

Citation-based Article Summarization using a Combination of Lexical Text Similarities: Evaluation with Computational Linguistics Literature Summarization Datasets

In-Su Kang*

Abstract

Citation-based article summarization is to create a shortened text for an academic article, reflecting the content of citing sentences which contain other's thoughts about the target article to be summarized. To deal with the problem, this study introduces an extractive summarization method based on calculating a linear combination of various sentence salience scores, which represent the degrees to which a candidate sentence reflects the content of author's abstract text, reader's citing text, and the target article to be summarized. In the current study, salience scores are obtained by computing surface-level textual similarities. Experiments using CL-SciSumm datasets show that the proposed method parallels or outperforms the previous approaches in ROUGE evaluations against SciSumm-2017 human summaries and SciSumm-2016/2017 community summaries.

▶ Keyword: Citation, Summarization, Text Similarity

I. Introduction

요약 대상 논문을 RP(Reference Paper)라고 할 때, 인용 기반 논문 요약은 RP의 인용텍스트 정보를 활용하여 논문을 자동 요약하는 것이다. 인용텍스트는 아래 가상 예시와 같이 특정 논문 P를 인용하는 논문(예: P1, P2) 내에 P에 대해 기술된 텍스트를 의미한다. 동일 피인용논문에 대한 서로 다른 인용텍스트들의 모음은 해당 피인용논문에 대한 다른 연구자들의 의견과 평가가 포함되어 있어 인용텍스트에 기반한 논문 요약은 요약 대상 논문의 내용과 기여점에 대한 새로운 관점을 제시하는 장점을 갖는다.

- 논문 P1: Brown 등(Brown et al., 1997)이 제시한 문서 색인법은 높은 시간복잡도의 한계를 안고 있다.

- 논문 P2: Brown 등(Brown et al., 1997)은 지식베이스와 대용량 웹 컬렉션 분석에 기반한 색인법을 통해 검색 성능을 크게 향상시켰다.

기존 인용 기반 논문 요약은 RP에 대한 인용문 집합으로부터 요약문을 추출하는 인용문 중심 요약 방법[1,2]과 인용문에 대응하는 RP 내 문장(이후 피인용문)을 중심으로 요약문을 추출하는 피인용(문) 중심 요약 방법으로 나뉜다. 이 중 피인용 중심의 요약법은 최근 3년간 국제 규모의 CL-SciSumm Shared Task로 진행되어 평가되었으며 다양한 접근법들이 시도되고 있다[3,4,5]. 이 연구에서는 피인용 중심 논문 요약법을 다룬다.

피인용 중심 논문 요약의 기존 대부분의 방법들은 각 RP 문장이 RP 및 인용문의 내용을 표현하는 정도를 텍스트 유사도 방식으로 계산하였으며, 이후 신경망 학습[6,7], 자질선형결합[8,9,10,11,12], DPP 샘플링[13], 그래프 기반 랭킹[6, 14,15,16], 클러스터링[12,15,17], MMR[16,17,18], 언어모델링[19,20] 등의 방법론을 단독 및 결합 적용하여 요약문 추출을 시도하였다.

*First Author: In-Su Kang, Corresponding Author: In-Su Kang

*In-Su Kang (dbaisk@ks.ac.kr), Dept. of Computer Science, Kyungsoong University

Received: 2019. 06. 05, Revised: 2019. 07. 26, Accepted: 2019. 07. 28.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2015R1D1A1A01060489).

이 연구에서는 기존 자질선형결합 방법론 내에서 표층 어휘 수준의 텍스트 유사도 계산에 기반한 추출식 논문 요약물 시도한다. 기존 자질선형결합 방법에서는 텍스트 유사도 이외에도 RP 문장에 대해 길이, 문장 내 위치, TextRank, Rhetorical 클래스 분류, 계층적 토픽 모델, 단서어구 정보 등에 기반한 문장 중요도 자질들이 추가 활용되었다. 이 연구에서는 기존 시도된 요약적합도들을 포함하여, 인용 기반 논문 요약에 특화된 문장 요약적합도들을 텍스트 유사도에 기반하여 새롭게 제안하고 이들의 선형결합을 통해 추출 요약문을 생성한다. 실험에서는 CL-SciSumm Shared Task용 요약 평가 집합을 사용하여 제안된 방법과 인용 기반 학술 논문 요약의 최신 방법들의 요약 성능을 비교 제시한다.

논문의 구성은 다음과 같다. 2장에서는 기존 연구에 대해 기술한다. 3장에서는 본 논문에서 제안하는 요약 방법에 대해 기술한다. 4장에서는 제안된 방법의 성능 평가 결과를 제시하고 5장에서 결론을 맺는다.

II. Related Works

전술한 바와 같이 인용 기반 논문 요약은 인용문 중심 요약과 피인용문 중심 요약으로 구분된다. 인용문 중심 요약[1,2]은 RP에 대한 인용문 집합에 군집법 및 기계학습 분류기를 적용하여 인용문들을 그룹으로 나누고 각 그룹의 대표 인용문을 추출하는 방식으로 요약문을 생성하였다.

피인용문 중심 요약은 CL-SciSumm Shared Task에서 평가하는 논문 요약법이다. 최근 3년간 개최된 CL-SciSumm Shared Task에서는 학술 논문 요약에 위한 세부 방법들로 피인용문 인식(Task-1A), 피인용문 패킷 결정(Task-1B), 피인용문 기반 논문 요약(Task-2)의 세 가지 태스크를 전산언어학 분야 논문 집합을 대상으로 평가하고 있다[3,4,5]. Task-1A는 RP에 대한 인용문에 대응하는 RP 내의 문장(들)을 결정하는 작업이며, Task-1B는 Task-1A에서 결정된 피인용문에 Aim, Hypothesis, Method, Result, Implication의 패킷을 부여하는 작업이고, Task-2는 이전 단계에서 결정된 패킷 부여된 피인용문 집합을 활용하여 요약문을 생성하는 작업이다.

SciSumm-2016 및 2017 요약 평가에서 높은 성능을 보인 CIST 팀은 자질선형결합방법과 DPP(Determinantal Point Processes) 기반 방법을 제안하였다[10]. 자질선형결합방법은 문장 길이, 문장 위치, 용어 빈도수 기반 문장-제목 코사인 유사도, 문장의 피인용텍스트 여부, 계층적 토픽 모델(hLDA) 기반 문장 점수 자질들의 가중합에 기반하여 상위 문장들을 요약문으로 추출한다. DPP 기반 방법은 자질선형결합에 기반한 문장 quality 벡터와 문장-문장 간 유사도 행렬을 입력 받아 DPP 샘플링을 적용하여 얻어지는 문장 부분 집합으로부터 요약문을 추출하는 방법으로 DPP를 통해 요약문의 내용 다양성을 높이고자 하였다.

SciSumm-2018 요약 평가에서 높은 성능을 보인 Abura'ed 등의 방법은 학습된 CNN(Convolutional Neural Network)을 통해 요약 대상 논문 내 각 문장에 대한 요약적합도를 계산하였다[6]. CNN의 입력으로는 문장의 단어 임베딩과 문맥자질 값들을 사용하였으며, 정답 출력 값으로는 문장과 정답 요약텍스트 내 각 문장 간 코사인 유사도들의 최대값을 사용하였다. 문맥자질 값으로는 문장-초록 유사도, 문장-논문 유사도, 문장-인용문 최소/최대/평균 코사인 유사도, 문장의 논문/섹션 내 위치, 문장 내 용어 tf-idf 값들의 총합의 정규화 값, 문장의 TextRank 점수 등을 활용하였다.

Ma 등은 non-negative matrix factorization을 통해 문장 벡터의 차원을 축소하고, bisecting K-means를 통해 문장 집합을 클러스터링한 다음, 군집 크기가 큰 순으로 각 군집에서 MMR에 기반하여 요약문을 추출하였다[17]. Ma 등은 초록 내 문장들을 motivation, approach, conclusion의 세 그룹 텍스트들로 분할한 후, 각 피인용문을 초록 내 그룹 텍스트와의 유사도(Jaccard, idf, tfidf 사용)에 따라 세 그룹 중 하나에 할당하였으며, 각 그룹으로부터 문장 점수(Jaccard/idf/tfidf 유사도, 문장길이, 문장위치 자질 선형결합)를 사용하여 요약문을 추출하였다[12].

Cao 등은 Manifold Ranking을 인용 기반 요약에 적용하였는데, 문장-문장 유사도 행렬과 문장-인용문 유사도 행렬의 일차결합에 대해 Random Walk를 통해 문장의 중요도를 결정하였다[14]. Lauscher 등은 Learning to Rank 모델을 통해 각 인용문에 대해 결정된 피인용문들의 모음 S를 클러스터링하고, S에 TextRank를 적용한 후, 군집 내 문장의 TextRank 평균 점수가 높은 군집 순으로 각 군집에서 최대 TextRank 점수를 갖는 문장을 요약문으로 추출하였다[15]. Cohan과 Goharian은 인용문을 질의로 사용하여 요약 대상 논문 내 각 문장을 검색하는 방식으로 피인용문을 인식하고, SVM 분류기를 통해 피인용문을 패킷 그룹에 할당한 다음, 패킷 그룹들로부터 PageRank 및 MMR에 기반하여 요약문을 추출하는 방법을 시도하였다[16].

Conroy와 Davis는 논문 요약 작업을 섹션 혼합 언어모델을 이용하여 논문 내용 포함성을 최대화하고 중복도를 최소화하는 RP 내 문장들의 최적 집합을 선택하는 문제로 다루었으며, 섹션들은 RP의 초록/결과/기타 섹션들 및 인용문 모음 섹션으로 구분하였다[20]. Mei와 Zhai는 요약 대상 논문 내 각 문장의 sentence 언어모델과 impact 언어모델 간에 계산되는 negative Kullback-Leibler divergence를 문장순위화 함수로 사용하여 상위 순위 문장들을 요약문으로 추출하였다[19]. Impact 언어모델은 개별 인용문의 중요도에 기반하여 인용텍스트 모음과 요약 대상 논문 텍스트로부터 추정되었는데, 인용문의 중요도는 인용문이 인용표지에 근접한 정도와 인용문이 출현한 논문이 대용량 논문-논문 인용 네트워크에서 갖는 페이지랭크 값의 곱으로 정의하였다.

Yasunaga 등은 요약문 후보로 피인용문들과 함께 초록 내 문장을 동시 고려하는 접근법을 시도하였는데, 이를 위해 1000

여 편 논문으로 구성된 논문 요약 학습 데이터셋으로부터 학습된 Graph Convolutional Networks(GCN)를 활용하였다[7].

Debnath 등은 OpenNMT 툴을 이용한 추상식 요약을 시도하였으나 학습데이터의 부족으로 인해 실효성을 얻지 못하였다[21].

기존 방법들을 교사 방법과 비교사 방법으로 구분할 때, 최근 신경망 기반의 교사 기반 요약법들이 높은 요약 성능을 보이고 있으나[6,7], 이 방법들은 학습 데이터 구축 및 신경망 학습 과정이 요구된다는 부담이 있다. 한편 기존 대부분의 비교사 방법들은 문장의 요약적합도 계산을 위해 텍스트 유사도와 함께 TextRank, 군집화, MMR 등의 방법론을 병행 활용하였다. 이에 비해 제안된 방법은 표층 수준의 텍스트 유사도만을 활용한다는 점에서 차이가 있다.

III. The Proposed Method

1. Citation-based Summarization employing Lexical Text Similarities

이 절에서는 본 논문에서 제안하는 텍스트 유사도 기반 논문 요약 방법에 대해 기술한다. 요약 시스템의 입력으로 요약 대상 논문 RP의 초록과 본문 텍스트, 그리고 RP의 인용텍스트 집합이 주어진다고 가정하면, 텍스트 유사도 기반 요약문 생성 절차는 다음과 같다. 먼저 요약 대상 논문의 본문 텍스트 내 각 문장 s 에 대해 요약적합도(salience score)를 계산한 후, 상위 요약적합도를 갖는 문장들을 요약텍스트 길이 제약 내에서 추출한다. 다음으로 추출된 문장들을 요약 대상 논문에서의 출현 순서 기준으로 재정렬하여 최종 요약텍스트를 생성한다.

문장 s 에 대한 요약적합도는 식 1과 같이 s 에 대한 서로 다른 유형의 총 8가지 요약적합도들의 가중합으로 정의된다.

$$Salience(s) = \sum_{t \in \{abs.cit, ABS, CIT, DOC, grp, seg, AD\}} \lambda_t \times Salience_t(s) \quad (1)$$

초록문장유사도 누적 요약적합도 $Salience_{abs}(s)$ 는 식 2와 같이 초록 텍스트(총 m 개 문장 가정)를 구성하는 개별 초록 문장 $abstractSentence_i$ 와 문장 s 와의 텍스트 유사도를 전체 초록 문장 집합에 대해 누적한 후, 비교 대상이 되는 RP 본문 텍스트 내 전체 문장들에 대해 정규화한 값으로 정의한다. 식 2를 통해 문장 s 가 논문 저자 관점의 선별적 핵심 내용들을 포함하는 정도를 계량화할 수 있다. 식 2의 유사도 누적과 달리, [12]에서는 문장 s 와 초록 내 부분 텍스트들과의 최대 유사도를 계산하였다.

$$Salience_{abs}(s) = \frac{\sum_{i=1 \sim m} sim(abstractSentence_i, s)}{\sum_{s'} \sum_{i=1 \sim m} sim(abstractSentence_i, s')} \quad (2)$$

인용문장유사도 누적 요약적합도 $Salience_{cit}(s)$ 는 식 3과 같이 인용텍스트 집합(총 n 개 문장 가정)를 구성하는 개별 인용문 $citation_i$ 과 문장 s 와의 텍스트 유사도를 전체 인용문 집합에 대해 누적한 후 정규화한 값이다. 이를 통해 여러 인용 논문들에서 반복적으로 인용되는 내용들을 많이 포함하는 문장의 요약적합도를 높일 수 있다. 식 3의 유사도 누적 방식과 달리, [6]에서는 문장 s 와 각 인용문 간 유사도들의 최대/최소/평균들을 활용하였고, [22]에서는 문장 s 와 각 인용문 간 유사도들의 평균을 요약적합도로 사용하였다.

많은 기존 연구들은 SciSumm의 Task-1A를 통해 결정된 피인용문 집합 정보를 활용하였는데, [9,10,13]에서는 문장 s 의 피인용문 여부를 요약적합도 계산에 활용하였으며, [12,15]에서는 요약문 추출 대상을 전체 RP 문장 집합 대신 피인용문 집합으로 제한하는 용도로 활용하였다. 한편, [14]에서는 Manifold Ranking을 위한 그래프 생성 시 문장 s 와 총 n 개 각 인용문 간에 유사도 부착 간선을 설정하였는데 이는 식 3과 개념적으로 유사하다.

$$Salience_{cit}(s) = \frac{\sum_{i=1 \sim n} sim(citation_i, s)}{\sum_{s'} \sum_{i=1 \sim n} sim(citation_i, s')} \quad (3)$$

초록 텍스트 기반 요약적합도 $Salience_{ABS}(s)$ 는 식 4와 같이 초록 전체를 하나의 텍스트로 고려하여 초록 전체 텍스트 $AbstractText$ 와 문장 s 와의 텍스트 유사도를 구한 후 정규화한 값으로, 문장 s 가 초록의 전체적인 내용을 표현하는 정도를 계산한다. [6]에서는 초록 텍스트를 초록 내 문장 벡터들의 평균 벡터로 변환한 후 문장 s 와의 유사도를 계산하였다.

$$Salience_{ABS}(s) = \frac{sim(AbtractText, s)}{\sum_{s'} sim(AbtractText, s')} \quad (4)$$

인용문 집합 텍스트 기반 요약적합도 $Salience_{CIT}(s)$ 는 식 5와 같이 인용문 집합 전체를 하나의 텍스트로 고려하여 인용문 집합 전체 텍스트 $AllCitationText$ 와 문장 s 와의 정규화된 텍스트 유사도로 정의되며, 문장 s 가 인용문 집합의 전체적인 내용을 담고 있는 정도를 의미한다.

$$Salience_{CIT}(s) = \frac{sim(AllCitationText, s)}{\sum_{s'} sim(AllCitationText, s')} \quad (5)$$

본문 텍스트 기반 요약적합도 $Salience_{DOC}(s)$ 는 식 6과 같이 요약 대상 논문의 본문 텍스트 전체를 하나의 단위로 고려하여 본문 전체 텍스트 $DocumentText$ 와 문장 s 와의 정규화된 유사도로 정의되며, 문장 s 가 요약 대상 논문의 전체적인 내용을 대표하는 정도를 의미한다. 식 6과 달리, [6]에서는 논문 텍스트를 그 문장 벡터들의 평균 벡터로 변환한 후 문장 s 와의 유사도를 계산하였다.

$$Salienc_{Doc}(s) = \frac{sim(DocumentText, s)}{\sum_{s'} sim(DocumentText, s')} \quad (6)$$

군집 기반 요약적합도 $Salienc_{grp}(s)$ 는 식 7과 같이 정의된다. 요약 대상 논문의 본문 텍스트 내 문장 집합을 군집화(총 k 개 군집 가정)하여 얻은 각 군집 내 문장들의 전체 텍스트 모음을 $GroupText_i$ 라고 할 때, 식 7은 개별 군집 텍스트 $GroupText_i$ 와 문장 s 와의 유사도를 전체 군집 집합에 대해 누적한 후 정규화한 값이다. 이는 학술 문헌의 내용이 여러 하위 토픽들로 구성되어 있다고 가정하고 문장 s 가 개별 토픽들의 내용을 포괄하는 정도를 계산한 것으로, 문장을 토픽에 대응시키기 위해 군집화를 적용한 것이다. [15,17]에서는 각 군집으로부터 요약문을 추출하기 위해 RP 문장 집합의 군집화를 적용하였다.

$$Salienc_{grp}(s) = \frac{\sum_{i=1 \sim k} sim(GroupText_i, s)}{\sum_{s'} \sum_{i=1 \sim k} sim(GroupText_i, s')} \quad (7)$$

세그먼트 기반 요약적합도 $Salienc_{seg}(s)$ 는 식 8과 같이 정의된다. 요약 대상 논문의 본문 텍스트를 j 개 부분 텍스트들로 선형 분할(총 j 개 분할 텍스트 가정)하여 얻은 각 부분 텍스트를 $SegmentText_i$ 라고 할 때, 식 8은 개별 분할 텍스트 $SegmentText_i$ 와 문장 s 와의 유사도를 전체 분할 텍스트들에 대해 누적한 후 정규화한 값이다. 이는 학술 문헌 텍스트가 IMRaD(Introduction, Method, Result, and Discussion)[23]와 유사한 정형화된 선형 구조를 갖는다고 가정하고 문장 s 가 서로 다른 각 영역 텍스트의 내용을 누적 표현하는 정도를 계산한 것이다.

$$Salienc_{seg}(s) = \frac{\sum_{i=1 \sim j} sim(SegmentText_i, s)}{\sum_{s'} \sum_{i=1 \sim j} sim(SegmentText_i, s')} \quad (8)$$

인접문맥 기반 요약적합도 $Salienc_{Adj}(s)$ 는 식 9와 같이 문장 s 의 인접 문장들의 모음을 $AdjacentText(s)$ 라고 할 때, 문장 s 와 인접텍스트와의 유사도를 정규화한 값으로, 문장 s 가 인접 문장들의 내용을 대표하는 정도를 계량화한 것이다. [9,10,13]에서는 피인용문 인식을 위해 문장 s 와 이전/이후 문장들과의 유사도 곱의 제곱근을 사용하였으며, 정규화 부분을 제외하면 식 9와 개념적으로 유사하다. [6]에서는 문장 s 와 s 의 이전/이후 3개씩 각 문장에 대해 초록/논문/제목/인용문 등과의 다양한 유사도 자질들을 CNN의 입력으로 사용하였다. 식 9에서는 문장 s 및 그 인접 문장들의 모음과 문장 s 와의 정규화된 유사도를 계산한다.

$$Salienc_{Adj}(s) = \frac{sim(AdjacentText(s), s)}{\sum_{s'} sim(AdjacentText(s'), s')} \quad (9)$$

위 식 2~9에서 두 텍스트 간 유사도를 계산하는 함수 $sim()$ 에 대해 본 연구에서는 식 10의 Jaccard 유사도 수식[24]을 사용한다. 식 10에서 X 는 식 2~9에서의 $abstractSentence_i$, $citation_i$, $AbstractText$, $AllCitationText$, $DocumentText$, $GroupText_i$, $SegmentText_i$, $AdjacentText(s)$ 에 대한 변수이며, $set(X)$, $set(s)$ 는 각각 X , s 내 출현 용어들의 집합 표현을 의미한다. Jaccard 유사도는 기존 대부분의 연구들에서 텍스트 유사도 계산을 위해 활용되었다[3,4,5].

$$sim(X, s) = \frac{|set(X) \cap set(s)|}{|set(X) \cup set(s)|} \quad (10)$$

2. Example of Calculating Sentence Salienc

이 절에서는 다음에 제시한 가상의 요약 대상 논문 및 인용문을 사용하여 전술한 텍스트 유사도 기반 논문 요약 방법의 동작 과정을 예시한다. 요약 대상 논문의 저자 및 출판년도는 각각 James Brown과 2016년으로 가정한다. 또한 아래 예에서는 설명의 편의상 요약 대상 논문이 3개 문장으로만 구성되어 있으며 초록은 누락되어 있다고 가정한다.

● 요약 대상 논문 내 문장

s₁: 이 연구에서는 단어 임베딩에 기반한 문서 색인 및 검색 방법을 제시한다.
s₂: 임베딩 벡터 생성을 위해 Word2Vec 임베딩 방법을 적용하였다.
s₃: 제안된 방법은 TREC 데이터셋을 사용한 평가에서 언어 모델 기반 방법과 대등한 성능을 보였다.

● 인용문

c₁: 최근 문서 색인 및 검색을 위한 임베딩 기반 방법(Brown, 2016)이 시도되었다.
c₂: 단어 임베딩은 문서 색인 및 검색(Brown, 2016), 분류(Williams, 2015) 등 다양한 분야에 적용되고 있다.

다음은 위 예시 논문 내 각 문장, 인용문, $DocumentText$, $AllCitationText$ 를 명사 중심의 용어 기반 표현으로 변환한 결과를 보인 것이다.

● 요약 대상 논문 내 문장의 용어 기반 표현

s₁: 연구, 단어, 임베딩, 기반, 문서, 색인, 검색, 방법, 제시
s₂: 임베딩, 벡터, 생성, Word2Vec, 임베딩, 방법, 적용
s₃: 제안, 방법, TREC, 데이터셋, 사용, 평가, 언어, 모델, 기반, 방법, 대등, 성능
DocumentText: 연구, 단어, 임베딩, 기반, 문서, 색인, 검색, 방법, 제시, 임베딩, 벡터, 생성, Word2Vec, 임베딩, 방법, 적용, 제안, 방법, TREC, 데이터셋, 사용, 평가, 언어, 모델, 기반, 방법, 대등, 성능

● 인용문의 용어 기반 표현

c₁: 최근, 문서, 색인, 검색, 임베딩, 기반, 방법, Brown, 2016, 시도
c₂: 단어, 임베딩, 문서, 색인, 검색, Brown, 2016, 분류, Williams, 2015, 분야, 적용
AllCitationText: 최근, 문서, 색인, 검색, 임베딩, 기반, 방법, Brown, 2016, 시도, 단어, 임베딩, 문서, 색인, 검색, Brown, 2016, 분류, Williams, 2015, 분야, 적용

텍스트 유사도 함수 $sim()$ 에 대해 식 10의 Jaccard 유사도를 사용하는 경우, 문장 s_1 에 대한 $Salienc(s_1)$ 계산과 관련된 텍스트 유사도들은 다음과 같이 계산된다.

$$\text{sim}(\text{AllCitationText}, s_1) = \frac{7}{18}, \text{sim}(\text{AllCitationText}, s_2) = \frac{3}{19}$$

$$\text{sim}(\text{AllCitationText}, s_3) = \frac{2}{25}, \text{sim}(\text{DocumentText}, s_1) = \frac{9}{22}$$

$$\text{sim}(\text{DocumentText}, s_2) = \frac{6}{22}, \text{sim}(\text{DocumentText}, s_3) = \frac{11}{22}$$

$$\text{sim}(c_1, s_1) = \frac{6}{13}, \text{sim}(c_1, s_2) = \frac{2}{14}, \text{sim}(c_1, s_3) = \frac{2}{19}$$

$$\text{sim}(c_2, s_1) = \frac{5}{16}, \text{sim}(c_2, s_2) = \frac{2}{16}, \text{sim}(c_2, s_3) = \frac{0}{23}$$

위 수식에서 $\text{sim}(\text{AllCitationText}, s_2)$ 의 분자 3은 집합 $\text{set}(\text{AllCitationText}) \cap \text{set}(s_2) = \{\text{임베딩, 방법, 적용}\}$ 의 크기이며 분모 19는 집합 $\text{set}(\text{AllCitationText}) \cup \text{set}(s_2)$ 의 크기이다. 위의 유사도 값들로부터 $\text{Saliency}_{\text{cit}}(s_1)$, $\text{Saliency}_{\text{CIT}}(s_1)$, $\text{Saliency}_{\text{DOC}}(s_1)$ 및 $\text{Saliency}(s_1)$ 의 계산 과정은 다음과 같다. 아래 계산에서는 식 1에서 $\lambda_{\text{abs}} = \lambda_{\text{ABS}} = \lambda_{\text{grp}} = \lambda_{\text{seg}} = \lambda_{\text{Adj}} = 0$, $\lambda_{\text{cit}} = \lambda_{\text{CIT}} = \lambda_{\text{DOC}} = 1$ 로 설정한 것이다.

$$\text{Saliency}_{\text{CIT}}(s_1) = \frac{\frac{7}{18}}{\frac{7}{18} + \frac{3}{19} + \frac{2}{25}}$$

$$\text{Saliency}_{\text{cit}}(s_1) = \frac{\frac{6}{13} + \frac{5}{16}}{\frac{6}{13} + \frac{5}{16} + \frac{2}{14} + \frac{2}{16} + \frac{2}{19} + \frac{0}{23}}$$

$$\text{Saliency}_{\text{DOC}}(s_1) = \frac{\frac{9}{22}}{\frac{9}{22} + \frac{6}{22} + \frac{11}{22}}$$

$$\text{Saliency}(s_1) = \text{Saliency}_{\text{cit}}(s_1) + \text{Saliency}_{\text{CIT}}(s_1) + \text{Saliency}_{\text{DOC}}(s_1)$$

IV. Experiments

1. Experimental Setup

제안된 방법의 평가를 위해 CL-SciSumm 데이터셋을 사용한다. CL-SciSumm 데이터셋은 전산언어학(Computational Linguistics) 분야 영어 학술논문들에 대한 요약 평가 집합이다. 실험에서는 CL-SciSumm Shared Task의 2016년, 2017년 데이터셋을 사용하였고, SciSumm 요약 성능 평가에 사용되는 community summary와 human summary 각각에 대해 요약 성능을 평가하였다. Community summary는 요약 대상 논문을 인용하는 각 인용문에 대응하는 피인용텍스트들을 수작업 인식하여 모아 둔 것이다. SciSumm-2016 테스트 데이터셋의 경우 총 10편의 요약 대상 논문 및 총 221편의 인용 논문들의 원문 텍스트가 xml 형식으로 제공된다. 또한 전체 인용 논문 집합 내 총 350개 각 인용텍스트에 대응하는 요약 대상 논문 내 피

인용텍스트들이 수작업 인식되어 있다. SciSumm-2017 테스트 데이터셋은 요약 대상 논문 10편, 인용 논문 88편과 인용텍스트 159개로 구성되어 있다.

요약 성능에 대한 평가 지표로 ROUGE-2 F1을 사용하였다. ROUGE-2는 단일 문서 요약에 대한 자동 평가 지표로 적합하다고 알려져 있으며[25], SciSumm Shared Task의 요약 성능 평가를 위한 공식 평가 지표에 포함된다.

요약 대상 논문 및 인용 논문 텍스트의 문장 단위 분할을 위해 SciSumm 데이터셋 내 기인식된 문장 단위를 사용하였다. 식 1의 요약적합도 계산은, 요약 대상 논문의 제목과 초록을 제외한 문장 길이 10이상 50이하 각 문장에 대해 적용하였다. 문장 내 출현 용어 인식을 위해, 먼저 문장을 소문자 변환한 후 영문자(a-z), 숫자(0-9), 대쉬(-) 문자들의 연속된 나열로 구성된 길이 2이상의 불용어가 아닌 문자열을 용어로 추출하였다. 불용어 제거를 위해 ROUGE 패키지에 포함된 SMART IR 시스템의 598개 불용어 목록을 사용하였다.

식 3의 개별 인용문 citation_i 의 표현은 Mei 등[19]의 연구를 따라 SciSumm 데이터셋 내 인용문에 대해 해당 인용문과 이전 이후 2개 문장을 결합한 문장들을 사용하였다. 식 7의 군집화를 위해 $k=5$ 로 설정하여 K-means 군집법을 적용하였으며, 식 8에서는 요약 대상 논문의 본문 텍스트를 4등분하였다. SciSumm 요약 평가에서는 자동 생성 요약문의 길이 제약으로 250 단어를 설정하고 있으므로, 이 연구의 실험에서도 자동 요약시스템이 250 단어를 초과하지 않는 요약텍스트를 생성하도록 하였다.

2. Experimental Results

표 1은 제안된 유사도 기반 요약 방법을 기존 연구에서 보고된 성능들과 비교 제시한 것이다. 이를 위해 2016 및 2017년 CL-SciSumm Shared Task 참가팀들의 최고 성능[26], impact 기반 요약법, CNN 기반 요약법과 GCN 기반 요약법의 성능들을 사용하였다. Impact 기반 요약법[19]은 문장 언어모델과 인용 언어모델 간 차이로부터 요약문을 선택하는 방법이고, CNN 기반 요약법[6]은 임베딩과 다양한 문맥자질을 활용하는 합성곱 신경망 기반 요약법이며, GCN 기반 요약법[7]은 요약문 후보로 초록 텍스트와 피인용 텍스트를 동시 고려하는 그래프 합성곱 신경망 기반 요약법이다. 기존 성능들은 impact 기반 요약법을 제외하면 해당 논문에서 발췌한 수치들이다. Impact 기반 요약법은 직접 구현하여 성능을 구하였다.

표 1에서 제안된 방법은 식 1의 가중치 λ_i 를 모두 1로 설정한 요약적합도를 적용한 것이며, 그 성능은 초록 텍스트를 활용한 경우와 그렇지 않은 경우로 나누어 제시하였다. 전자(w/abstract)는 식 1의 8가지 모든 요약적합도들을 선형 결합한 것이며, 후자(w/o abstract)는 초록 텍스트와 관련된 abs, ABS의 두 유형을 제외한 나머지 요약적합도들을 결합한 것에 해당한다. 표 1의 성능 수치는 ROUGE 패키지를 실행하여 얻은 값이다.

실험 결과 human summary 평가에서 제안된 방법은 초록을 활용한 경우 2016년 데이터셋에서는 기존 최고 성능에 미치지 못하였으나, 2017년 데이터셋에서 기존 방법들보다 높은 성능을 보였으며, 초록 활용을 배제한 경우에는 두드러진 성능 저하를 보였다. Community summary 평가에서 제안된 방법은 초록 활용 유무에 무관하게 SciSumm-2016, SciSumm-2017년 최고 시스템들의 성능을 능가하였으며, 초록이 사용된 경우 기존 최고 방법들보다 높은 성능을 보였다. Impact 기반 방법과 비교할 때, 제안된 방법은 2017년 community summary 평가셋의 경우를 제외하면 초록 활용 유무에 무관하게 대등하거나 높은 성능을 보였다.

이러한 결과는 제안된 방법이 human summary에 비해 community summary에서 상대적으로 적합한 요약문을 추출하며, community summary를 통해 평가되는 인용 내용 중심 논문 요약에서 ROUGE-2 F1 기준으로 기존 방법들보다 더 적절한 요약텍스트를 생성함을 의미한다. 이는 2016년 human summary 평가를 제외하면 식 2~9의 8개 요약적합도들의 결합을 통한 문장중요도 계산 기법이 SciSumm 방식의 논문 요약에서 효과가 있음을 시사한다. 현재 제안된 방법은 최신 딥러닝 방법론이나 개념 및 의미 수준의 텍스트 유사도에 의존하지 않고 표층 어휘 수준의 텍스트 유사도 자질들만을 사용하여 표 1의 성능 결과를 보였다는 점에서 그 의의가 크다.

Table 1. ROUGE-2 F1 Performances of Different Summarization Methods

Summarization method	Community summary		Human summary	
	Dataset -2016	Dataset -2017	Dataset -2016	Dataset -2017
SciSumm-16 best	0.25	n/a	0.22	n/a
SciSumm-17 best	n/a	0.204	n/a	0.275
Impact-based	0.2438	0.3162	0.1220	0.2134
CNN-based	n/a	0.274	n/a	0.288
GCN-based	n/a	n/a	0.3154	n/a
Proposed method (w/ abstract)	0.2826	0.3556	0.1813	0.2950
Proposed method (w/o abstract)	0.2610	0.2907	0.1293	0.2175

V. Conclusions

이 연구에서는 인용 기반 학술 문헌 요약을 위해 인용 논문 요약에 특화된 문장 요약적합도들의 결합에 기반한 추출식 요약법을 제시하였다. SciSumm-2016/2017 데이터셋을 사용한 community summary 평가에서 제안된 방법은 ROUGE-2 F1 기준으로 기존 방법들보다 높은 요약 성능을 보였으며, human summary 평가에서는 SciSumm-2017 데이터셋에서 기존 최고 성능에 뒤지지 않는 성능을 보였다.

현재 제안된 방법은 문장 요약적합도 계산을 위해 표층 어휘 수준의 단순 텍스트 유사도를 사용하고 있어 향후 개념, 토픽, 의미 수준의 새로운 텍스트 유사도 적용을 통한 성능 향상이 기대된다.

REFERENCES

- [1] V. Qazvinian, and D. Radev, "Scientific Paper Summarization Using Citation Summary Networks," Proceedings of COLING-2008, pp. 689-696, 2008.
- [2] A. Abu-Jbara, and D. Radev, "Coherent Citation-Based Summarization of Scientific Papers," Proceedings of ACL-2011, pp. 500-509, 2011.
- [3] K. Jaidka, M. Chandrasekaran, S. Rustagi, and M.-Y. Kan, "Overview of the CL-SciSumm 2016 Shared Task," Proceedings of BIRNDL-2016, pp. 93-102, 2016.
- [4] K. Jaidka, M. Chandrasekaran, D. Jain, and M.-Y. Kan, "The CL-SciSumm Shared Task 2017: Results and Key Insights," Proceedings of BIRNDL-2017, pp. 1-15, 2017.
- [5] K. Jaidka, M. Yasunaga, M. Chandrasekaran, D. Radev, and M.-Y. Kan, "The CL-SciSumm Shared Task 2018: Results and Key Insights," Proceedings of BIRNDL-2018, pp. 74-83, 2018.
- [6] A. Abura'ed, A. Bravo, L. Chiruzzo, and H. Saggion, "LaSTUS/TALN+ INCO @ CL-SciSumm 2018 - Using Regression and Convolutions for Cross-document Semantic Linking and Summarization of Scholarly Literature," Proceedings of BIRNDL-2018, pp. 150-163, 2018.
- [7] M. Yasunaga, J. Kasai, R. Zhang, A. Fabbri, I. Li, D. Friedman, and D. Radev, "ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks," Proceedings of AAAI-2019, 2019.
- [8] H. Saggion, A. AbuRa'ed, and F. Ronzano, "Trainable Citation-enhanced Summarization of Scientific Articles," Proceedings of BIRNDL-2016, pp. 175-186, 2016.
- [9] L. Li, L. Mao, Y. Zhang, J. Chi, T. Huang, X. Cong, and H. Peng, "CIST System for CL-SciSumm 2016 Shared Task," Proceedings of BIRNDL-2016, pp. 156-167, 2016.
- [10] L. Li, Y. Zhang, L. Mao, J. Chi, M. Chen, and Z. Huang, "CIST@CLSciSumm-17: Multiple Features Based Citation Linkage, Classification and Summarization," Proceedings of BIRNDL-2017, pp. 43-54, 2017.
- [11] A. Abura'ed, L. Chiruzzo, H. Saggion, P. Accuosto, and A. Bravo, "LaSTUS/TALN @ CLSciSumm-17: Cross-document Sentence Matching and Scientific Text

- Summarization Systems," Proceedings of BIRNDL-2017, pp. 55-66, 2017.
- [12] S. Ma, H. Zhang, J. Xu, and C. Zhang, "NJUST @ CLSciSumm-18," Proceedings of BIRNDL-2018, pp. 114-129, 2018.
- [13] L. Li, J. Chi, M. Chen, Z. Huang, Y. Zhu, and X. Fu, "CIST@CLSciSumm-18: Methods for Computational Linguistics Scientific Citation Linkage, Facet Classification and Summarization," Proceedings of BIRNDL-2018, pp. 84-95, 2018.
- [14] Z. Cao, W. Li, and D. Wu, "PolyU at CL-SciSumm 2016," Proceedings of BIRNDL-2016, pp. 132-138, 2016.
- [15] A. Lauscher, G. Glavas, and K. Eckert, "University of Mannheim @ CLSciSumm-17: Citation-Based Summarization of Scientific Articles Using Semantic Textual Similarity," Proceedings of BIRNDL-2017, pp. 33-42, 2017.
- [16] A. Cohan, and N. Goharian, "Scientific Document Summarization via Citation Contextualization and Scientific Discourse," International Journal on Digital Libraries, Vol. 19, No. 2-3, pp. 287-303, 2018.
- [17] S. Ma, J. Xu, J. Wang, and C. Zhang, "NJUST @ CLSciSumm-17," Proceedings of BIRNDL-2017, pp. 16-25, 2017.
- [18] B. Malenfant, and G. Lapalme, "RALI System Description for CL-SciSumm 2016 Shared Task," Proceedings of BIRNDL-2016, pp. 146-155, 2016.
- [19] Q. Mei, and C. Zhai, "Generating Impact-Based Summaries for Scientific Literature," Proceedings of ACL-2008, pp. 816-824, 2008.
- [20] J. Conroy, and S. Davis, "Section Mixture Models for Scientific Document Summarization," International Journal on Digital Libraries, Vol. 19, No. 2-3, pp. 305-322, 2018.
- [21] D. Debnath, A. Achom, and P. Pakray, "NLP-NITMZ @ CLSciSumm-18," Proceedings of BIRNDL-2018, pp. 164-171, 2018.
- [22] L. Moraes, S. Baki, R. Verma, and D. Lee, "University of Houston at CL-SciSumm 2016: SVMs with tree kernels and Sentence Similarity," Proceedings of BIRNDL-2016, pp. 113-121, 2016.
- [23] R. Day, "The Origins of the Scientific Paper: The IMRAD Format," American Medical Writers Association Journal, Vol. 4, No. 2, pp. 16-18, 1989.
- [24] P. Jaccard, "Nouvelles recherches sur la distribution florale," Bull. Soc. Vaud. Sci. Nat., Vol. 44, pp. 223-270, 1908.
- [25] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," Proceedings of Workshop on Text Summarization Branches Out, pp. 74-81, 2004.
- [26] P. Mayr, M. Chandrasekaran, and K. Jaidka, "Report on the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2018)," SIGIR Forum 52(2), pp. 105-110, 2018.

Authors



In-Su Kang received his bachelor's degree from Kyungpook National University in 1995, and master's and doctoral degrees from POSTECH, in 1999, and 2006, respectively. He is currently an associate professor in the Department of Computer

Science, Kyungsoo University. He is interested in natural language processing and information retrieval.