KIPS Trans. Softw. and Data Eng.
Vol.10, No.11 pp.535~540
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.11.535

Multi-Document Summarization Method of Reviews Using Word Embedding Clustering

Pil Won Lee† · Yun Young Hwang† · Jong Seok Choi†† · Young Tae Shin†††

ABSTRACT

Multi-document refers to a document consisting of various topics, not a single topic, and a typical example is online reviews. There have been several attempts to summarize online reviews because of their vast amounts of information. However, collective summarization of reviews through existing summary models creates a problem of losing the various topics that make up the reviews. Therefore, in this paper, we present method to summarize the review with minimal loss of the topic. The proposed method classify reviews through processes such as preprocessing, importance evaluation, embedding substitution using BERT, and embedding clustering. Furthermore, the classified sentences generate the final summary using the trained Transformer summary model. The performance evaluation of the proposed model was compared by evaluating the existing summary model, seq2seq model, and the cosine similarity with the ROUGE score, and performed a high performance summary compared to the existing summary model.

Keywords : Muti-document, Text Summarization, Transformer, Word Embedding

워드 임베딩 클러스터링을 활용한 리뷰 다중문서 요약기법

이 필 원† · 황 윤 영† · 최 종 석†† · 신 용 태†††

요 약

다중문서는 하나의 주제가 아닌 다양한 주제로 구성된 문서를 의미하며 대표적인 예로 온라인 리뷰가 있다. 온라인 리뷰는 정보량이 방대하기 때문에 요약하기 위한 여러 시도가 있었다. 그러나 기존의 요약모델을 통해 리뷰를 일괄적으로 요약할 경우 리뷰를 구성하고 있는 다양한 주제가 소실되는 문제가 발생한다. 따라서 본 논문에서는 주제의 손실을 최소화하며 리뷰를 요약하기 위한 기법을 제시한다. 제안하는 기법은 전처리, 중요도 평가, BERT를 활용한 임베딩 치환, 임베딩 클러스터링과 같은 과정을 통해 리뷰를 분류한다. 그리고 분류된 문장은 학습된 Transformer 요약모델을 통해 최종 요약을 생성한다. 제안하는 모델의 성능 평가는 기존의 요약모델인 seq2seq 모델과 ROUGE 스코어와 코사인 유사도를 평가하여 비교하였으며 기존의 요약모델과 비교하여 뛰어난 성능의 요약을 수행하였다.

키워드 : 다중문서, 텍스트 요약, 트랜스포머, 워드 임베딩

1. 서 론

다중문서는 다수의 문서가 하나의 주제가 아닌 다양한 주제로 이루어진 문서를 의미한다. 대표적으로 온라인 쇼핑몰의 리뷰는 하나의 상품에 대해 소비자의 다양한 의견을 나타내고 구매 결정에 주요한 영향을 미치는 다중문서이다[1]. 그러나 리뷰는 하나의 상품에 대해 많게는 약 10,000건이 넘는 의견들로 이루어져 있어 소비자가 모든 리뷰를 읽어보고 정보를 습득하는 것은 효율적이지 않다. 이러한 문제를 해결하기 위해 문서의 정보를 효율적으로 압축하여 정보를 나타내려는 문서요약에 관한 여러 연구가 진행되었다[2].

문서 요약은 일반적으로 추출(Extractive)요약과 추상(Abstractive)요약으로 나누어진다[3]. 추출요약은 문서 내부의 중요한 문장을 알고리즘을 통해 중요도를 평가하여 해당 문장을 추출하는 방법이다. 추상요약은 문서의 전체적인 글의 내용을 고려하여 학습한 내용을 바탕으로 새로운 문장을 생성하는 요약하는 방법이다. 다중문서의 경우는 추출요약과 생성요약 하나만 적용하여 일괄적으로 요약을 수행할 경우 다중문서의 다양한 주제가 소실되어 만족스러운 요약결과가

※ 이 논문은 2021년도 과학기술정보통신부 및 정보통신기획평가원의 대학ICT 연구센터지원사업 지원에 의하여 수행된 것임(IITP-2020-2020-0-01602, 지능형 사이버 위협 대응 기술 개발 및 인력양성).
※ 이 논문은 2021년 한국정보처리학회 춘계학술발표대회에서 “워드 임베딩 유사도 클러스터링을 통한 다중 문장 생성 요약기법”의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 숭실대학교 컴퓨터학과 석사과정
†† 준 회 원 : 숭실대학교 스파르탄SW교육원 교수
††† 종신회원 : 숭실대학교 컴퓨터학부 교수
Manuscript Received : June 28, 2021
First Revision : July 14, 2021
Accepted : July 14, 2021
* Corresponding Author : Young-Tae Shin(shin@ssu.ac.kr)

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/ licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

도출되기 어렵다[4].

따라서 본 논문에서는 다중문서의 형태소 분석을 통해 명사를 추출한 뒤 미리 학습된 BERT 모델을 통해 워드 임베딩을 도출한다. 도출된 워드 임베딩의 클러스터링을 수행하여 다중문서가 가지고 있는 주제를 추출하고 각각의 주제에 해당하는 단어 세트를 구성한다. 각각의 주제의 단어 세트를 통해 하나의 문장이 어떤 주제인지 판별하여 분류한다. 그리고 각 주제에 해당하는 문장을 감정분석을 통해 긍정, 부정을 분류하고 문장의 중요도를 평가하는 TextRank 알고리즘을 통해 높은 중요도의 문장을 추출한다. 최종적으로 각각의 주제마다 긍정, 부정으로 분류된 다수의 문장이 추출되며 이를 하나의 문장으로 요약하기 위해 Transformer 모델을 활용하여 생성요약을 수행한다.

본 논문의 구성은 다음과 같다. 2장 관련 연구는 기존의 문서 요약 연구와 다중문서 요약 연구 대해 서술한다. 3장에서는 본 논문에서 제안하는 워드 임베딩의 유사도 클러스터링을 통한 다중문서 요약 생성 기법에 관해 설명한다. 4장에서는 제안하는 기법의 요약 성능을 분석하고, 마지막 5장에서는 결론을 서술한다.

2. 관련 연구

2.1 문서 요약

문서 요약은 문서 내에서 중요한 문장을 포함하여 요약을 생성하는 작업이다[5]. 일반적으로 문서요약은 크게 추출요약과 추상 요약으로 나누어진다. 추출요약은 문서내의 문장의 중요도를 평가하는 알고리즘을 통해 높은 중요도를 갖는 원래의 문장을 수정 없이 추출한다[6]. 최근 다양한 분야에서 추출요약을 활용하여 문서를 요약하는 연구가 진행되었다. 그러나 온라인 리뷰와 같은 하나의 문서에 여러 주제를 가지고 있는 다중문서에 대해 오직 추출요약만 활용하여 문석요약을 수행하면 일부 주제가 소실되는 문제가 존재한다. 추상요약은 문서의 정보를 참고하여 요약된 새로운 문장을 생성하는 작업이다[7]. 추상요약은 인공신경망의 발전에 따라 기계번역 분야에서 높은 성과를 보이고 있다. 신경망 기반 문서요약 모델은 기본적으로 인코더-디코더 구조를 바탕으로 구성된다. 인코더에 입력문장을 입력하면 문장의 단어가 순차적으로 학습된 인공신경망에 입력되어 결과적으로 하나의 벡터값 Context Vector가 도출된다. Context Vector는 디코더에 입력되어 단어를 순차적으로 하나씩 단어를 도출하여 문장을 완성한다. 요약모델 중 가장 대표적인 Sequence-to-Sequence 모델은 LSTM 또는 GRU로 구성된 인코더-디코더 구조 모델이다. 이후 RNN 모델의 단점인 기울기 소실문제로 입력문장이 길어지면 요약 성능이 떨어지는 것을 보완하기 위해 주위 집중 매커니즘인 Attention[8]이 모델에 결합되어 성능을 높였다. 최근에는 RNN을 활용하지 않고 Attention만 활용하여 인코더-디코더를 구성하는 트랜스포머(Transformer) 방식이 연구되고 있으며 성능이 Sequence-to-Sequence보다 우수한 것으로 나타났다[9]. 따라서 본 논문에서는 추출요약과 추상요약 기법의 이점을 모두 활용하기 위해 두 가지 요약기법을 모두 사용하여 요약기법을 설계한다.

2.2 다중문서 요약

다중문서는 하나의 문서에 한가지의 주제가 아니라 여러 가지의 주제를 내포하는 문서를 의미한다. 대표적으로 온라인 리뷰는 다수의 짧은 문장에 소비자의 반응을 표현하는 문서이다. 또한 온라인 리뷰는 매출에 주요한 영향을 미치는 eWOM(Electronic Word-of-Mouse)의 종류 중 하나다. 그러나 리뷰는 하나의 제품에 대해 많으면 수만 개의 리뷰가 존재하기 때문에 소비자가 모든 정보를 획득하는 것이 어렵다. 따라서 리뷰의 정보 중에 소비자에게 도움이 되는 정보를 요약하려는 연구가 진행되어왔다. Tan et al.(2017)[10]은 주제 고정 검토 요약(TARS)라는 2단계의 요약 방법을 제시한다. 첫 번째는 주제 측면 감성 모델(TASM)을 활용하여 리뷰 세트에서 세분화된 감성 주제를 식별한다. 두 번째에서는 TASM에서 식별된 주제의 중요도를 평가하여 리뷰의 순위를 평가하여 요약하는 기법을 TripAdvisor 리뷰를 통해 실험하였으며 기존의 요약보다 뛰어난 결과를 보였다.

Ma and Li(2019)[11]은 긴 문서를 요약할 때 가독성과 감성을 유지하는 요약기법을 제안한다. 위 논문에서는 심층 신경망 모델을 활용하여 계층적 문서 인코더와 문장 추출기, 감정 분류기 및 GAN 기반 판별기로 구성된 약하게 지도(supervised)되는 추출 프레임워크를 제안한다. 실험결과 효과적인 압축 비율로 요약을 생성하였다.

3. 제안하는 다중문서 요약기법

제안하는 기법은 Fig. 1과 같이 4단계로 이루어진 전체적인 구성도이다. 첫 번째는 요약에 도움이 되는 유용한 리뷰를 분류하기 위해 형태소 분석과 감성어 사전을 활용하여 전처리를 수행한다. 두 번째는 리뷰가 가지고 있는 주제를 추출하기 위해 학습된 언어모델 BERT를 활용하여 워드 임베딩을 추출하여 핵심 주제 후보들의 클러스터링을 수행한다. 세 번째는 현재 단계까지 분류된 결과를 바탕으로 중요 문장을 평가하고 상위 스코어의 문장을 추출한다. 최종적으로 요약문을 생성하는 Transformer 모델을 기반으로 주제 및 감정별 요약문을 생성한다.

3.1 데이터 수집 및 전처리

데이터는 일반적인 소비자와 접근성이 높은 네이터 쇼핑몰 상품 리뷰를 수집하였다. 리뷰 정보에는 게시 날짜, 별점, 리뷰 내용이 포함되어있다. 리뷰 내용의 맞춤법 교정은 Naver의 맞춤법 검사기 기반의 py-hanspell을 활용하여 진행하였다. 형태소 분석기는 실험결과 오탈자에 강인한 KOMORAN

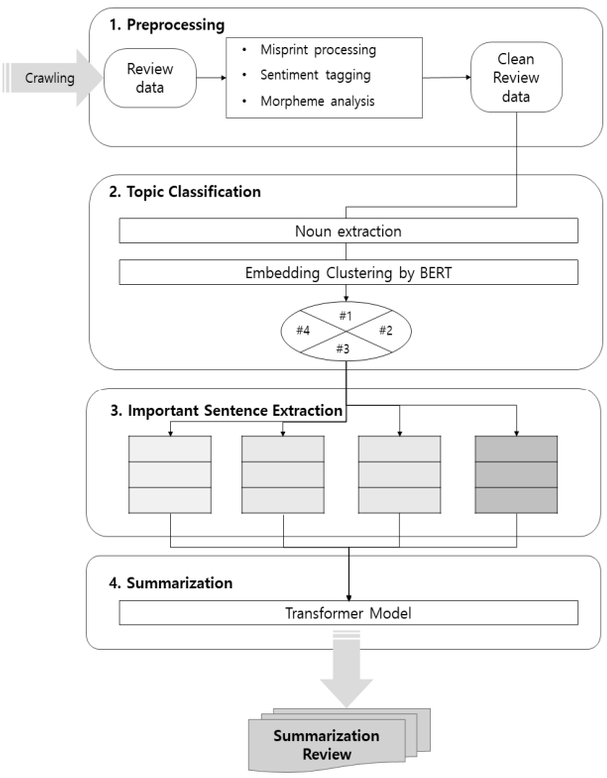

Fig. 1. Proposed Multi-document Summarization Method Architecture

을 활용하여 각각의 단어에 형태소를 태깅하였다. 또한 상품의 구매에 영향을 미치는 요소로 리뷰에 내용에 감정어가 포함되면 리뷰의 유용성(helpfulness)이 높아진다[12]. 따라서 군산대학교의 KNU 한국어 감성사전을 활용하여 감정어가 존재하지 않는 리뷰는 제외하였으며 해당 리뷰가 어떤 감정을 가지는 리뷰인지 정수(양의 정수는 긍정, 음의 정수는 부정)로 태깅하였다.

3.2 주제 분류

먼저 리뷰가 내포하고 있는 주제를 추출하기 위해 단어 빈도수를 활용하였다. 주제의 형태소는 명사로 지정하여 추출하였으며 전체 문서에서 단어의 빈도수와 상대적인 중요도를 평가할 수 있는 그래프 기반의 TextRank 알고리즘의 키워드 추출을 활용하였으며 Equation (1)과 같다.

$$TR(V_i) = (1-d) + d \times \sum_{V_j \in in(V_i)} \frac{w_{ji}}{\sum_{V_j \in out(V_i)} w_{jk}} TR(V_j) \qquad (1)$$

V_i 는 현재 문장을 의미하고 $in(V_i)$는 현재 문장과 같은 단어를 포함하는 다른 문장의 집합을 의미한다. $Out(V_j)$는 V_j와 같은 단어를 포함하는 다른 문장의 집합을 의미한다. w_{ij}는 V_i와 V_j의 가중치를 의미한다. d는 무작위로 접근될 확률값으로 기본값으로 0.85를 할당한다. 그리고 $TR(V_i)$을 임의로 설정한 위 수식을 통해 일정한 값에 수렴할 때까지 반복한다. 그 결과, 문서 한 개 즉, 상품 하나의 리뷰를 바탕으로 추출된 명사의 중요도를 산출한 결과 상위 10개를 내림차순으로 정렬한 결과는 다음 Table 1과 같다. 명사의 종류는 한국어 형태소 품사 태그 기준에 따라 NNG(Common nouns), NNP(Proper nouns)로 태깅하였다. 위 결과를 바탕으로 주제를 분류하기 위해 단어를 BERT(Bidirectional Encoder Representations Transformers) 단어 모델을 활용하여 워드 임베딩 벡터로 치환하였다. 미리 학습된 언어모델을 통하여 워드 임베딩을 추출하면 학습한 문장을 기반으로 문장 내에서 문맥을 고려한 벡터값을 얻을 수 있다.

Table 1. Calculation Result of Extracted Keyword

Word	TR
사용(use)/NNG	38.45
구매(purchase)/NNG	29.90
향(scent)/NNG	27.70
가격(price)/NNG	24.35
용량(capacity)/NNP	22.71
스킨(liquid lotion))/NNP	15.43
피부(skin)/NNG	14.20
제품(product)/NNG	12.12
느낌(feeling)/NNG	11.42
대용량(bulk)/NNG	11.34

본 논문에서는 SKTBrain이 한국어 위키를 기반으로 학습한 KoBert 언어모델을 활용하여 임베딩 벡터를 추출하였다. 그리고 추출된 벡터의 군집화를 실시하여 벡터를 k개의 클러스터로 분류한다. 여기서 클러스터 개수의 기준은 실험결과 저자가 가장 적절하게 분류되었다고 판단한 수로 지정하였으며 군집화의 결과는 Table 2와 같다.

클러스터를 수행한 결과 의미가 비슷한 단어가 같은 클러스터에 존재하는 것을 확인했다. 위 결과를 통해 리뷰에서 각각의 클러스터에 존재하는 단어를 2개 이상 포함하는 문장을 해당 클러스터의 주요한 내용을 포함하고 있다고 판단하여 분류한다. 만약 문장이 해당하는 클러스터가 두 개 이상이라면 클러스터의 종류를 모두 표기한다.

Table 2. Result of Clustering

Cluster	Word
#1	사용(use), 용량(capacity), 성분(ingredient), 효과(effect)
#2	구매(purchase). 가격(price), 구입(purchase), 주문(order)
#3	향(scent), 냄새(smell), 향도(scent too)
#4	느낌(feeling), 생각(thinking), 대비(comparition), 것(thing), 때(moment), 전(previous time)
#5	피부(skin), 폼(foam), 화장(makeup), 얼굴(face)

Table 3. Summarization Result of Classified Review

Cluster	Word		Summarization
#1	사용(use), 용량(capacity), 성분(ingredient), 효과(effect)	positive	'양도 많고 가격도 저렴하고 해서 구입했어요. 좋아요. 양도 많은데 한 개 가격에 두 개나 받아서 좋아요'
		negative	'확실히 피부가 맑아 보이고 트러블이 좀 자잘하게 나는 편인데 이걸로 하고서부터는 뭐가 나는 일이 보기 드문 것 같아 아쉬운 점이 있다면 향이 조금 순하고 연하고 할 줄 알았는데 순한거는 모르겠지만 연하지는 않은 것 같아요'
#2	구매(purchase), 가격(price), 구입(purchase), 주문(order)	positive	'재구매 의사 백프로로 민감성 피부인데 진짜 좋아서 착한 가격에 대용량 스킨을 구매해서 판매자님께 감사하게 생각하고 계속 구매할께요'
		negative	'양도 정말 많은데 그냥저냥 괜찮아요. 막 와닿게 좋은 건 아니에요.'
#3	향(scent), 냄새(smell), 향도(scent too)	positive	'향도 진하지 않고 편하게 사용하기 좋습니다. 받자마자 써봤는데 향이 너무 좋아요'
		negative	'향에 민감하시거나 거부감이 드시는 분들도 가볍게 쓰시기 괜찮을 것 같고 재구매는 안 할 것 같네요.'

3.3 중요 문장 추출

앞서 클러스터링을 통해 분류된 문장 중에 중요한 문장을 추출하기 위해 TextRank 알고리즘을 활용한다. TextRank 알고리즘은 키워드 추출뿐 아니라 Equation (2)를 통해 문장의 유사도를 간선의 가중치로 설정하면 문장의 중요도 산출이 가능하다.

$$Similarity(V_i, V_j) = \frac{|\{w_k | w_k \in V_i \& w_k \in V_j\}|}{\log(|V_i|) + \log(|V_j|)} \quad (2)$$

따라서 본 논문에서는 분류된 문장을 앞서 키워드 추출에 활용했던 TextRank 알고리즘에 문장의 유사도를 적용하여 분류된 클러스터별로 문장의 중요도를 산출하여 계산한다.

3.4 최종 요약 생성

마지막으로 분류와 중요도 평가가 완료된 리뷰를 대상으로 요약 생성을 수행한다. 요약의 기법은 Multi-head Attention 기반의 Transformer 모델을 활용한다. Transformer 모델은 기존의 요약 생성 기법인 LSTM 기반의 인코더-디코더 모델과 비교하여 우수한 성능을 나타냈다[13]. Transformer 모델의 핵심은 인공신경망을 배제하고 Attention 기법만을 활용하여 병렬 컴퓨팅을 가능하게 하고 인공신경망과 비교해 빠른 연산이 가능하다는 점이다.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (3)$$

Equation (3)은 Attention 기법을 표현하는 수식으로 Query, Key, Value 3가지의 가중치 행렬로 구성된다. 입력되는 워드 임베딩을 통해 가중치 행렬을 학습하여 어떤 단어에 집중할지 고려한 출력값을 얻을 수 있다. 이와 같은 하나의 Attention 과정을 모델 내부에서 Attention head라고 하며 여러 개의 Attention head가 동시에 수행되어 집중하는 단어가 편향되지 않게 보정 하는 것을 Multi-head Attention이라고 한다. Transformer 모델은 인코더-디코더의 구조가 Multi-head Attention으로 이루어진 것을 의미하며 요약 문제 해결을 위해 활용이 가능하다. 본 논문에서는 Transformer 모델을 통해 인코더에 요약 전 원문을 입력하고 요약된 문장을 디코더에서 출력하는 Transformer 모델을 설계하고 학습한다. 본 논문에서 Query, Key, Value 행렬 벡터의 차원은 64로 지정하였으며 인코더와 디코더의 레이어는 12개의 레이어로 구성하고 input 및 output의 최대 길이는 512로 지정하였으며 multi-head의 개수는 12로 지정하여 Transformer 모델을 구성했다. 요약 학습 데이터는 DACON의 한국어 문서 추출요약 AI 경진대회[14]에서 제공하는 뉴스 요약 데이터를 활용하여 학습을 진행했다. 학습 데이터는 42,803건의 데이터로 이루어졌으며 데이터의 크기는 112MB이다. 데이터 속성은 기사 출판 미디어, 데이터 고유 번호, 전체 기사 내용, 사람이 생성한 요약문, 사람이 추출한 요약문의 인덱스로 총 5개의 속성으로 구성된다. 마지막으로 학습된 모델로 최종 분류된 리뷰를 요약한 결과는 Table 3과 같다.

4. 성능 평가

제안하는 기법의 성능을 평가하기 위해 LSTM 기반의 인코더-디코더 구조에 Attention 기법을 추가적인 필터 없이 적용한 seq2seq 요약모델과 성능을 비교하였다. 성능평가는 Transformer 모델을 학습할 때 사용했던 학습 데이터에서 테스트 데이터를 임의로 추출하여 정답 요약과 비교하여 평가하였으며 지표로는 ROUGE[15]를 활용한다. ROUGE 스코어는 n-gram을 기반으로 문장의 재현율 검증하기 위해 추측 요약과 정답 요약의 겹치는 수를 정답 요약의 단어의 수로 나누는 Recall과 문장의 정밀도를 검증하기 위해 추측 요약과 정답 요약의 겹치는 수를 추측 요약의 단어의 수로 나누는

Table 4. ROUGE Score Measure Result

Model	ROUGE-1	ROUGE-2	ROUGE-L
Vanilla seq2seq	30.96	18.23	22.03
Proposed	38.73	28.18	35.08

Table 5. Calculation Result of Cosine Similarity

Model	Cosine Similarity
Vanilla seq2seq	76.61
Proposed	88.77

Precision과 두 가지를 모두 종합한 F1-score로 평가된다. ROUGE 스코어의 종류는 ROUGE-1, ROUGE-2, ROUGE-L로 이루어져 있으며 각각 unigram, bigram, 최장 길이로 매칭되는 문자열을 측정을 수행한다. 성능평가의 결과는 Table 4와 같다.

성능평가 결과 기존의 seq2seq 모델과 비교하여 뛰어난 성능의 요약 성능을 나타냈다. 본 논문에서 제안하는 기법은 정답 요약이 존재하지 않는 태스크이기 때문에 더욱 다양한 방법으로 요약결과를 평가하기 위해 요약 원본과 추측 요약의 워드 임베딩을 활용하여 코사인 유사도(Cosine Similarity)의 평균을 산출하여 요약문이 원문과 얼마나 비슷한지 실험을 진행하였다. 코사인 유사도 결과는 Table 5와 같다.

실험결과 제안하는 모델이 생성한 추측 요약은 seq2seq 모델과 비교하여 상대적으로 원문과 비슷한 문장을 생성했다는 결과를 도출했다.

5. 결 론

본 논문에서는 다중문서가 가지고 있는 다양한 주제를 요약할 때 주제의 손실을 최소화하며 요약을 수행하기 위한 기법을 제안하였다. 다중문서를 형태소 분석을 통해 명사를 추출하고 문서 전체를 고려하여 중요한 명사를 평가하여 도출하였다. 도출된 단어는 단어의 문맥을 고려하기 위해 학습된 BERT 모델을 통해 워드 임베딩으로 치환하여 클러스터링을 수행하여 단어의 의미별로 클러스터를 나누어 단어 세트를 구성하였다. 단어 세트를 활용하여 각각의 문장이 어떤 클러스터에 해당하는지 판별하고 감정어 사전을 통해 긍정, 부정 문장을 판별하여 문장을 분류하였다. 최종적으로 분류된 문장은 Transformer 모델을 학습하여 요약을 수행하였다. 수행결과 기존의 seq2seq 요약모델과 비교하여 우수한 성능을 나타냈다. 그러나 요약 생성은 요약하려는 원문의 영향을 많이 받는 것으로 나타났으며 온라인 리뷰와 같이 맞춤법 및 미등록 단어가 빈번한 문장에 대해 요약 성능이 크게 저하되었다. 향후 연구에서는 원문의 오탈자 및 미등록 단어에 강인한 요약 방법에 관해 연구를 진행할 계획이다.

References

[1] T. Liu, "The support of online reviews on user shopping process," Master's Thesis, Kyung Hee University, 2016.

[2] S. Harer, S. Kadam, and R. Kaptein, "Mining and summarizing movie reviews in mobile environment," *International Journal of Computer Science & Information Technologies*, Vol.5, No.3, pp.3912-3916, 2014.

[3] C. V. Gupta, and G. S. Lehal, "A survey of text summarization extractive techniques," *Journal of Emerging Technologies in Web Intelligence*, Vol.2, No.3, pp.258-268, 2010.

[4] J. U. Heu, I. Qasim, and D. H. Lee, "FoDoSu: Multi-document summarization exploiting semantic analysis based on social Folksonomy," *Information Processing & Management*, Vol.51, No.1, pp.212-225, 2015

[5] K. S. Jones, "Automatic summarising: The state of the art," *Information Processing & Management*, Vol.43, No.6, pp.1449-1481. 2007.

[6] M. Allahyari, et al., "Text summarization techniques: A brief survey," *International Journal of Advanced Computer Science and Applications*, Vol.8, No.10, pp.397-405, 2017.

[7] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," *Proceedings of The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.93-98, pp.93-98, 2016.

[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *3rd ICLR 2015 as Oral Presentation*, 2015.

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL-HLT*, No.1, pp.4171-4186, 2019.

[10] J. Tan, A. Kotov, R. P. Mohammadiani, and Y. Huo, "Sentence retrieval with sentiment-specific topical anchoring for review summarization," *ACM Conference on Information and Knowledge Management,* pp.2323-2326, 2017.

[11] Y. Ma and Q. Li, "A weakly-supervised extractive framework for sentiment-preserving document summarization," *World Wide Web*, Vol.22, No.4, pp.1401-1425, 2019.

[12] H. Lee, "The Relational Analysis between Types of Online Hotel Review and Usefulness according to the Hotel Class," *Korean Management Review*, Vol.46, No.1, pp.137-156, 2017.

[13] A. Vaswani, et al., "Attention is all you need," In *Advances in Neural Information Processing Systems*, pp.6000-6010. 2017.

[14] Dacon. Korea Data Competition Platform. Extracting Summary of Korean Document Contest [Internet], https://dacon.io/competitions/official/235671/overview/description

[15] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," *Proceeding of the Workshop on Text Summarization Branches Out,* 2004.

이 필 원

https://orcid.org/0000-0003-4092-8658
e-mail : pwlee@soongsil.ac.kr
2020년 ~ 현 재 숭실대학교 컴퓨터학과 석사과정
관심분야 : 자연어처리, 인공지능, IoT, 클라우드 컴퓨팅

최 종 석

https://orcid.org/0000-0002-2959-678X
e-mail : jschoi@ssu.ac.kr
2012년 숭실대학교 컴퓨터학과(석사)
2015년 숭실대학교 컴퓨터학과 (박사수료)
2019년 ~ 현 재 (주)공감하다 대표
2020년 ~ 현 재 숭실대학교 스파르탄SW교육원 교수
관심분야 : 데이터 분석, IoT, 네트워크, 클라우드 컴퓨팅

황 윤 영

https://orcid.org/0000-0002-5239-3796
e-mail : doublewhy@soongsil.ac.kr
2018년 성결대학교 도시계획부동산학부 (학사)
2020년 ~ 현 재 숭실대학교 컴퓨터학과 석사과정
관심분야 : IoT, 빅데이터, 클라우드 컴퓨팅, 네트워크

신 용 태

https://orcid.org/0000-0002-1199-1845
e-mail : shin@ssu.ac.kr
1985년 한양대학교 산업공학과(학사)
1990년 Univ. of lowa, 컴퓨터학과(석사)
1994년 Univ. of lowa, 컴퓨터학과(박사)
1995년 ~ 현 재 숭실대학교 컴퓨터학부 교수
관심분야 : 정보보호, 인터넷 프로토콜, IoT, 클라우드 컴퓨팅