KIPS Trans. Softw. and Data Eng.
Vol.9, No.8 pp.251~258 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2020.9.8.251

Building a Korean Text Summarization Dataset Using News Articles of Social Media

Gyoung Ho Lee† · Yo-Han Park†† · Kong Joo Lee†††

ABSTRACT

A training dataset for text summarization consists of pairs of a document and its summary. As conventional approaches to building text summarization dataset are human labor intensive, it is not easy to construct large datasets for text summarization. A collection of news articles is one of the most popular resources for text summarization because it is easily accessible, large-scale and high-quality text. From social media news services, we can collect not only headlines and subheads of news articles but also summary descriptions that human editors write about the news articles. Approximately 425,000 pairs of news articles and their summaries are collected from social media. We implemented an automatic extractive summarizer and trained it on the dataset. The performance of the summarizer is compared with unsupervised models. The summarizer achieved better results than unsupervised models in terms of ROUGE score.

Keywords : Korean Text Summarization Dataset, Description, Headline, Subhead, Automatic Extractive Summarization

신문기사와 소셜 미디어를 활용한 한국어 문서요약 데이터 구축

이 경 호† · 박 요 한†† · 이 공 주†††

요 약

문서 요약을 위한 학습 데이터는 문서와 그 요약으로 구성된다. 기존의 문서 요약 데이터는 사람이 수동으로 요약을 작성하였기 때문에 대량의 데이터 확보가 어려웠다. 그렇기 때문에 온라인으로 쉽게 수집 가능하며 문서의 품질이 우수한 인터넷 신문기사가 문서 요약 연구에 많이 활용되어 왔다. 본 연구에서는 언론사가 소셜 미디어에 게시한 설명글과 제목, 부제를 본문의 요약으로 사용하여 한국어 문서 요약 데이터를 구성하는 것을 제안한다. 약 425,000개의 신문기사와 그 요약데이터를 구축할 수 있었다. 구성한 데이터의 유용성을 보이기 위해 추출 요약 시스템을 구현하였다. 본 연구에서 구축한 데이터로 학습한 교사 학습 모델과 비교사 학습 모델의 성능을 비교하였다. 실험 결과 제안한 데이터로 학습한 모델이 비교사 학습 알고리즘에 비해 더 높은 ROUGE 점수를 보였다.

키워드 : 한국어 문서 요약 데이터 집합, 설명글, 제목, 부제, 자동 추출 문서 요약

1. 서 론

자동 문서 요약(automatic text summarization)은 자연언어 처리 분야에서 많은 관심을 받아온 응용 분야이다. 자동 문서 요약은 사람이나 시스템이 다양한 텍스트 데이터를 효과적으로 처리할 수 있도록 도움을 줄 수 있다.

자동 문서 요약 시스템 개발을 위해 문서와 그 문서에 대한 요약이 필요하다. 문서 또는 문서 군집을 여러 사람이 읽고 주관에 따라 다양한 형식과 길이로 만든 요약이 있다면, 자동 문서 요약 연구를 위한 활용도 높은 데이터가 될 수 있다. 문서 요약 연구 초창기에 개발된 DUC 데이터[1]의 경우 한 개의 문서 또는 문서 집합에 대해 3~4명의 평가자가 문서를 읽고 미리 정해진 길이와 목적에 따른 요약을 작성하여 이를 활용하도록 하였다. 하지만 이러한 방식의 문서 요약 데이터 구축은 많은 비용과 시간이 필요하기 때문에 대량의 데이터를 확보하기 어렵다.

인터넷 신문기사는 온라인으로 쉽게 수집할 수 있고 문서의 품질이 우수하여 문서 요약 연구에 많이 활용되어왔다. 이러한 연구들은 주로 영어를 대상으로 이루어졌으며[2-6], 최근 일본어[7], 중국어[8], 체코어[9] 등 다양한 언어에서도 신문기사를 기반으로 한 문서 요약 연구들이 진행되고 있다. 본 연구에서는 기존의 문서 요약 데이터들과 같이, 한국어로 작성된 인터넷 신문기사를 기반으로 한국어 문서 요약 데이터

※ 이 논문은 2019년 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2019R1F1A1053136).
† 준 회 원 : 드라마앤컴퍼니 연구원
†† 비 회 원 : 충남대학교 전파정보통신공학과 학사과정
††† 종신회원 : 충남대학교 전파정보통신공학과 교수
Manuscript Received : May 18, 2020
Accepted : June 10, 2020
* Corresponding Author : Kong Joo Lee(kjoolee@cnu.ac.kr)

를 구축하는 방안에 대해 제안한다.

신문기사와 관련된 다양한 텍스트를 기사의 요약으로 활용할 수 있다. 본 연구에서는 소셜 미디어에서 기사에 대해 설명하는 글을 수집하여 기사 본문에 대한 요약으로 사용하는 것을 제안한다. 최근 소셜 미디어의 발달로 여러 언론사들이 자신들의 기사를 소셜 미디어를 통해 유통하고 있다. 이 과정에서 언론사들은 소셜 미디어 사용자의 흥미를 끌기 위하여 기사에 대한 간략한 정보를 담은 글을 기사와 함께 등록한다. 본 연구에서는 이러한 기사 본문에 대한 설명글을 기사 본문에 대한 요약으로 간주하여 사용하였다. 또한 제목(headline)과 부제(subhead)를 기사에 대한 포괄적 정보를 제공하는 요약으로 간주하여 설명글과 함께 기사 본문에 대한 요약으로 사용하여 한국어 문서 요약 데이터를 구성하는 것을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구과 관련된 연구에 대하여 소개하였다. 3장에서는 본 연구에서 제안하는 제목 요약과 설명 요약의 유형에 대해 구체적으로 살펴보고 4장에 데이터의 수집 방법에 대해 나타내었다. 5장에서는 수집된 데이터의 특성에 대해 분석하였고 6장에서는 수집한 데이터의 유효성을 검증하였다.

2. 관련 연구

딥러닝 기반의 문서 요약 연구가 일반화되면서 대량의 문서 요약 데이터의 필요성이 커지고 있다. 영어권 연구에서는 언어모델이나 정보검색에서 활용하던 Gigaword 코퍼스[2], 질의응답 연구용으로 개발된 CNN/DailyMail 코퍼스[10] 등 기존의 다른 분야에서 사용하던 대량의 코퍼스를 문서 요약 데이터로 활용하였다. 또한 New York times Annotated 코퍼스[6], Newsroom 코퍼스[5] 등과 같이 문서 요약을 목적으로 코퍼스가 구축되기도 하였다. 본 연구에서는 CNN/DailyMail 코퍼스와 Newsroom 코퍼스의 특징과 유사한 특징을 가지는 한국어 문서 요약 코퍼스 구축을 시도하였다.

영어 이외에도 여러 언어에 대한 문서 요약 코퍼스 구축이 이루어져 왔다. [7]의 연구에서는 추상 문서 요약(Abstractive summarization)을 위하여 일본어 인터넷 신문 사이트인 livedoor.com의 신문기사와 요약을 수집하였다. 이 사이트에서는 신문기사의 요지를 3문장으로 작성한 요약을 제공한다. [9]의 연구에서는 체코어에 대한 문서 요약 코퍼스를 구축하였다. 코퍼스 구축을 위해 체코어로 기사를 작성하는 5개의 언론사 사이트로부터 기사의 제목, 초록(abstract), 본문을 수집하였다. 이 연구에서 요약으로 사용한 텍스트는 신문기사의 전문(leading sentence)을 이용한 것으로 보인다. 본 연구에서는 기사의 전문 외에, 제목과 부제, 그리고 소셜 미디어의 설명글을 요약으로 활용하였다.

한국어 문서 요약 연구에도 다양한 형태의 문서 요약 데이터가 사용되었다. [11]의 연구에서는 한국어 위키피디아 페이지를 이용하여 문서 요약 모델을 평가하였다. [12]의 연구에서는 문서 요약 모델 연구를 위하여 기자가 작성한 요약이 있는 경우 요약을 사용하고 없을 경우 전문을 요약으로 사용하였다. [13]의 연구도 이와 유사하게 기사의 나머지 부분으로부터 첫 문단에 위치한 전문을 생성하는 연구를 진행하였다.

한국어 추상 요약을 목적으로 구축된 [12]과 [13]의 데이터 및 [9]의 체코어 문서 요약 데이터는 기사의 전문 역할을 하는 텍스트를 요약으로 간주하고 문서 요약에 사용하였다. 본 연구에서는 신문기사에서 제목과 더불어 부제가 기사 내용을 대표하는 요약으로 가치가 있다고 판단하고 이를 요약에 활용하였다. 또한 소셜 미디어에서 뉴스 유통을 위한 언론사 공식 계정의 신문기사 설명글을 기사의 내용을 대표하는 짧은 글이라 판단하여 요약으로 활용하였다.

3. 한국어 문서 요약 구성

기사 본문에 대해 작성된 글은 그 본문의 요약으로 활용할 수 있는 자원이다. 영어 문서 요약 연구를 위한 CNN/DailyMail 코퍼스와 뉴스룸 코퍼스의 구축 철학도 이에 기반한다. 본 연구에서도 이들과 유사한 형식의 데이터를 수집하고 한국어 문서 요약 연구에 활용하였다. 본 연구에서는 한국어 문서와 그 문서의 요약 쌍을 대량으로 수집하기 위하여 한국어로 작성된 인터넷 신문기사의 소셜미디어에 등록된 기사에 대한 설명글과 제목, 부제를 수집하여 요약으로 활용하였다.

3.1 설명글(description)[1)]

최근 소셜 미디어가 뉴스 기사의 주요 유통 채널로 자리매김하고 있다. 언론들은 자신의 기사가 소셜 미디어에 공유될 때, 기사 링크에 들어가지 않아도 소셜 미디어 상에서 기사 내용을 파악할 수 있도록 기사에 대한 짧은 설명을 제공하고 있다(Fig. 1). 이러한 글은 주로 언론사 시스템이나 소셜 미디어 담당자의 주관에 따라 기사 제목, 본문의 첫 문장이나 주요 문

Fig. 1. Example of Description[1)]

1) https://twitter.com/hanitweet/status/1158241218830188544

장, 또는 기사 내용에 기반한 새로운 글 등 다양한 형식으로 작성된다. [14]의 연구에서는 기사 웹 페이지에 포함된 기사의 설명을 요약으로 간주하여 뉴스룸 코퍼스 구축에 사용하였다. 본 연구에서도 [14]의 연구와 같이 소셜 미디어 사용자를 위해 제공되는 기사 설명 정보가 한국어 기사에 대한 요약으로 가치가 있다고 판단하고 이를 수집하여 요약으로 사용하였다.

3.2 제목(headline)

일반적으로 신문기사는 독자의 관심을 유발하고 기사 내용을 효과적으로 전달하기 위해 관습적으로 제목(부제 포함), 전문, 본문의 3가지 구조를 가진다[15]. 제목은 1)기사가 제공하는 정보를 독자에게 빠르고 인상적으로 알리기 위한 광고 및 색인 기능 2)뉴스의 비중을 독자에게 간접적으로 알리는 가치 판단 기능 3)기사의 내용을 압축적으로 전달하기 위한 압축 전달 기능 4)신문의 시각적 효과를 위한 지면의 미적 균형 기능을 가진다. 제목이 가지는 광고 및 색인 기능과 압축 전달 기능, 가치 판단의 기능은 문서의 요약이 가져야하는 중요한 기능이다[15]. 또한 [15]의 연구에서 한국어 신문기사의 제목 유형 중 정보 전달형과 관심 유도형의 비율이 86.3%와 13.7%로 조사되었다. 이러한 한국어 기사 제목에 관한 연구를 근거로, 한국어 신문기사의 제목을 기사의 요약으로 활용한다.

3.3 부제(subhead)

영어 문서 요약 연구에서 많이 사용되는 CNN/DailyMail 코퍼스는 미국 CNN과 영국 DailyMail의 인터넷 신문기사를 이용한 데이터이다. 이들 기사에는 기자나 에디터가 작성한 기사의 요지를 담은 스토리 하이라이트(story highlights)를 포함하고 있다(Fig. 2A). 스토리 하이라이트는 기사의 개요를 담은 3~4줄의 문장으로 구성된다[16]. 작성자에 따라 기사 본문의 문장을 발췌하거나 약간의 변화, 또는 완전 새로운 문장으로 구성하는 경우가 있어 문서 요약 연구에서 다양하게 활용되고 있다. 현재 몇몇 한국어 신문기사 사이트에서는 스토리 하이라이트의 역할을 하는 부제를 제공하고 있다(Fig. 2B). 본 연구에서는 이러한 부제를 수집하여 요약으로 사용한다.

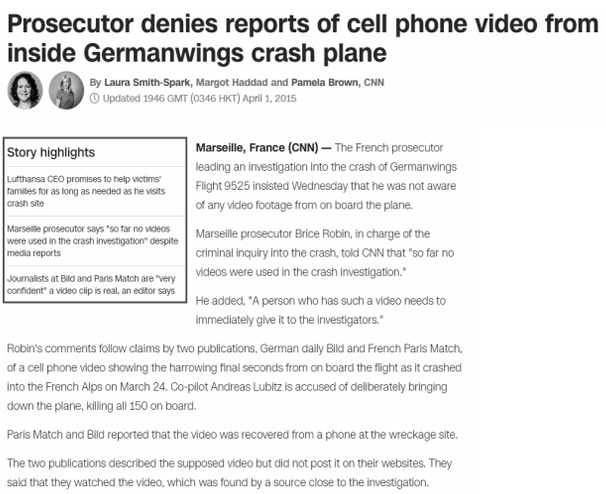
제목은 대부분 기사와 함께 제공되지만 요약으로 활용하기에는 길이가 짧고 함축적이다. 이를 보강하기 위해 부제를 함께 사용하지만 기사에 따라 부제가 없는 경우도 있다. 본 연구에서는 소셜 미디어에 등록된 기사의 설명글과 제목, 부제를 함께 해당 기사의 요약으로 사용한다.

4. 기사-요약 수집 방법

본 연구에서 제안하는 요약의 수집 방법은 다음과 같다.

4.1 언론사 목록 수집

많은 언론사들이 한국어로 작성된 기사를 인터넷으로 배포하고 있다. 이들로부터 일정 수준의 기사와 요약을 수집하기 위해 언론사를 선별하였다. 언론사의 주요 기사 유형, 소셜


Prosecutor denies reports of cell phone video from inside Germanwings crash plane

By Laura Smith-Spark, Margot Haddad and Pamela Brown, CNN

Updated 1946 GMT (0346 HKT) April 1, 2015

Story highlights

Lufthansa CEO promises to help victims' families for as long as needed as he visits crash site

Marseille prosecutor says "so far no videos were used in the crash investigation" despite media reports

Journalists at Bild and Paris Match are "very confident" a video clip is real, an editor says

Marseille, France (CNN) — The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane.

Marseille prosecutor Brice Robin, in charge of the criminal inquiry into the crash, told CNN that "so far no videos were used in the crash investigation."

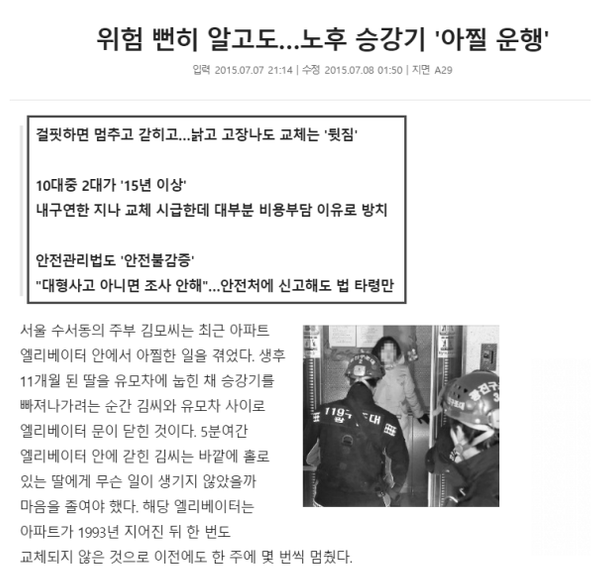
He added, "A person who has such a video needs to immediately give it to the investigators."

Robin's comments follow claims by two publications, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board the flight as it crashed into the French Alps on March 24. Co-pilot Andreas Lubitz is accused of deliberately bringing down the plane, killing all 150 on board.

Paris Match and Bild reported that the video was recovered from a phone at the wreckage site.

The two publications described the supposed video but did not post it on their websites. They said that they watched the video, which was found by a source close to the investigation.

Fig. 2A. Example of CNN Story Highlights[2)]


위험 뻔히 알고도...노후 승강기 '아찔 운행'

입력 2015.07.07 21:14 | 수정 2015.07.08 01:50 | 지면 A29

걸핏하면 멈추고 갇히고...낡고 고장나도 교체는 '뒷짐'

10대중 2대가 '15년 이상'
내구연한 지나 교체 시급한데 대부분 비용부담 이유로 방치

안전관리법도 '안전불감증'
"대형사고 아니면 조사 안해"...안전처에 신고해도 법 타령만

서울 수서동의 주부 김모씨는 최근 아파트 엘리베이터 안에서 아찔한 일을 겪었다. 생후 11개월 된 딸을 유모차에 눕힌 채 승강기를 빠져나가려는 순간 김씨와 유모차 사이로 엘리베이터 문이 닫힌 것이다. 5분여간 엘리베이터 안에 갇힌 김씨는 바깥에 홀로 있는 딸에게 무슨 일이 생기지 않았을까 마음을 졸여야 했다. 해당 엘리베이터는 아파트가 1993년 지어진 뒤 한 번도 교체되지 않은 것으로 이전에도 한 주에 몇 번씩 멈췄다.

Fig. 2B. Example of Subhead[3)]

미디어 기사 제공 여부, 데이터 수집의 난이도 등을 고려하여 총 25개 한국어 언론사를 선정하여 이들이 등록한 데이터를 수집하였다.[2)3)]

4.2 기사 URL 수집

기사 본문 주소(URL)와 설명글은 소셜 미디어를 기반으로 수집한다. 언론사가 소셜 미디어에 작성한 글을 수집하기 위하여 앞서 선정한 언론사들의 트위터[4)] 공식 계정을 수집하고 이 계정이 작성한 글을 수집하였다. 이를 통해 약 770,000개 기사에 대한 기사 설명글과 기사 본문 URL을 수집하였다. 영어 문서 요약 연구 데이터인 뉴스룸 코퍼스는 html 문서의

2) https://edition.cnn.com/2015/04/01/europe/france-germanwings-plane-crash-main/
3) https://www.hankyung.com/society/article/2015070749801
4) http://www.twitter.com

Table 1. Examples of Preprocessing Description

1)	Before Preprocessing	경찰, '불법업소 논란' 대성 건물 압수수색…관련자료 확보 #대성 #강남 #빅뱅 #성매매http://omn.kr/1ka7x
	After Preprocessing	경찰, '불법업소 논란' 대성 건물 압수수색 관련자료 확보
2)	Before Preprocessing	MBC 기자가 최근 회사에 사표를 내고 외교부 대변인실로 이직했다. 7일부터 외교부 대변인실 정책홍보담당관으로 근무한다. 보도국 구성원들은 이직을 만류했으나 끝내 마음을 되돌리진 못했다. http://www.mediatoday.co.kr/news/articleView.html?idxno=201560 …
	After Preprocessing	MBC 기자가 최근 회사에 사표를 내고 외교부 대변인실로 이직했다. 7일부터 외교부 대변인실 정책홍보담당관으로 근무한다. 보도국 구성원들은 이직을 만류했으나 끝내 마음을 되돌리진 못했다.

Table 2. Example of Summaries and Bodies

1	Headline	"이혼위기 극복" '인생술집' 홍지민이 밝힌, '다섯가지 ♡의 언어'
	Subhead	
	Description	'이혼 위기' 맞이한 홍지민과 남편이 이를 극복한 방법 5가지
	Body	홍지민이 전한 '다섯가지 사랑의 언어'가 시청자들이게 깊은 여운을 남겼다. 30일 방송된 tvN 예능 '인생술집'에서 뜻밖의 절친특집으로 홍지민, 소이현, 정애연이 출연했다. 세 사람은 드라마 '비포 &에프터 성형외과'에서 친해지게 됐다고 했다. 모이면 그릇 브레이커라고. …
2	Headline	MBC 기자, 외교부 대변인실로 이직
	Subhead	대변인실 정책홍보담당관 개방형 공모 합격
	Desciptioin	MBC 기자가 최근 회사에 사표를 내고 외교부 대변인실로 이직했다. 7일부터 외교부 대변인실 정책홍보담당관으로 근무한다. 보도국 구성원들은 이직을 만류했으나 끝내 마음을 되돌리진 못했다.
	Body	MBC 기자가 최근 회사에 사표를 내고 외교부 대변인실로 이직했다. A 기자는 7일부터 외교부 대변인실 정책홍보담당관으로 근무한다. 개방형 직위로 지정된 외교부 정책홍보담당관 공모에 지원해 합격했다. …

모 기술을 위해 사용되는 메타 태그 정보를 기사의 요약으로 간주하여 문서 요약 데이터를 수집하였다. 국내 언론사들의 소셜 미디어 작성 글을 관찰한 결과, 국내 언론사들은 이러한 메타 태그의 활용보다 소셜 미디어에 정보를 직접 입력하는 경향이 있는 것을 확인하였다. 그렇기 때문에 메타 태그 대신, 각 언론사의 공식 소셜 미디어 계정이 작성한 기사 URL과 설명글을 수집하였다.

이러한 방식은 신문기사 URL 목록 수집에도 이점이 있다. 한국어 웹 페이지의 경우, 뉴스룸 코퍼스 수집에 사용된 Archive.org와 같은 웹 페이지 아카이빙 서비스가 부족하다. 그렇기 때문에 각 언론사의 기사 목록을 수집하기 위해서는 개별 언론사의 웹 페이지를 분석하여 기사 목록 페이지나 URL 패턴을 찾아야 한다. 하지만 소셜 미디어를 활용하면 소셜 미디어에 등록된 글에서 직접 기사 URL을 찾아낼 수 있다.

4.3 설명글 수집

수집한 언론사의 트위터 텍스트에는 트위터 사용자들 사이에 통용되는 규칙이나 기사 링크, 이미지 링크 등이 포함되어 있다. 기사 설명글을 요약으로 활용하기 위하여 이러한 내용들을 제거하였다. 또한 설명글의 길이가 짧거나 기사 URL이 없는 글을 제외하였다.

Table 1의 1) 과 2)는 수집된 설명글의 전처리 결과 예이다. 전처리 과정에서 트위터에서 사용되는 특수 태그, URL, 말줄임 기호 등이 제거된다. 1)의 전처리 결과는 함께 첨부된 링크의 신문기사 제목과 동일하다. 이러한 경우 소셜 미디어를 통해 수집한 설명글이 제목 요약과 같다. 2)의 글은 작성자가 원본 기사에서 중요한 문장 몇 개를 선택하고 이 문장에 대한 추가적인 수정을 하여 만든 글이다. 이러한 글은 좋은 설명글의 예이다.

4.4 제목과 부제 수집

앞서 단계를 통해 설명글과 신문기사의 URL을 수집하였다. 수집한 기사 URL 페이지를 분석하여 기사의 제목, 부제, 본문을 추출한다. 수집된 기사 중, 기사의 올바른 페이지를 찾을 수 없는 경우, 기사의 제목이나 내용이 너무 짧은 경우, 동일한 제목 또는 동일한 URL을 가진 경우, 한국어로 작성된 기사가 아닌 경우 등 요약을 추출하기 적합하지 않은 문서를 제거하였다. 또한 기사 본문 중 시작과 끝 위치에 언론사 소개 및 기자 소개, 기타 기사 내용과 관계 없는 부분을 규칙을 통해 제거하였다.

Table 3. Number Documents IN Korean Document Summary Data

	Train	Test	Validation	Total
With Subhead	205,454	25,486	23,037	253,977 (60%)
Without Subhead	138,744	17,008	15,208	170,960 (40%)
Total	344,198	42,494	38,245	**424,937**

Table 4. Korean News Article Information

Average # of Sentences per Article	Average # of Words per Article
28.6	385.5

이렇게 수집된 기사의 제목과 부제를 함께 수집하여 요약으로 사용하였다. Table 2은 수집된 문서 요약 데이터의 예이다.

5. 수집데이터 분석

앞서 설명한 과정을 통해 총 424,937개의 기사를 수집하였다. 수집된 데이터의 10%를 평가 데이터로 남겨두고 남은 데이터를 9:1의 비율로 학습과 검증 데이터로 나누었다[5]. 이 비율에 따라 학습, 평가, 검증 데이터는 각각 344,198개, 42,494개, 38,245개의 문서로 구성된다(Table 3). 이들 문서 중 부제가 있는 문서는 각각 206,454개, 25,486개, 23,037개로 약 60%의 문서가 부제를 포함하고 있다.

학습 데이터를 구성하는 기사의 정보는 Tabel 4와 같다. 수집된 신문기사는 평균 28.6 문장, 385.5어절로 구성된다.

Table 5은 각 요약들의 특성에 대한 분석 결과이다. 특성을 비교하기 위하여 기사의 첫 문장(전문 LEAD)을 함께 분석하였다.

4가지 요약 중, 제목이 평균 7.3어절로 다른 요약들보다 더 짧은 것으로 나타났다. 문장 구성 성분에서는 제목과 부제가 비슷한 특성을 보였고, 다른 요약들과는 다른 특성을 보였다. 설명글과 전문을 구성하는 단어들의 형태소 중 명사류의 비율은 각각 39.1%와 40.1%이다. 반면 제목과 부제의 경우 명사류 비율이 약 45%, 47%로 다른 두 요약보다 명사류의 비율이 더 높다. 조사와 어미의 비율에서도 제목과 부제가 다른 요약과 다른 특성을 보였다. 제목과 부제는 약 8.0%, 8.5%의 조사와 10.2%, 9.6%의 어미로 구성되지만 설명글은 13.6%과 14.4%, 전문은 15.2% 과 14.3%로 제목과 부제가 다른 요약들에 비해 조사와 어미를 덜 포함하고 있는 것이 확인됐다. 이는 기사 제목과 부제가 압축적인 정보 전달을 위해 여러 가지 생략 과정을 거쳐 만들어졌기 때문으로 보인다. 또한 제목과 부제에서 심볼의 사용 비율이 다른 요약에 비해 더 높았다. 제목, 부제에서 본문에 등장하는 인명이나 지명 등을 한자로 대치해서 보여주거나 기사 내용의 일부 내용을 인용부호를 통해 인용하는 등의 이유로 심볼의 비율이 다른 두 요약에 비해 높게 나온 것으로 보인다.

종결형 어미로 끝난 평서문의 비율에서 4가지 요약이 모두 다른 특성을 나타냈다. 전문의 경우 기사의 첫 문장을 가져왔기 때문에 대체적으로 일반적인 평서문의 비율이 높다(약 87%). 반면 제목과 부제는 10.6%, 5.8%로 평서문의 비율이 낮았다. 문장 구성성분 중 명사류 비율과 함께 살펴보면 제목과 부제가 주요 명사형 단어들의 나열식으로 구성된다는 점을 알 수 있다. 설명글의 경우 종결형 어미로 끝나는 문장의 비율이 31.5%로 앞서 다른 두 요약과 다른 양상을 보이고 있다. 이는 설명글이 복합적인 형식으로 구성되어 있는 것에 기인한 것으로 보인다.

Table 5의 ROUGE는 문서 본문과 각 요약간의 ROUGE 점수이다. LEAD의 경우에는 첫 문장을 제외한 본문과 비교하였다. ROUGE-1 정확률 점수는 제목 81.4, 부제 83.8 설

Table 5. Summary Characteristics

		Headline	Subhead	Description	Lead
Average # of Words		7.3	13.4	13.4	13.3
Average # of Characters		23.6	42.6	45.1	46.7
Declarative Sentence Ratio		10.6	5.8	31.5	86.8
ROUGE-1	Precision	81.4	83.8	85.3	78.4
	Recall	1.9	2.7	3.8	4.0
ROUGE-2	Precision	36.3	40.0	52.9	38.5
	Recall	0.8	1.3	2.5	1.8
Ratio of POS tags	Noun/Proper Noun	44.8	47.1	39.1	40.1
	Verb	7.4	7.1	9.5	8.4
	Postposition	8.0	8.5	13.6	15.2
	Verbal Ending	10.2	9.6	14.4	14.3
	Symbol	21.8	18.9	14.3	13.0
	Etc	8.0	8.7	9.1	9.0

5) https://github.com/gyholee/CNUKorSummData

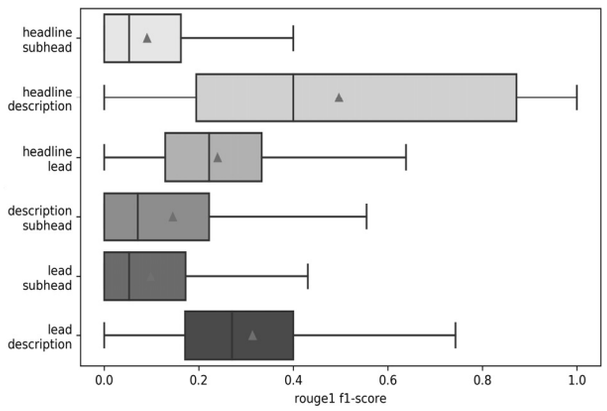


Fig. 3. ROUGE-1 F1-score between Summaries

명글 85.3, 전문 요약 78.4로 모두 높은 수준의 정확률을 나타낸다. 이는 요약에서 사용된 단어들이 높은 비율로 본문에서 사용되었음을 의미한다. 그렇기 때문에 이들 요약은 문서의 재해석을 통해 새로운 표현으로 만들어진 요약보다 본문의 내용과 표현에 기반하여 만들어진 요약에 더 가깝다고 볼 수 있다. ROUGE-2 정확률의 경우 설명글이 다른 요약들보다 더 높은 점수를 나타낸다. ROUGE-2 정확률은 요약의 bi-gram이 본문에 등장한 비율로 볼 수 있다. 이를 기반으로 설명글이 다른 요약들보다 본문에 사용된 문장과 좀 더 유사할 것으로 짐작할 수 있다.

요약의 내용적인 특성은 구성 성분 특성과 다른 성향을 보이고 있다. Fig. 3은 4가지 요약 간의 ROUGE-1 F1 점수의 분포, Table 6은 요약 데이터의 예시이다. 부제는 구성 성분에서는 비슷한 성향을 보였던 제목뿐만 아니라, 설명글, 첫 문장과의 F1 점수가 낮았다. Table 6의 예제 (1)과 (2) 에서와 같이 부제가 제목, 설명글과 다른 내용을 시사하고 있는 것이 여러 차례 확인됐다. 반면 제목과 설명글의 F1 점수에서 50% 이상의 기사가 0.4 이상의 점수를 받았다. 이는 전체의 약 40% 설명글 요약이 제목과 동일하거나 제목을 포함하고 있기 때문이다. 예제 (2)의 설명글은 제목을 포함하고 있다.

6 실험 및 분석

다음 실험의 목적은 본 연구에서 구축한 한국어 문서 요약 데이터의 유용성을 살펴보기 위함이다. 기존에 널리 사용되고 있던 추출 기반 요약 시스템을 이용하여 실험을 진행하였다. 추출 기반 요약 시스템은 문서 내용 중 일부를 추출하여 요약으로 사용한다.

6.1 정답 요약 구축

추출 요약 시스템의 개발 및 평가를 위해서는 문서의 일부 문장을 정답 요약으로 결정해야 한다. 정답 요약의 구축은 다음과 같다. 본문 문장을 순회하면서 제목, 부제, 설명글과의 ROUGE 점수가 가장 높은 문장을 부분 정답 요약으로 선택한다. 다시 본문 문장을 순회하면서 앞서 선택한 부분 정답 요약과 함께 ROUGE 점수를 계산하여 가장 높은 점수의 문장을 선택한다. 이때 선택된 문장을 추가했을 때의 ROUGE 점수가 이전의 부분 정답 요약의 ROUGE 점수보다 높아질 경우 해당 문장을 부분 정답 요약에 추가한다. 이를 반복하다가 ROUGE 점수가 더 높아지지 않을 때 순회를 종료한다 [17]. 각 신문기사당 추출 문장 개수는 평균 2.1개이다.

6.2 평가 방법

정답 요약으로 모델을 학습한 뒤 학습된 모델이 생성한 추출요약과 {제목, 부제, 설명글}과의 ROUGE-score를 계산하

Table 6. Examples of Summary Data

(1)	Headline	여름 블록버스터 끝나면 추석엔 사극 '3파전'
	Subhead	'물괴' 시작으로 추석 대목에 개봉하는 사극 영화들 '크리처'부터 '풍수지리'까지…다채로운 소재 가져와
	Description	'물괴'를 실체화시켜야 하는 문제 때문에 영화 제작이 어려울 것으로 생각했다고. 여기에 사극이라는 장르까지 더해져 '흥행 위험성'이 더욱 높았다는 설명이다.
	Body	여름 블록버스터 대전이 끝나면 추석 사극 '3파전'이 시작된다. 9월 중순 추석을 맞이해 극장가에는 가족관객을 겨냥한 각기 다른 매력의 사극 영화들이 개봉한다. 9월 22일부터 시작되는 5일 간의 연휴 동안 관객들을 찾아 올 추석 사극 영화들을 정리해봤다. ...
(2)	Headline	박원순, 대중교통 무료 논란에 "50억보다 시민 생명 우선"
	Subhead	"남경필 지사는 무엇을 하셨냐" 경기도에 역공…"교통량 감소 두 자릿수대 목표" "내년 서울 개최 100주년 전국체전, 평양서 동시 개최 제안"
	Description	박원순, 대중교통 무료 논란에 "50억보다 시민 생명 우선" "50억원을 선택할 것인가, 시민의 생명을 선택할 것인가"
	Body	박원순 서울시장은 미세먼지 비상저감조치인 '출·퇴근 시간 대중교통 무료' 시행과 관련, "경기도가 참여했다면 그 효과가 훨씬 높았을 것"이라고 17일 밝혔다. 서울시가 지난 15일 처음 시행한 출·퇴근 시간대 대중교통 무료 조치를 '포퓰리즘'이라고 비판한 남경필 경기지사에 대한 박 시장의 반박이다. ...

Table 7. Experimental Results

	ROUGE-1			ROUGE-2			ROUGE-L		
	F1	P	R	F1	P	R	F1	P	R
Text Rank	18.71	11.81	53.79	6.75	4.24	19.80	13.29	8.34	38.93
LEAD - 3	22.76	15.64	52.13	9.53	6.47	22.64	17.04	11.65	39.65
SummaRuNNer	25.13	17.80	52.54	10.91	7.63	23.54	18.75	13.18	39.87
Bert + SummaRuNNer	25.41	18.08	52.73	11.13	7.82	23.88	19.02	13.44	40.15

여 평가를 진행하였다. 데이터의 유용성을 평가하기 위해 비교사(unsupervised learning) 모델과도 비교하였다.

실험 모델은 다음과 같다.

TEXT RANK : Text RANK[18]알고리즘은 PAGE RANK 알고리즘을 기반으로 한 비교사 학습 추출 기반 문서 요약 알고리즘이다. 문서 내의 문장 간 유사도를 통해 각 문장의 중요도를 계산하고, 중요도가 높은 문장을 가지고 문서를 요약하는 알고리즘이다. 실험을 위해 gensim[19]의 summarization을 사용하였으며, 문장 간의 유사도는 조사, 접사, 어미, 심볼을 제거하고 계산하였다. 알고리즘 수행 후 중요도 상위 3문장으로 요약을 구성하였다.

LEAD-3 : LEAD-3는 기사의 처음 3문장을 선택하여 요약으로 사용하는 모델이다. 신문기사는 주로 두괄식으로 작성되기 때문에 기사의 앞부분에 주요한 내용들이 분포하는 특성을 보인다. 이러한 특성에 의해 LEAD-3는 여러 문서 요약 연구에서 강력한 기준 모델(baseline model)로 사용되어 왔다.

SummaRuNNer : 추출 기반의 문서 요약을 위해 기존의 영어권 문서 요약 연구에서 활용되었던 SummaRuNNer[17]을 한국어 문서 요약 모델로 활용하였다. 영어의 경우 단어 단위의 입력을 사용하였다. 본 연구에서는 이를 한국어에 맞도록 형태소 단위로 수정하였다. 또한 [20]의 연구에서와 같이, 기본적인 언어 분석 정보를 통해 문서 요약의 성능을 높이고자 품사 태그, 문서 내 단어 빈도수, 개체명 여부를 자질로 함께 사용하여 학습하였다. 이 모델에서 형태소는 미리 학습된 128차원의 word2vec 단어 임베딩을 사용하여 표현하였고 다른 자질들은 64차원의 벡터로 표현하여 사용하였다. 단어로부터 문장 표현을 생성하기 위한 Word Layer와 문서의 문맥을 반영한 문장 표현을 생성하는 Sentence Layer는 각각 1개 층의 bi-directional GRU를 사용하였다. 모델을 통해 각 문장의 추출 확률을 계산한 후, 확률 상위 3문장을 선택하여 요약을 구성한다.

BERT+SummaRuNNer : 이 모델은 위에서 설명한 SummaRuNNer 모델의 형태소와 품사 태그 정보 대신 BERT의 출력과 언어 분석 자질을 함께 사용한 모델이다. 본 연구에서는 BERT모델의 Fine-tuning은 수행하지 않았다.

SummaRuNNer 모델과 BERT+SummaRuNNer모델 학습을 위한 배치 사이즈는 각각 32와 4이고 이를 학습률(learning rate) 0.01의 ADAM 알고리즘으로 학습하였다. 학습 과정에서 일정 주기로 모델 파라미터를 저장하였고 검증 데이터를 이용한 실험에서 가장 좋은 성능을 나타낸 모델 파라미터를 평가에 사용하였다.

Table 7은 실험 결과이다. 이 결과에서 비교사 학습 알고리즘의 TEXT RANK 알고리즘과 LEAD-3의 ROUGE-2 F1 점수는 각각 9.53과 6.75의 점수로 본 연구에서 제안하는 데이터로 학습한 SummaRuNNer의 10.91보다 낮은 점수를 나타내었다. 이를 통해 본 연구에서 제안하는 데이터가 자동 문서 요약 모델의 학습에 유용하다는 것을 알 수 있다. 또한 BERT+SummaRuNNer모델이 BERT를 사용하지 않은 모델보다 약간의 성능 향상을 보였다.

7. 결 론

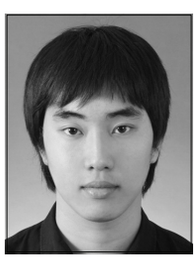
본 연구에서는 한국어 자동 문서 요약 개발을 위한 한국어 문서 요약 데이터 구축 방안에 대해 제안하였다. 한국어 신문기사에 대한 소셜 미디어 상의 설명글을 제목, 부제와 함께 수집하여 약 424,000개 문서로 구성된 한국어 문서 요약 데이터를 구축하였다. 수집된 데이터의 분석을 통해 각 요약의 특성을 살펴보았고 이를 통해 이들 데이터의 활용 방안에 대해 탐색하였다. 또한 이들 데이터가 교사 학습 방식의 자동 문서 요약 모델에 유효함을 실험을 통해 증명하였다. 본 연구의 이러한 결과가 향후 한국어 문서 요약 연구의 기초로 활용될 수 있기를 기대한다.

References

[1] P. Over, H. Dang, and D. Harman, “DUC in context,” *Information Processing & Management*, Vol.43, No.6, pp.1506-1520, 2007.

[2] C. Napoles, M. Gormley, and B. Van Durme, “Annotated gigaword,” in *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, Association for Computational Linguistics, 2012.

[3] J. G. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in SIGIR, 1998.
[4] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," arXiv preprint arXiv: 1603.07252, 2016.
[5] M. Grusky, M. Naaman, and Y. Artzi, "Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies," arXiv preprint arXiv:1804.11283, 2018.
[6] E. Sandhaus, "The new york times annotated corpus. Linguistic Data Consortium," *Philadelphia*, Vol.6, No.12, p.e26752, 2008.
[7] T. Kodaira and M. Komachi, "The Rule of Three: Abstractive Text Summarization in Three Bullet Points," arXiv preprint arXiv:1809.10867, 2018.
[8] B. Hu, Q. Chen, and F. Zhu, "Lcsts: A large scale chinese short text summarization dataset," arXiv preprint arXiv: 1506.05865, 2015.
[9] M. Straka, N. Mediankin, T. Kocmi, Z. Žabokrtský, V. Hudeček, and J. Hajic "SumeCzech: Large Czech News-Based Summarization Dataset," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. 2018.
[10] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom "Teaching machines to read and comprehend," in *Advances in Neural Information Processing Systems.*, 2015.
[11] Su-Jin Baek, "Multi-Document Summarization Method Based on Semantic Relationship using VAE," *Journal of Digital Convergence*, Vol.15, No.12, pp.341-347, 2017.
[12] Kyoung-Ho Choi and Chang-Ki Lee, "End-to-end Korean Document Summarization using Copy Mechanism and Input-feeding," *Journal of KIISE*, Vol.44, No.5, pp.503-509, 2017.
[13] Tae-Hyeong Kim, Ahyoung Kim, Yunseok Noh, Seong-Bae Park, and Seyoung Park "Generation of News Article Dataset Using LEAD for Neural Summarization Model," *Korea Software Congress 2017*, pp.688-690, 2017.
[14] M. Grusky, M. Naaman and Y. Artzi, "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long Papers). 2018.
[15] Yeo-Hoon Jeong, "A Study on the Types of Newspaper Headlines and their Realizations," *The Sociolinguistic Journal of Korea*, Vol.14, No.1, pp.85-113, 2006.
[16] K. Woodsend and M. Lapata, "Automatic generation of story highlights," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2010.
[17] R. Nallapati, F. Zhai and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
[18] F. Barrios, F. López, L. Argerich and R. Wachenchauzer "Variations of the similarity function of textrank for automated summarization," arXiv preprint arXiv:1602.03606, 2016.
[19] Gensim [Internet], https://github.com/summ anlp/gensim.
[20] G. H. Lee and K. J. Lee, "Single Document Extractive Summarization Based on Deep Neural Networks Using Linguistic Analysis Features," *KIPS Transactions on Software and Data Engineering*, Vol.8, No.8, pp.343-348, 2019.



이 경 호
https://orcid.org/0000-0002-3639-3155
e-mail : gyholee@gmail.com
2011년 충남대학교 정보통신공학과(학사)
2013년 충남대학교 정보통신공학과(석사)
2020년 충남대학교 정보통신공학과(박사)
2020년 ~ 현 재 드라마앤컴퍼니 연구원
관심분야 : 자연언어처리, 기계학습, 인공지능

박 요 한
https://orcid.org/0000-0002-5023-5604
e-mail : happy005012@naver.com
2016년 ~ 현 재 충남대학교
전파정보통신공학과 학사과정
관심분야 : 자연언어처리, 기계학습,
인공지능

이 공 주
https://orcid.org/0000-0003-0025-4230
e-mail : kjoolee@cnu.ac.kr
1992년 서강대학교 전자계산학과(학사)
1994년 한국과학기술원 전산학과(공학석사)
1998년 학국과학기술원 전산학과(공학박사)
1998년 ~ 2003년 한국마이크로소프트(유)
연구원
2003년 이화여자대학교 컴퓨터학과 대우전임강사
2004년 경인여자대학 전산정보과 전임강사
2005년 ~ 현 재 충남대학교 전파정보통신공학과 교수
관심분야 : 자연언어처리, 기계학습, 인공지능, 정보검색