력 모듈(200)은 사용자 또는 외부로부터 원문 및 요약문 문장 개수를 입력받는다.

[0046] 상기 모델 결정 모듈(210)은 상기 입력 모듈을 통해 입력된 요약문 문장 개수에 따라 짧은 요약문 생성 모델 및 긴 요약문 생성 모델 중 하나를 결정하여 선택하고, 선택된 모델로 원문을 제공한다. 여기서, 상기 모델 결정 모듈은 요약문 문장 개수가 사전 설정된 기준값(자연수 N임)이하이면 짧은 요약문 생성 모델을 선택하고, 요약문 문장 개수가 상기 기준값을 초과하면 긴 요약문 생성 모델을 선택하게 된다. 상기 기준값은 짧은 요약문 생성 모델의 학습 과정에서 사용된 훈련 데이터의 고정된 정답 문장 개수와 동일하게 설정되는 것이 바람직하다. 따라서, 본 발명에서는 상기 기준값을 '3'으로 설정하였다.

[0047] 상기 요약문 생성 장치(220)은 상기 모델 학습 장치(10)에 의해 학습된 짧은 요약문 생성 모델 및 긴 요약문 생성 모델을 구비하고, 상기 모델 결정 모듈에 의해 선택된 모델로 원문이 제공되고, 선택된 모델이 요약문을 생성하여 상기 요약문 완성 모듈(230)로 출력한다. 이 경우, 짧은 요약문 생성 모델은 고정된 개수의 문장들로 이루어진 요약문을 생성하여 출력하게 되고, 긴 요약문 생성 모델은 일정하지 않는 개수의 문장들로 이루어진 요약문을 생성하여 출력하게 된다.

[0048] 상기 요약문 완성 모듈(230)은 상기 요약문 생성 모듈에 의해 생성된 요약문을 제공받고, 상기 생성된 요약문의 문장 개수를 측정하고, 측정된 요약문 문장 개수와 상기 입력 모듈에 의해 입력된 요약문 문장 개수를 비교한다. 비교 결과, 측정된 요약문 문장 개수가 입력된 요약문 문장 개수를 초과하면, 요약문의 앞에서부터 상기 입력된 요약문 문장 개수만큼 문장들을 획득하고 상기 입력된 요약문 문장 개수를 초과하는 문장들은 버림으로써, 요약문의 앞에서부터 문장들을 추출하여 요약문을 완성함으로써, 입력된 요약문 문장 개수의 문장들로 이루어진 요약문을 최종 완성하여 출력한다.

[0049] 전술한 본 발명에 따른 문서 자동 요약 시스템에 따른 문서 자동 요약 방법을 순차적으로 설명한다. 도 2는 본 발명의 바람직한 실시예에 따른 문서 자동 요약 시스템에 따른 문서 자동 요약 방법을 순차적으로 도시한 흐름도이다.

[0050] 도 2를 참조하면, 본 발명에 따른 문서 자동 요약 방법은, 크게 사전 학습 단계와 요약문 생성 단계로 구성된다. 먼저, 사전 학습 단계에서, 훈련 데이터를 이용하여 신경망 모델들을 사전에 학습하여 짧은 요약문 생성 모델 및 긴 요약문 생성 모델을 완성한다(단계 200 및 단계 202). 제1 학습 단계에서는 훈련 데이터를 이용하여 사전에 학습하여 짧은 요약문 생성 모델을 완성하며, 제2 학습 단계에서는 훈련 데이터를 이용하여 사전에 학습하여 긴 요약문 생성 모델을 완성한다. 구체적인 학습 과정은 문서 자동 요약 시스템의 모델 학습 장치에서 설명한 바와 동일하므로, 중복되는 설명은 생략한다.

[0051] 여기서, 상기 짧은 요약문 생성 모델은 훈련 데이터를 이용하여 제1 신경망 모델을 사전 학습시켜 완성된 신경망 모델로서, 원문에 대하여 고정된 개수의 문장들로 이루어진 요약문을 생성하도록 학습된다. 상기 긴 요약문 생성 모델은 훈련 데이터를 이용하여 제2 신경망 모델을 사전 학습시켜 완성된 신경망 모델로서, 원문에 대하여 일정하지 않은 개수의 문장들로 이루어진 요약문을 생성하도록 학습된다.

[0052] 다음, 요약문 생성 단계는 아래와 같이 진행된다.

[0053] 먼저, 사용자 또는 외부로부터 요약할 문서인 원문 및 요약문 문장 개수 정보를 입력받는다(단계 210).

[0054] 다음, 상기 입력된 요약문 문장 개수 정보에 따라 상기 짧은 요약문 생성 모델 및 상기 긴 요약문 생성 모델 중 하나를 선택하고, 상기 선택된 모델로 상기 원문을 제공한다(단계 220). 여기서, 모델 선택을 위한 사전 설정된 기준값을 저장하고, 상기 입력된 요약문 문장 개수 정보가 상기 기준값 이하이면 짧은 요약문 생성 모델을 선택하고, 상기 기준값을 초과하면 긴 요약문 생성 모델을 선택하도록 구성할 수 있다. 여기서, 상기 기준값은 짧은 요약문 생성 모델에 대한 학습 단계에서 사용된 훈련 데이터를 구성하는 고정된 정답 문장 개수와 동일하도록 설정하는 것이 바람직하다.

[0055] 다음, 상기 선택된 모델을 이용하여 요약문을 생성한다(단계 230).

[0056] 다음, 상기 선택된 모델에 의해 생성된 요약문을 구성하는 문장 개수가 상기 입력된 요약문 문장 개수 정보보다 크면, 요약문의 앞에서부터 요약문 문장개수정보에 대응되는 문장들까지 추출하여 요약문을 완성한다(단계 240).

[0057] 이상에서 본 발명에 대하여 그 바람직한 실시예를 중심으로 설명하였으나, 이는 단지 예시일 뿐 본 발명을 한정하는 것이 아니며, 본 발명이 속하는 분야의 통상의 지식을 가진 자라면 본 발명의 본질적인 특성을 벗어나지 않는 범위에서 이상에 예시되지 않은 여러 가지의 변형과 응용이 가능함을 알 수 있을 것이다. 그리고, 이러한 변형과 응용에 관계된 차이점들은 첨부된 청구 범위에서 규정하는 본 발명의 범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0058] 1 : 문서 자동 요약 시스템

10 : 모델 학습 장치

100 : 제1 학습 모듈

110 : 제2 학습 모듈

20 : 요약문 생성 장치

200 : 입력 모듈

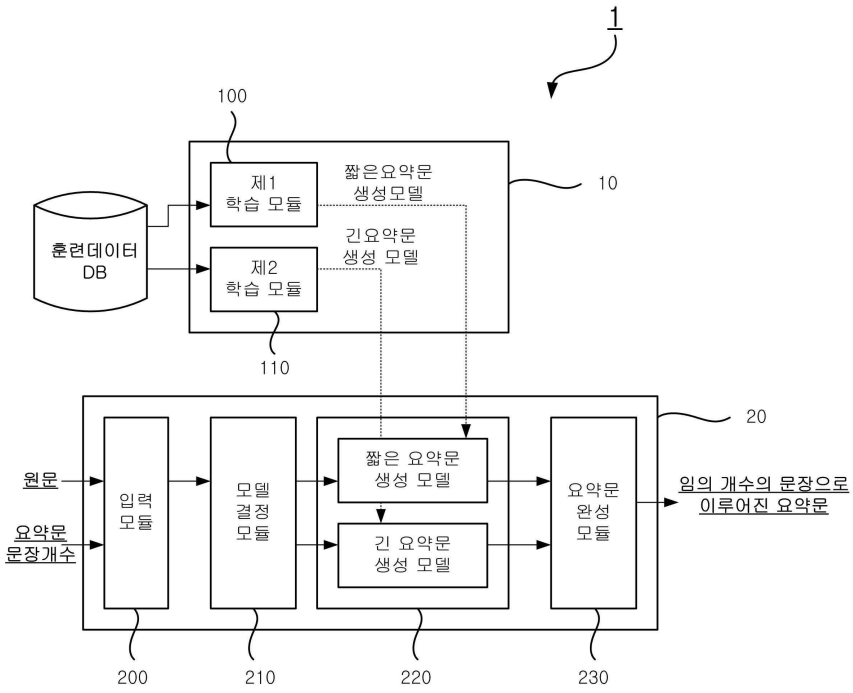
210 : 모델 결정 모듈

220 : 요약문 생성 모듈

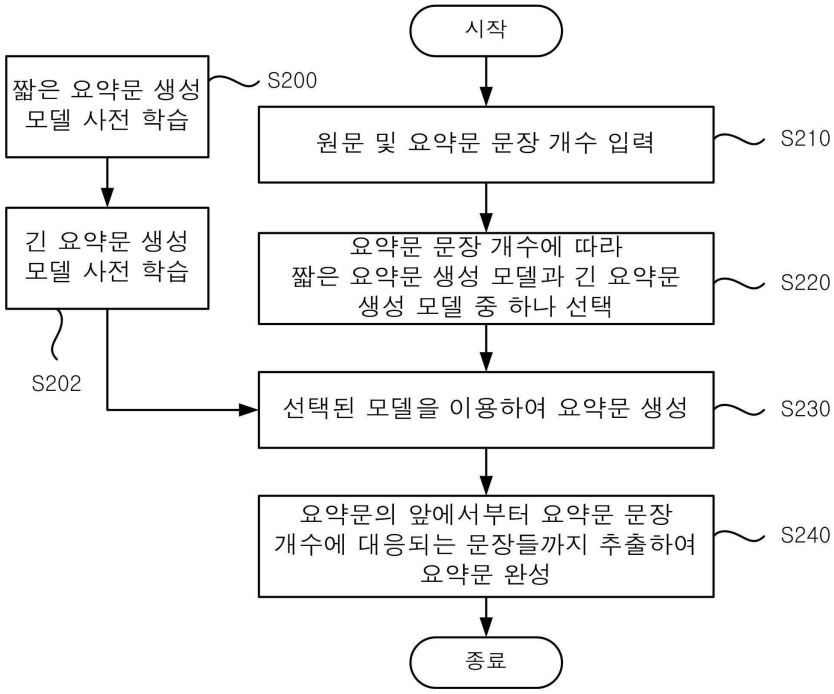
230 : 요약문 완성 모듈

도면

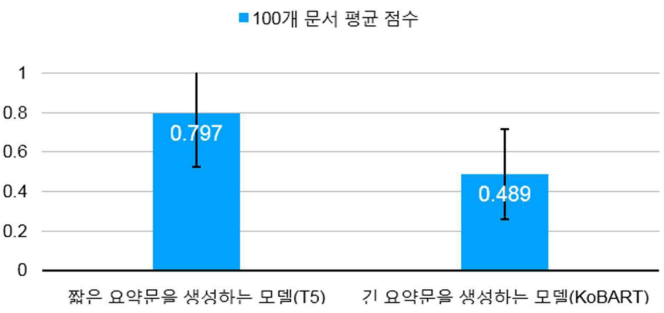
도면1


1
100
제1
학습 모듈
짧은요약문
생성모델
10
훈련데이터
DB
제2
학습 모듈
긴요약문
생성 모델
110
20
원문
요약문
문장개수
입력
모듈
모델
결정
모듈
짧은 요약문
생성 모델
긴 요약문
생성 모델
요약문
완성
모듈
임의 개수의 문장으로
이루어진 요약문
200
210
220
230

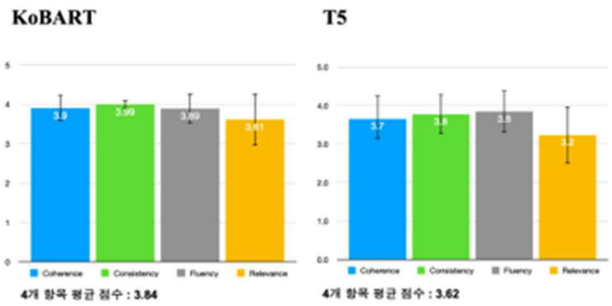
도면2


시작
짧은 요약문 생성
모델 사전 학습
S200
원문 및 요약문 문장 개수 입력
S210
긴 요약문 생성
모델 사전 학습
요약문 문장 개수에 따라
짧은 요약문 생성 모델과 긴 요약문
생성 모델 중 하나 선택
S220
S202
선택된 모델을 이용하여 요약문 생성
S230
요약문의 앞에서부터 요약문 문장
개수에 대응되는 문장들까지 추출하여
요약문 완성
S240
종료

도면3


100개 문서 평균 점수
1
0.8
0.6
0.4
0.2
0
0.797
0.489
짧은 요약문을 생성하는 모델(T5)
긴 요약문을 생성하는 모델(KoBART)

도면4


KoBART
T5
4개 항목 평균 점수 : 3.84
4개 항목 평균 점수 : 3.62