JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 26, No. 7: 949~955, Jul. 2022

딥러닝 텍스트 요약 모델의 데이터 편향 문제 해결을 위한 학습 기법

조준희[1] · 오하영[2*]

Training Techniques for Data Bias Problem on Deep Learning Text Summarization

Jun Hee Cho[1] · Hayoung Oh[2*]

[1]Undergraduate Student, Web Progamming, Korea Digital Media High School, Ansan, 15255 Korea
[2*]Associate Professor, College of Computing and Informatics, Sungkyunkwan University, Seoul, 03063 Korea

요 약

일반적인 딥러닝 기반의 텍스트 요약 모델은 데이터셋으로부터 자유롭지 않다. 예를 들어 뉴스 데이터셋으로 학습한 요약 모델은 커뮤니티 글, 논문 등의 종류가 다른 글에서 핵심을 제대로 요약해내지 못한다. 본 연구는 이러한 현상을 '데이터 편향 문제'라 정의하고 이를 해결할 수 있는 두 가지 학습 기법을 제안한다. 첫 번째는 고유명사를 마스킹하는 '고유명사 마스킹'이고 두 번째는 텍스트의 길이를 임의로 늘이거나 줄이는 '길이 변화'이다. 또한, 실제 실험을 진행하여 제안 기법이 데이터 편향 문제 해결에 효과적임을 확인하며 향후 발전 방향을 제시한다. 본 연구의 기여는 다음과 같다. 1) 데이터 편향 문제를 정의하고 수치화했다. 2) 요약 데이터의 특징을 바탕으로 학습 기법을 제안하고 실제 실험을 진행했다. 3) 제안 기법은 모든 요약 모델에 적용할 수 있고 구현이 어렵지 않아 실용성이 뛰어나다.

ABSTRACT

Deep learning-based text summarization models are not free from datasets. For example, a summarization model trained with a news summarization dataset is not good at summarizing other types of texts such as internet posts and papers. In this study, we define this phenomenon as *Data Bias Problem (DBP)* and propose two training methods for solving it. The first is the 'proper nouns masking' that masks proper nouns. The second is the 'length variation' that randomly inflates or deflates the length of text. As a result, experiments show that our methods are efficient for solving *DBP*. In addition, we analyze the results of the experiments and present future development directions. Our contributions are as follows: (1) We discovered *DBP* and defined it for the first time. (2) We proposed two efficient training methods and conducted actual experiments. (3) Our methods can be applied to all summarization models and are easy to implement, so highly practical.

키워드 : 딥러닝, 요약 모델, 휴리스틱 알고리즘, 학습 기법

Keywords : Deep learning, Summarization Model, Heuristic algorithms, Training techniques

Received 3 May 2022, **Revised** 1 June 2022, **Accepted** 10 June 2022

*** Corresponding Author** Hayoung Oh (E-mail: hyoh79@gmail.com, Tel:+82-2-583-8585)
Associate Professor, College of Computing and Informatics, Sungkyunkwan University, Seoul, 03063 Korea

Open Access **http://doi.org/10.6109/jkiice.2022.26.7.949** print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

1.1. 연구 배경

최근 딥러닝을 활용한 텍스트 요약이 활발하게 연구되고 있다. 텍스트 요약은 텍스트에서 중요한 정보를 추출하는 과정을 의미한다. 보통의 요약 모델은 원문과 정답 요약문으로 구성된 데이터셋을 통해 학습한다. 그러나 이렇게 학습한 요약 모델은 데이터셋으로부터 자유롭지 않다. 예를 들어 뉴스 데이터셋으로 학습한 요약 모델의 경우, 뉴스는 잘 요약해내지만 논문이나 커뮤니티의 글은 제대로 요약하지 못한다. 학습 데이터의 범주를 조금만 벗어나도 모델이 제 기능을 못 하는 것이다. 실제로 표1은 뉴스 데이터셋으로 학습한 모델에 대해 요약 성능을 평가한 결과이다. 뉴스를 제외한 다른 텍스트는 제대로 요약하지 못하는 것을 확인할 수 있다. 본 연구에선 이 현상을 '데이터 편향 문제'라고 정의한다.

Table. 1 Performance of models

Text type	Performance score (ROUGE-2 [1])
News article	**15.12**
Paper abstract	3.76
Internet post	2.69

1.2. 연구 동향

데이터 편향 문제가 없는 요약 모델은 데이터셋에 국한되지 않고 다양한 종류의 글을 요약할 수 있어야 한다. 이와 관련한 선행 연구로는 [2], [3]이 있다. 참고 [2]은 단어를 유의어로 교체하고, 두 단어의 위치를 바꾸는 등의 텍스트 증강법을 제안한다. 그러나 데이터 증강은 단순히 주어진 데이터의 양을 늘리는 것이 목적이기 때문에 데이터 편향 문제를 효과적으로 해결하지 못한다. 참고 [3]은 트랜스포머 구조[4]를 개선하고 학습 과정에서 텍스트를 마스킹하여 모델의 성능을 향상했다. 해당 연구는 다양한 자연어 처리 태스크에 적용할 수 있는 범용 모델을 만드는 것을 목적으로 한다. 하지만 데이터 편향 문제는 요약이라는 단일 태스크 내에서 발생한다는 점에서 이 접근은 데이터 편향 문제 해결과는 거리가 멀다. 그뿐만 아니라 모델 구조를 통한 접근법은 구현이 어렵고 다양한 구조에 적용할 수 없다는 점에서 한계를 갖는다.

상술한 선행 연구는 대부분 단순 성능 향상이 목표였으며 데이터 편향 문제를 직접적으로 해결하려는 시도는 아니었다. 따라서 본 연구는 데이터 편향 문제를 해결하기 위한 학습 기법을 제안하고 실제 실험을 진행하여 그 실효성을 검토한다. 첫 번째는 데이터의 간극을 줄이기 위해 고유명사를 마스킹하는 '고유명사 마스킹'이고, 두 번째는 학습 데이터의 길이를 임의로 늘이거나 줄이는 '길이 변화'이다. 두 학습 기법은 구현이 쉽고 어떤 모델에든 적용 가능하다는 점에서 큰 장점을 갖는다.

Ⅱ. 본 론

2.1. 정의

본 절에서는 언어 모델과 데이터 편향 문제를 정의한다. 논문에서 사용되는 주요 기호는 표2와 같다.

Table. 2 Symbols used in this paper

Variable	Description
x	the original text (input)
y	the target summary (target)
X	the set of original texts
Y	the set of target summaries
$X^{'}$	the another set of original texts
$Y^{'}$	the another set of target summaries
G_{θ}	the summarization model
θ	the parameters of G_{θ}
$G_{\theta}(x)$	the summary of x generated by G_{θ}
B_{θ}	*data bias* of the model
$S(\cdot)$	the similarity evaluation metric of two texts
r	the inflation rate

데이터 편향 문제는 서론에서 언급했듯이 '요약 모델이 학습 데이터셋으로부터 자유롭지 않은 현상'을 의미한다. 즉, 데이터셋 A로 모델을 학습했을 때 데이터셋 B에서는 성능이 제대로 나오지 않는 현상이 데이터 편향 문제이다. 과대적합(overfitting)은 학습 데이터에서는 좋은 성능을 보이나, 테스트 데이터에서는 낮은 성능을 내는 경우를 뜻한다. 언뜻 보면 데이터 편향 문제는 과대적합과 비슷해 보일 수 있다. 하지만 과대적합은 한 데이터셋 내의 학습 데이터와 테스트 데이터 간의 성능 차이를 기준으로 하는 반면, 데이터 편향 문제는 한 데이터셋과 또 다른 데이터셋 간의 성능 차이를 기준으로

한다. 다시 말해, 과대적합은 하나의 데이터셋에 국한된 일반화 문제를, 데이터 편향 문제는 여러 데이터셋에 대한 더 넓은 의미의 일반화 문제를 나타낸다. 데이터 편향 문제를 제대로 파악하기 위해선 이를 수치화 할 필요가 있다. 따라서 데이터 편향 문제를 측정하는 방법으로 '데이터 편향도(B_θ)'를 정의한다. 데이터 편향도는 데이터 편향 문제가 심한 정도를 나타내며, 본 연구의 목적은 이 값을 최소화하는 것이다.

$$B_\theta = \left(\frac{score_\theta(X, Y) - score_\theta(X', Y')}{score_\theta(X, Y) + \varepsilon} \right)^2 \quad (1)$$

수식 (1)의 $score_\theta(X, Y)$는 데이터 X, Y에 대한 모델의 성능을 의미한다. 데이터 내의 각 원문(x)과 정답 요약문(y)에 대한 예측 성능의 평균이 $score_\theta(X, Y)$ 값이 된다 (2).

$$score_\theta(X, Y) = \frac{1}{n} \sum_{i=1}^{n} S\left(G_\theta(X_i), Y_i\right) \quad (2)$$

수식 (2) 시그마 안의 함수 $S(\cdot)$는 두 텍스트의 유사도를 평가하는 지표이다. ROUGE[1] 등의 평가 지표가 $S(\cdot)$로 사용될 수 있다. $G_\theta(X_i)$는 요약 모델이 생성한 요약문으로, $S\left(G_\theta(X_i), Y_i\right)$는 모델이 생성한 요약문과 정답 요약문 간의 유사도를 나타낸다. 한편, 수식 (1)의 분자는 학습 데이터셋과 또 다른 데이터셋 간의 성능 차이를 뜻한다. 이 값이 작을수록, 다시 말해 데이터셋 간의 성능 차이가 작을수록 데이터 편향도(B_θ)는 감소한다. 만약 모델 자체의 성능이 감소하면 $score_\theta(X, Y)$ 값과 $score_\theta(X', Y')$ 값이 동시에 작아져 0에 가까워진다. 결과적으로 분자의 값이 줄어들어 데이터 편향도가 감소한다. 그러나 이 경우에는 데이터 편향 문제가 해결됐다고 보기 어렵다. 모델의 품질은 최대한 유지하면서 데이터셋 간의 성능 차이를 줄이는 것이 본 연구의 목적이기 때문이다. 따라서 모델 자체의 성능을 분모 자리에 위치시켜 이를 함께 고려하도록 했다. 또한 분모가 0이 되는 것을 방지하기 위해 절댓값이 매우 작은 양의 상수 ε을 더했다. 본 연구에서 찾고자 하는 최적의 요약 모델은 데이터 편향도가 최소가 되도록 하는 모델이다.

2.2. 제안 기법

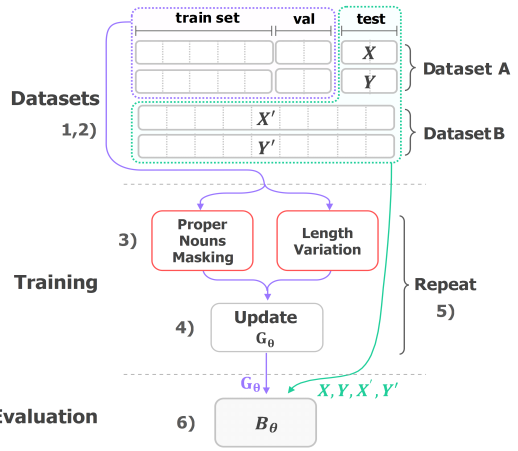

Fig. 1 The overall training process using our methods

본 연구의 제안 기법은 그림1에 붉은색 테두리로 표시된 '고유명사 마스킹 (Proper Nouns Masking)'과 '길이 변화 (Length Variation)'이다. 제안 기법을 적용한 전체적인 학습 과정은 다음과 같다. 1) 뉴스 요약 데이터, 논문 요약 데이터 등 서로 다른 두 종류의 요약 데이터를 준비한다. 2) 첫 번째 요약 데이터를 학습 데이터와 검증 데이터, 테스트 데이터(X, Y)로 분할한다. 3) 제안 기법을 적용하여 학습 데이터를 변형시킨다. 4) 변형된 학습 데이터로 요약 모델(G_θ)을 업데이트 한다. 5) 단계 3-4를 일정 횟수만큼 반복한다. 6) 수식 (1)을 바탕으로 요약 모델의 데이터 편향도 B_θ를 계산한다. 위에서 단계 3)을 제외한 대부분의 단계는 평범한 모델 학습 과정에 속한다. 바꿔 말해, 평범한 학습 과정 중간에 단계 3)만 추가하면 기존 구조를 바꿀 필요 없이 본 제안 기법을 적용할 수 있다. 이처럼 본 제안 기법은 기존 학습 과정에 쉽게 적용할 수 있고 요약 모델의 형태에도 구애받지 않는다는 장점이 있다.

가. 고유명사 마스킹

글에서 등장하는 고유명사의 종류는 데이터셋마다 다르다. 예를 들어 커뮤니티 글에서는 일상적인 고유명사가 자주 등장하지만 논문에서는 전문적인 고유명사가 주를 이룬다. 자주 등장하는 고유명사가 다르면 데이터셋 간의 간극이 커져 데이터 편향 문제가 발생할 수 있다. 따라서 데이터셋 간의 간극을 줄이기 위해 표3처

럼 고유명사를 마스킹 한다.

Table. 3 The example of proper nouns masking

original	Sejong was the fourth king of the Joseon dynasty.
masking	#1 was the fourth king of the #2 dynasty.

나. 길이 변화

일반적으로 원문이 길어지면 요약문도 이에 비례하여 길어진다. 논문의 초록은 뉴스 기사보다 짧기 때문에 각 요약문끼리도 길이에 차이가 생긴다. 표4는 데이터셋 별 원문 평균 길이(단어 수)를 비교한 것이다. 데이터셋에 따라 평균 길이가 크게 차이나는 것을 확인할 수 있다.

Table. 4 Average text length.

Dataset	Text type	# of words
SciTLDR[5]	Paper abstract	**159**
Xsum[6]	News article	358
Reddit TIFU[7]	Internet post	384

일반적인 방법으로 학습한 요약 모델은 특정 길이의 텍스트에만 적응하게 된다. 때문에 길이가 다른 텍스트에 대해선 적당한 길이의 요약문을 생성하지 못하는 데이터 편향 문제가 발생한다. 이를 방지하기 위해 본 연구는 학습 데이터의 길이를 임의로 늘이거나 줄이는 길이 변화 기법을 적용하였다. 아래 표5는 문장에 길이 변화 기법을 적용한 예시이다.

Table. 5 The example of length variation

original	Sejong was the fourth king of the Joseon dynasty.
deflated	Sejong was king of Joseon.
inflated	Sejong was the fourth **the dynasty** king of the Joseon **of** dynasty.

텍스트의 길이를 늘이는 방법은 크게 문장 단위와 단어 단위로 나눌 수 있다. 문장 단위 늘이기는 텍스트 내 임의의 문장을 복사, 삽입하는 방식으로 진행된다. 단어 단위 늘이기도 이와 비슷하게 임의의 단어를 텍스트 내에서 복사, 삽입하는 방식을 의미한다. 본 제안 기법에서는 가장 먼저 문장 단위로 텍스트를 늘인 뒤, 단어 단위로 다시 한번 텍스트를 늘이는 방법을 사용한다. 이에 대한 구체적인 과정은 다음과 같다. 1) 텍스트 내 문장을 무작위로 하나 선택하고 임의의 위치에 삽입한다. 2) 이를 K번 반복한다. 3) 이번에는 텍스트 내 단어를 무작위로 하나 선택하고 임의의 위치에 삽입한다. 3) 마찬가지로 이를 L번 반복한다. 반복 횟수인 K, L은 길이 증가율(r)을 통해 결정한다. 길이 증가율은 길이 조절 기법을 통해 도달하고자 하는 목표 길이 비율이다. 원본 텍스트와 길이를 늘인 텍스트를 각각 x와 x'이라 할 때, 길이 증가율 r은 (3)과 같이 정의할 수 있다.

$$r = \frac{N(x')}{N(x)} \tag{3}$$

수식 (3)에서 $N(\cdot)$는 텍스트의 길이를 나타낸다. 만약 텍스트의 길이가 1.5배 증가했다면 길이 증가율(r)은 1.5가 된다. 길이 조절 기법에서는 길이 증가율을 입력으로 받아 이에 맞춰 K와 L값을 정한다.

길이 늘이기와 반대로, 텍스트의 길이를 줄이는 것은 중요하지 않은 단어나 문장을 삭제하는 방식으로 진행된다. 이때 중요하지 않은 단어와 문장은 TF-IDF와 ROUGE[1]를 바탕으로 결정한다. TF-IDF는 단어 등장 빈도를 바탕으로 특정 문서에서 해당 단어의 중요도를 계산하는 통계 방법이며, ROUGE[1]는 두 텍스트 간의 유사도를 평가하는 대표적인 지표이다.

텍스트의 길이를 줄이는 방법은 중요하지 않은 문장을 삭제하는 문장 단위와, 중요하지 않은 단어를 삭제하는 단어 단위로 나눌 수 있다. 전체 과정은 길이 늘이기와 유사하게 진행되는데, 가장 먼저 문장 단위로 그 다음 단어 단위로 텍스트의 길이를 줄인다. 이를 구체적으로 표현하면 다음과 같다. 1) ROUGE를 기반으로 각 문장의 중요도를 계산한다. 이때 요약문과의 ROUGE 점수가 높은 문장일수록 중요도가 커진다. 2) 가장 중요도가 낮은 문장 하나를 삭제한다. 3) 이를 K번 반복한다. 4) TF-IDF를 기반으로 각 단어의 중요도를 계산한다. 5) 가장 중요도가 낮은 단어를 하나 삭제한다. 6) 이를 L번 반복한다. 텍스트의 길이를 늘일 때와 마찬가지로 K와 L은 길이 증가율(r)에 따라 계산된다. 다만, 이번에는 길이를 줄이는 것이 목적이기 때문에 길이 증가율(r)은 0과 1 사이의 값을 가져야 한다. 예를 들어 원본의 절반만큼 텍스트의 길이를 줄이려 하는 경우, 길이 증가율 r은 0.5가 된다.

이처럼 길이 변화 기법은 길이 늘이기와 길이 줄이기로 구성된다. 하나의 텍스트에 길이 변화 기법을 적용할 때에는 길이 늘이기와 길이 줄이기 중 하나를 임의로 선

택하여 사용한다. 또한 이에 맞춰 길이 증가율도 무작위 값으로 설정한다.

길이 변화 기법은 그 과정에서 문법을 고려하지 않는다. 표5처럼 문법적으로 어색한 문장이 생성될 수 있다. 요약 모델이 학습 도중 이런 어색한 문장을 많이 접하게 되면 모델의 품질도 감소할 수밖에 없다. 따라서 실제 학습 시에는 전체 데이터의 일부에만 길이 변화 기법을 적용하여 모델이 자연스러운 문장을 함께 학습할 수 있도록 했다. 이때 길이 변화 기법이 적용되는 비율은 하이퍼파라미터 p로 나타낸다. 만약 p값이 0.7이라면 전체 데이터의 70%에만 길이 변화 기법을 적용하게 된다.

2.3. 실험

본 절에서는 제안 기법을 실제로 적용한 실험 결과를 소개한다.

가. 데이터셋

실험을 위한 요약 데이터셋은 SciTLDR[5], XSum[6], Reddit TIFU(이하 Reddit)[7]를 사용했다. SciTLDR[5]은 논문의 초록을 요약한 데이터셋이고, XSum[6]은 뉴스 요약 데이터셋이며, Reddit[7]은 커뮤니티 글을 요약한 데이터셋이다.

나. 모델

실험용 요약 모델로는 사전학습 된 BART[8]와 T5[9]를 사용했다. BART는 인코더-디코더(encode-decoder) 계열의 트랜스포머[4] 모델이며, 텍스트 요약 뿐만 아니라 질의응답 등의 분야에서 최고 수준의(state-of-the-art) 성능을 달성했다. T5도 인코더-디코더 기반의 모델로 BART와 마찬가지로 대부분의 텍스트 분야에서 최고 수준의 성능을 달성했다. 본 실험에서는 각각 139M, 60M 크기의 모델을 사용했다.

다. 평가지표

요약 모델의 데이터 편향도(B_θ)를 계산하기 위해선 (1)에서 $S(\cdot)$로 사용될 평가 지표가 필요하다. 본 실험에서는 평가 지표로 ROUGE 점수[1]를 사용했다. ROUGE 점수[1]는 ROUGE-N, ROUGE-L 등으로 구성된 대표적인 요약 평가 지표이다. ROUGE-N은 두 텍스트의 겹치는 n-gram 개수를 기반으로 요약 품질을 평가한다. 반면 ROUGE-L은 가장 길게 겹치는 텍스트의 길이를 기반으로 요약 품질을 평가한다. 본 실험에서는 ROUGE-1, ROUGE-2, ROUGE-L을 사용했다. 하나의 요약 모델에 대한 데이터 편향도를 계산할 때는 세 평가 지표를 각각 사용하여 총 세 개의 데이터 편향도를 구한 뒤, 평균을 냈다.

라. 실험 결과

실험은 제안 기법을 적용한 모델과 그렇지 않은 모델을 학습시키고 성능을 측정하는 방식으로 진행됐다. 길이 변화 기법을 적용하는 비율(p)은 0.75로 설정했으며, 공정한 결과를 위해 모델은 충분히 학습을 진행한 뒤 검증 손실(validation loss)을 기준으로 가장 성능이 좋았던 체크포인트를 평가에 사용했다. 이에 따른 실험 결과는 표6과 같다. 가독성을 위해 테스트 데이터에 대한 평가 점수는 ROUGE-2 값만 표시했다. 표6에서 *masking*은 고유명사 마스킹 기법을, *length*는 길이 변화 기법을 가리킨다.

표6에 따르면 고유명사 마스킹 기법을 적용한 경우(B,F) 모델 종류에 상관 없이 데이터 편향도가 감소하는 것으로 나타났다. 따라서 고유명사 마스킹 기법은 데이터 평향 문제 해결에 도움이 된다고 결론지을 수 있다. 한편, 길이 변화 기법을 적용한 경우(C)와 두 기법을 모두 적용한 경우(D)도 데이터 편향도가 감소하는 경향을 보인다. 그러나 T5 모델의 경우 오히려 데이터 편향도가 증가하기도 했다(G,H). 이는 길이 변화 기법이 모델 구조와 사전 학습 데이터셋 종류에 영향을 받는 것으로 예상되며 향후 연구가 필요해 보인다. 그럼에도 불구하고 길이 변화 기법을 적용했을 때 (C,D,G,H) 데이터 편향도가 증가한 경우, 혹은 변화가 없는 경우보다 감소한 경우가 많다는 점에서 길이 변화 기법도 데이터 편향 문제 해결에 도움이 될 가능성이 있다.

Table. 6 Experimental results
제안 기법에 대한 모델의 성능 및 데이터 편향도를 나타낸 표이다. 데이터 편향도가 낮은 곳을 굵게 표시했다.

methods	train dataset	test dataset			B_θ
		SciTLDR	Xsum	Reddit	
BART					
(A) none	SciTLDR	8.56	2.36	2.08	0.223
	XSum	2.73	17.06	2.82	0.472
	Reddit	5.64	5.15	8.82	0.114

methods	train dataset	test dataset			B_θ
		SciTLDR	Xsum	Reddit	
BART					
(B) masking	SciTLDR	8.25	2.53	2.08	**0.209**
	XSum	1.56	12.54	2.2	**0.469**
	Reddit	5.59	5.12	8.55	**0.110**
(C) length	SciTLDR	8.55	2.66	2.13	**0.207**
	XSum	3.59	15.64	2.84	**0.431**
	Reddit	5.47	4.99	8.49	**0.112**
(D) masking + length	SciTLDR	8.52	2.55	2.08	**0.215**
	XSum	2.99	12.82	1.75	0.478
	Reddit	5.74	4.97	8.79	0.115
T5					
(E) none	SciTLDR	8.05	2.22	3.37	0.151
	XSum	2.46	13.37	1.86	0.455
	Reddit	3.64	2.49	7.06	**0.232**
(F) masking	SciTLDR	7.86	2.33	3.22	**0.148**
	XSum	2.65	9.95	2.07	**0.356**
	Reddit	3.61	2.53	7.07	**0.228**
(G) length	SciTLDR	8.11	2.29	3.29	0.153
	XSum	2.62	12.77	1.27	0.482
	Reddit	3.01	2.52	7.09	0.263
(H) masking + length	SciTLDR	7.59	2.34	3.26	**0.138**
	XSum	2.7	9.58	1.11	**0.415**
	Reddit	2.96	2.44	6.95	0.264

바. 한계 및 향후 연구

본 연구는 크게 두 가지 한계를 갖고 있다. 첫 번째는 데이터 다양성 문제이다. 실험에서는 SciTLDR[5], XSum[6], Reddit[7] 데이터셋을 사용했다. 각각 논문 초록, 뉴스, 커뮤니티 글을 요약한 데이터셋이지만 세 종류만으로 제안 기법의 성능을 완벽히 평가하기엔 무리가 있었다. 따라서 향후에는 리뷰 요약 데이터셋, 대화 요약 데이터셋 등 더 다양한 데이터셋으로 실험을 진행할 필요가 있다. 두 번째는 길이 변화 기법 개선이다. 제안 기법 중 하나인 길이 변화 기법은 그 과정에서 문법을 고려하지 않아 어색한 텍스트가 생성될 우려가 있다. 따라서 향후에는 길이 변화 기법에서 문법 요소를 고려할 필요가 있어 보인다. 뿐만 아니라, 길이 변화 기법이 악영향을 준 경우에 대한 추가 연구도 필요하다. 모델 구조, 사전 학습 데이터셋, 파라미터 개수 등 길이 변화 기법 작동에 영향을 주는 요소를 파악하여 개선하면 더욱 효율적으로 데이터 편향 문제를 해결할 수 있을 것이다.

본 연구의 제안 기법, 특히 길이 변화 기법은 데이터 편향 문제 해결을 위해서 뿐만 아니라 데이터 증강을 위해서도 사용될 수 있다. 거꾸로, 기존의 데이터 증강 기법을 데이터 편향 문제 해결을 위한 기법으로 개선할 수도 있다. 더 나아가 데이터 편향 문제는 텍스트 요약뿐만 아니라 감성 분석, 질의 응답, 음성, 이미지 등의 영역까지도 확장될 수 있다. 이처럼 본 연구는 단일 연구에서 그치는 것이 아닌, '범용 모델'이라는 인공지능의 궁극적 목표를 향한 또 하나의 연장선으로 볼 수 있다.

Ⅲ. 결 론

본 연구는 데이터 편향 문제를 정의하고 이를 해결하기 위한 두 가지 학습 기법을 제안했다. 또한 실험을 통해 본 제안 기법이 데이터 편향 문제를 해결하는 데 효과가 있음을 확인했다. 하지만 실험에 사용된 데이터가 다양하지 않고, 제안 기법이 문법성을 고려하지 않는다는 한계도 존재했다. 결론적으로 본 연구는 데이터 편향 문제가 없는 일반화된 요약 모델의 필요성을 인식하고, 실용적인 학습 기법을 제안하여 문제 해결에 기여했다는 점에서 의의가 있다. 그뿐만 아니라 데이터 편향 문제는 음성, 이미지 등 다양한 분야로 확장될 수 있다. 본 연구를 시작으로 관련 연구가 지속적으로 이루어져 다양한 분야에서 데이터 편향 문제가 해결되길 바란다.

ACKNOWLEDGEMENT

This work was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2022R1F1A1074696).

REFERENCES

[1] C. Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in *Proceedings of the Workwhop on Text Summarization Branches Out*, Barcelona, Spain, pp. 74-81, 2004.

[2] J. Wei and K. Zou, “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification

Tasks," in *Proceeding of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 6382-6388, 2019. DOI: 10.18653/v1/D19-1670.

[3] Z. Liu, J. Li, and M. Zhu, "Improving Text Generation with Dynamic Masking and Recovering," in *International Joint Conference on Artificial Intelligence*, Online, pp. 3878-3884, 2021. DOI: 10.24963/ijcai.2021/534.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *31st Conference on Neural Information Processing Systems*, Long Beach: CA, USA, 2017.

[5] I. Cachola, K. Lo, A. Cohan, and D. Weld, "TLDR: Extreme Summarization of Scientific Documents," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, pp. 4766-4777, 2020. DOI: 10.18653/v1/2020.findings-emnlp.428.

[6] S. Narayan, S. B. Cohen, and M. Lapata, "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization," in *Proceeding of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 1797-1807, 2018. DOI: 10.18653/v1/D18-1206.

[7] B. Kim, H. Kim, and G. Kim, "Abstractive Summarization of Reddit Posts with Multi-level Memory Networks," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis: MN, USA, pp. 2519-2531, 2019. DOI: 0.18653/v1/N19-1260.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 7871-7880, 2020. DOI: 10.18653/v1/2020.acl-main.703.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1-67, Jun. 2020.

조준희(Jun Hee Cho)

한국디지털미디어고등학교 웹프로그래밍과 (2020~)
※관심분야 : 딥러닝, 자연어 처리

오하영(Hayoung Oh)

성균관대학교 소프트웨어융합대학 (2020.03~)
※관심분야 : 딥러닝, 자연어 처리