JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 22, No. 9, pp. 1487-1491, Sep. 2021

Check for updates

Attention 알고리즘 기반 요약 콘텐츠 생성 방안 연구

이 소 연[1] · 최 지 은[2] · 유 선 용[3*]
[1]전남대학교 ICT융합시스템공학과 석사과정 [2]전남대학교 컴퓨터정보통신공학과 학사 [3*]전남대학교 ICT융합시스템공학과 교수

A Study on the Content Summary Based on Attention Algorithm

Soyeon Lee[1] · Ji-Eun Choi[2] · Sunyong Yoo[3*]

[1]Master's Course, Department of ICT Convergence System Engineering, Chonnam National University, Gwangju 61186, Korea
[2]Bachelor's Degree, School of Electronics and Computer Engineering, Chonnam National University, Gwangju 61186, Korea
[3*]Professor, Department of ICT Convergence System Engineering, Chonnam National University, Gwangju 61186, Korea

[요 약]

최근 바쁜 현대인들에게 뉴스, 도서, 영화, TV 프로그램 등 각종 콘텐츠를 요약해 제공하는 '요약 콘텐츠(Summary Contents)' 시장이 주목받고 있다. 기존 대부분의 콘텐츠 요약 기법은 문장을 분석하여 통계적으로 의미있는 단어를 추출하는 것에 집중하였다. 하지만 단순히 단어의 구문적 특징만을 고려할 경우 단어들 간의 연관성과 내재된 의미를 놓치는 경우가 많다. 따라서, 문장의 복잡한 구조와 의미를 고려하여 핵심 요소를 추출하고 추상적 요약을 만들기 위한 방법이 필요하다. 본 연구는 영문 리뷰 데이터와 국문 신문 기사 데이터에 attention 알고리즘 기반 딥러닝 모델을 적용하여 핵심 문맥을 반영한 추상적 요약문을 생성한다. 실험 결과, 제안하는 모델은 단어의 의미를 중점적으로 해석해 성공적으로 영문 리뷰 데이터의 요약 예측문을 생성하였다. 국문 텍스트의 경우 전처리가 까다로움에도 실제와 유사한 예측 요약문을 생성하는 유의미한 결과를 보였다. 수기 확인(manual curation) 및 설문조사 결과, 생성된 요약 콘텐츠는 주요 단어 및 추상적 개념을 효과적으로 생성하여 문장을 요약하는 것을 확인할 수 있었다. 본 연구는 향후 현대인들에게 정보를 전달하는 과정에서 시간 단축 및 편리성을 제공할 수 있을 것이다.

[Abstract]

Recently, the 'content summary' market, which summarizes various contents such as news, books, movies, and TV programs to busy people, is drawing attention. Most existing content summarization techniques focused on analyzing sentences to extract statistically meaningful words. However, simply considering the syntactic features of words often misses the associations and intrinsic meanings between words. Therefore, a method for extracting key elements and making abstract summaries is needed considering the complex structure and meaning of the sentence. This study applies an attention algorithm-based deep learning model to English review and Korean newspaper article data to generate abstract summaries reflecting the core context. The model in this study successfully generated summary prediction of the English review data by interpreting the meaning of the words. In the case of korean text, although preprocessing is difficult, the results showed significant results in generating predictive summaries similar to actual summaries. The results of manual curation and surveys showed that the inferred summary content effectively generated key words and abstract concepts to summarize sentences. This study will be able to provide time reduction and convenience in the process of delivering information to modern people in the future.

색인어 : 텍스트 요약, 콘텐츠 요약, 텍스트 전처리, Attention 알고리즘, 자연어처리
Keyword : text summary, content summary, text preprocessing, attention algorithm, natural language processing

http://dx.doi.org/10.9728/dcs.2021.22.9.1487

Received 16 August 2021; **Revised** 27 August 2021
Accepted 07 September 2021

***Corresponding Author; Sunyong Yoo**

Tel: 062-530-1761
E-mail: syyoo@jnu.ac.kr

Ⅰ. 서 론

최근 뉴스, 도서, 영화, TV 프로그램 등 각종 콘텐츠를 요약해 제공하는 '요약 콘텐츠(content summary)' 시장이 각광받고 있다. 따라서, 다양한 주제의 원문 데이터로부터 추출 요약문과 생성요약문을 도출해 소비자에게 정보를 전달하는 과정에서 시간 단축 및 편리성을 제공할 필요가 있다.

텍스트 요약은 크게 추출적 요약과 추상적 요약으로 나눌 수 있다. 추출적 요약은 원문에서 중요한 핵심 문장 또는 단어 구를 몇 개 추출하고 이들로 구성된 요약문을 만드는 방법이다. 따라서 추출적 요약의 결과로 나온 요약문의 문장이나 단어 구들은 전부 원문에 있는 문장들로 구성된 것이 큰 특징이다. 추출적 요약의 대표적인 알고리즘으로 머신러닝 알고리즘인 텍스트 랭크(TextRank)가 있다[1]. 반대로 추상적 요약은 원문에 없던 문장이라도 핵심 문맥을 반영해 새로운 문장을 생성해서 원문을 요약한다. 추출적 요약보다는 구현 난이도가 높으며 대표적인 알고리즘으로 Seq2Seq(Sequence to Sequence) 모델이 있다[2]. 하지만 RNN(Recurrent Neural Netwok)에 기반한 Seq2Seq 모델은 하나의 고정된 크기의 벡터에 모든 정보를 압축하고 있어 정보 손실이 발생하며, RNN의 기존 문제였던 기울기 소실 문제가 발생해 입력 문장이 길어지게 되면 정확도가 다소 떨어지게 된다[3]. 반면 attention 알고리즘의 경우, 하나의 고정된 크기 벡터를 사용하는 것이 아닌, 중요한 단어에 집중하여 가변 크기의 벡터를 디코더에 전달하는 기법으로 입력 문장이 길어지게 되더라도 정확도가 떨어지지 않는다[4].

따라서 본 논문에서는 영문 데이터와 국문 데이터에 대해 attention 알고리즘 기반의 딥러닝 모델을 적용하여 추출적 요약을 도출하였다. Attention 알고리즘을 이용해 요약문을 추출하는 여러 선행 연구들이 존재한다. 초기 연구에서는 attentive CNN encoder를 사용하여 텍스트를 압축하고 neural network 언어 모델이 요약을 생성하는 모델을 사용하였다[5]. 그 후에는 원문을 잘 표현하도록 gated attention encoder를 도입해 생성된 요약문이 원본 텍스트와 높은 의미적 관련성을 갖도록 했다[6]. 본 연구에서는 seq2seq에 attention layer를 추가한 모델을 도입했다. 특히 seq2seq에서는 RNN 에 기반하지 않고 LSTM을 사용해 기울기 소실 문제를 해결했다. 거기에 attention layer을 더하여 제안한 모델이 성공적으로 추출적 요약을 하도록 도왔다.

국문 데이터의 경우, 최근 정부의 디지털 뉴딜 핵심사업 '데이터 댐' 사업의 추진과 더불어 그 양이 크게 증가하고 있는 상황이다. 이에 4차 산업혁명의 핵심기술 중 하나인 딥러닝 기술을 기반으로 국문 텍스트를 효과적으로 분석하여 핵심 내용을 발췌함으로써 국문 기반 인공지능의 발전에 기여하고자 한다.

Ⅱ. 본 론

2-1 실험 데이터 및 구현 환경

영문 데이터는 Kaggle의 'Amazon fine food review'로부터 100,000개 샘플을 수집하여 연구를 진행하였다[7]. 국문 데이터의 경우 AI-HUB에서 제공해주는 국문 문서 요약 텍스트 데이터 16,983개를 수집하여 연구를 진행하였다[8]. 해당 데이터는 다양한 국문 원문 데이터로부터 정제된 추출 및 생성 요약문으로서 검증된 것이며, 요약문 재사용에 제한이 없기 때문에 저작권 문제가 발생하지 않는다.

사용언어는 딥러닝 모델링을 구현할 수 있도록 다양한 기능을 제공해주는 구글(Google)에서 파이썬(Python) 기반 딥러닝 오픈소스 패키지 텐서플로(TensorFlow)를 이용하여 구현하였다.

2-2 방법

1) 전처리 과정

데이터 전처리 과정은 다음과 같이 진행하였다. 첫 번째, 원본 내용과 생성요약 내용을 각각 분리한 다음 중복된 정보 및 NULL 값을 가지는 행을 제거해 주었다. 최종적으로 사용된 영문 데이터의 샘플 수는 88,425개이며, 국문 데이터의 샘플 수는 16,973개이다. 두 번째, 영문 데이터에서 y'all : you all, you're : you are, you'll : you will, you've: you have 등의 줄임말과 약어들을 풀어서 변환한 다음 자연어 처리 및 분석, 텍스트 마이닝을 위한 파이썬 패키지 NLTK(Natural Language TooKit)에서 지정한 불용어 179개를 제거해 주었다[9]. 국문 데이터는 '많', '만들', '등', '또' 등 95개의 불용어를 직접 지정해 제거했다. 세 번째, 각각의 summary 텍스트 맨 앞에 시작 토큰 'sostoken', 맨 뒤에 종료 토큰 'eostoken'을 추가해 디코더의 입력과 출력으로 새 컬럼을 만들어준 뒤 토큰화를 진행했다. 네 번째, 데이터셋을 학습 데이터셋과 테스트 데이터셋으로 8:2의 비율로 나눠주고 텍스트형을 숫자로 바꾸는 정수 인코딩 작업을 거쳐 컴퓨터가 해당 데이터를 인식할 수 있도록 해주었다. 마지막으로 모든 텍스트가 같은 길이를 유지할 수 있도록 'zero padding'을 추가해 텍스의 길이가 최대 문장 길이보다 작으면 나머지 부분은 0으로 채워주었다.

2) Attention 기반 요약 컨텐츠 생성 모델링

딥러닝 모델 구조를 시각화해보면 [그림 1]과 같다. 인코더는 embedding layer와 3개의 LSTM(Long short-term memory) layer로 구성된다[10]. Embedding layer는 인코더의 단어를 벡터로 표현하는 단어 임베딩을 수행한다. LSTM의 모두 드롭아웃은 전체의 40%, 순환 드롭아웃은 전체의 40%로 설정하였다.

디코더는 embedding layer와 LSTM, attention layer로 구성된다. Embedding layer는 디코더의 단어를 벡터로 표현하는 단어 임베딩을 수행한다. LSTM의 드롭아웃은 40%, 순환 드롭아웃은 20%로 설정하였다. 여기까지는 기존의 Seq2Seq 구조와 유사하나 attention Layer를 추가한다는 점에서 차이가 있다. 마지막으로 출력층의 활성화 함수인 softmax를 사용하여 0~1 사이의 정규화된 값으로 출력하도록 설정하였다[11].

3) 모델 매개변수

딥러닝 모델 구조를 인코더 디코더 방식으로 만든 다음, 영어 텍스트 요약 모델과 한글 텍스트 요약 모델 모두 epoch를 100, batch 크기를 128로 설정한 후, early stopping을 사용하였으며 최적화 함수는 RMSProp를 적용하여 모델을 학습시켰다[12].

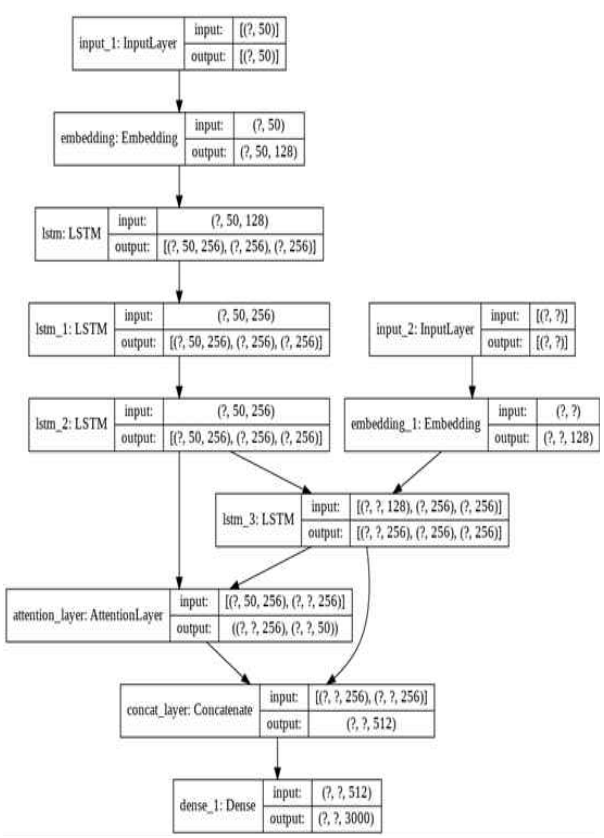


그림 1. Attention 기반 모델 구조 시각화
Fig. 1. Attention-based model structure

Ⅲ. 실험 결과

영문 텍스트와 국문 텍스트 특징을 비교하여 확인하였다. [그림 2]와 같이 영문 원문 데이터는 문장 길이가 10~200 사이로 이루어져 있지만, 국문 원문 데이터의 경우 100~500 사이로 이루어져 있어 데이터 전처리 과정에 영문 데이터보다 다소 많은 시간이 소비되었다. 추가로, 영문은 띄어쓰기 단위로 비교적 명확한 의미가 있어 구분이 쉽지만, 국문은 조사와 어미의 변형으로 인해 의미 단위를 쉽게 구분하기 어려워 KoNLPy(Korean natural language processing in Python)에서 제공하는 패키지 중 Okt와 Mecab를 이용해 국문의 형태소를 분석하고 토큰화를 진행해 지속적으로 보완해야 한다[13].

영문 텍스트 원문으로부터 실제 요약문과 예측된 요약문을 함께 본 결과, [그림 3]과 같이 예측 요약문에서는 원문에 존재하지 않지만 비슷한 의미인 다른 단어로 요약되어 추출되거나, [그림 4]와 같이 실제 요약문과 유사한 강조 표현이 사용되는 것을 확인할 수 있었다. 하지만 [그림 5]와 같이 모든 예측 원문에서 실제 요약문과 유사하게 추출되지 않고 실제 요약문보다 단순하게 요약이 되거나, [그림 6]과 같이 실제 의미와 전혀 다른 예측 요약이 추출되는 경향도 보였다. [그림 7]과 같이 국문 텍스트 원문으로부터는 실제 요약문과 유사한 예측 요약문을 추출했지만, [그림 8]과 같이 중복된 단어와 의미 없는 문장들이 추출되는 경우도 존재하는 것을 확인하였다. 이는 학습에 사용된 영문 텍스트보다 한글 텍스트의 수가 훨씬 적기 때문이며 향후 학습 데이터를 추가하여 개선하는 작업이 필요하다고 사료된다. 추가로 20명의 응답자에게 예측 요약문이 원문의 내용을 의미론적으로 반영하고 있는지 정성적인 평가를 수행하였다. 질문은 총 5개 문항으로 1) 요약문이 실제 텍스트의 의미를 잘 포함하고 있는가? 2) 요약문의 중요 요소들 사이의 연관성 관계가 잘 표현되어 있는가? 3) 추상적 표현으로 요약문이 적절하게 작성되었는가? 4) 의미론적으로 내용을 파악하기에 어려움이 있는가? 5) 요약문 전달이 효과적인가?; 로 구성된다. 각 문항은 2점 만점으로 구성하였으며, 항목별 평균을 낸 결과 10점 만점에 평균 7.53이라는 의미있는 결과를 얻을 수 있었다.

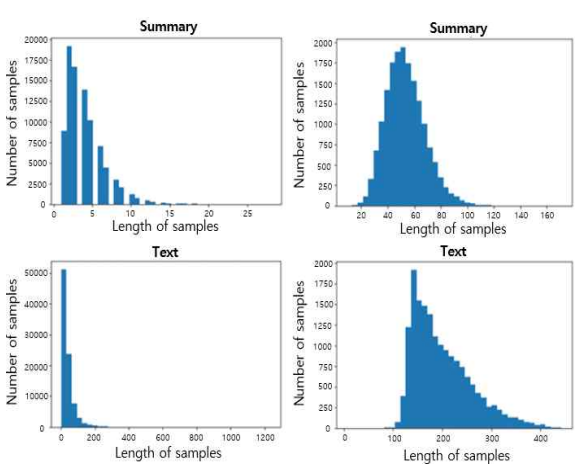


그림 2. (좌) 영어 원문 문장 길이 히스토그램, (우) 한글 원문 문장 길이 히스토그램
Fig. 2. (Left) Histogram of English sentence length; (Right) Histogram of Korean sentence length

▸Original: two **pug**s major rawhide bone pig ear fans reason interested chews one eat getting anything else even bother
▸Actual summary: pugs not interested
▹Predicted summary: not for **small dogs**

그림 3. 원문에 존재하지 않지만 비슷한 의미인 다른 단어로 요약문이 생성된 경우
Fig. 3. Case in which a summary was generated with other words that does not exist in the original sentence but has a similar meaning

▸Original: chips really good taste better name brands cheaper salty like brands flavors blend really well right amount sour cream onion flavor little salt mixed start eating hard stop definitely buying potato chips rather another name brand sure state name brand
▸Actual summary: these are **soooo** good
▹Predicted summary: **love love love** these chips

그림 4. 실제 요약문과 유사한 강조 표현이 사용되어 요약문이 생성된 경우
Fig. 4. Case in which a summary was generated with highlights similar to actual summary statement

▸Original: thrilled find bunches honey bunches oats offered online since cannot find stores anymore make sure realize size boxes purchasing small almost sample size something size purchased store
▸Actual summary: **delicious with vanilla yogurt and frozen blueberries**
▹Predicted summary: **yummy**

그림 5. 실제 요약문보다 단순하게 요약되는 경향을 보이는 경우
Fig. 5. Case that tends to be simpler than the actual summary statement

▸Original: recently tried pocket coffee friend brought back italy one best chocolate ever searched brought product amazon extremely disappointed get italy instant maxwell coffee middle hershey chocolate compare real thing
▸Actual summary: **do not waste your money**
▹Predicted summary: **simply the best**

그림 6. 실제 의미와 전혀 다른 요약문이 생성되는 경우
Fig. 6. Case in which a summary statement is generated that is completely different from the actual meaning

▸Original (Kor.): 기아자동차는 27일 일반 고객 40명을 초청해 신기술 교육을 진행하는 이벤트를 실시했다고 28일 밝혔다 . 기아 자동차가 처음으로 자동차를 구매하려는 고객들을 대상으로 이들의 안전한 자동차 생활을 지원하기 위해 마련한 행사다.
▸Original (Eng.): Kia Motors announced on the 28th that it invited 40 general customers to train on new technologies on the 27th. It is an event organized by Kia Motors to support their safe car life for customers who are planning to own cars for the first time.
▸Actual summary (Kor.): **기아차 초청 실시**
▸Actual summary (Eng.): **Kia Motors Invited**
▹Predicted summary (Kor.): **기아차 이벤트**
▹Predicted summary (Eng.): **Kia Motors Event**

그림 7. 실제 요약문과 다소 유사한 요약문을 예측하는 경우
Fig. 7. Case predicting summary statements somewhat similar to actual summary statements

▸Original (Kor.): 8일 전남도의 '2019년 추진 결과'에 따르면 문화재 저수지 4곳 교량 3곳 대한 필요한 것으로 조사됐다. 전남도 관계자는 '필요한 공공시설물 정비를 위해 행정안전부에 신속하게 정비할 수 있도록 지원하겠다'고 말했다.
▸Original: According to 'the 2019 Project Results of Jeollanam-do Province' on the 8th, three bridges of four cultural heritage reservoirs were found to be necessary. An official from Jeollanam-do Province said, 'We supported the Ministry of Public Administration and Security to quickly maintaining necessary public facilities.'
▸Actual summary (Kor.): 전남 안전관리 시급
▸Actual summary (Eng.): **Jeollanam-do safety management urgent**
▹Predicted summary (Kor.): **전남도 전남 지자체 증가**
▹Predicted summary (Eng.): **Jeollanam-do Provincial Government Increase**

그림 8. 중복된 단어와 의미 없는 문장으로 요약문이 생성되는 경우
Fig. 8. Case in which a summary statement is generated with duplicate words and meaningless sentences

Ⅳ. 결 론

본 연구에서는 요약 컨텐츠를 생성하기 위해 attention 알고리즘을 기반으로 한 모델을 제안하였다. 제안된 모델은 기존 seq2seq 구조에 attention layer를 추가한 형태이다. 실제 요약문과 견주어 봤을 때 모델은 데이터 요약을 효과적으로 수행하였기 때문에 이를 이용하면 성공적인 데이터 추출적 요약 서비스를 제공 할 수 있을거라 기대된다.

국문 추출적 요약 서비스는 크게 정성적・정량적 기대효과 측면에서 살펴볼 수 있다. 먼저, 추출적 요약 서비스의 정성적 효과는 특정 분야에 한정된 요약 서비스와 달리, 신문 기사・논문・기고문・판결문 등 전 분야의 요약 서비스를 통합적으로 제공하여 교육과 업무의 편의성을 향상하는데 그 의의가 있다. 구체적으로 문서 요약 시간을 단축하여 업무 효율성을 극대화할 수 있을 뿐만 아니라, 요약된 정보를 제공하여 빅데이터 접근성을 높일 수 있다.

추출적 요약 서비스의 정량적 효과는 정부・기업・개인의 관점에서 접근할 수 있다. 정부는 2020년 기준 340개 공공기관의 경영 평가 준비에 기관별 경영 평가팀 직원 17.5명, 인건비 9억 5,451만 6,500원(평균 임금 5,454만 3,800원 *17.5명) [14]의 막대한 비용을 지불하고 있다. 그러나 추출적 요약 서비스의 보고서 요약이 활용된다면, 기관별 경영 평가팀 직원을 17.5명에서 10명 이내로 감축할 수 있으며 인건비도 9억 5,451만 6,500원(평균 임금 5,454만 3,800원 x 17.5명)에서 5억4,543만8,000원(평균 임금 5,454만3,800원*10명)으로 절약할 수 있을 것으로 예상한다. 기업은 작년(2019년) 기준 중소기업의 빅데이터 활용률이 14% [15]에 불가하였으나, 각종 문서를 요약한 빅데이터를 제공한다면 기업의 빅데이터 활용률이 20%까지 증가할 것으로 전망된다. 또한, 빅데이터를 활용하기 위한 기술인력 확보에 어려움을 겪고 있는 54%의 중소기업에서 경영의 효율성을 높이며, 마

케팅 효과를 극대화의 추진할 수 있을 것이다. 개인은 기존의 약 2,000자에 해당하는 기사를 500자 이내로 축소되고 읽기 소요 시간도 약 5분에서 2분 이내로 단축되어 보다 효율적인 정보 습득을 이룰 것으로 예상한다.

감사의 글

본 연구는 2020년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(NRF-2020R1C1C1006007)로서, 관계부처에 감사드립니다.

참고문헌

[1] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[2] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." arXiv preprint arXiv:1409.3215, 2014.

[3] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078, 2014.

[4] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473, 2014.

[5] S.C. Alexander M. Rush, J. Weston, A neural attention model for abstractive sentence summarization, in : EMNLP, pp. 379-389, 2015.

[6] Z.L.S.W. JQian Chen, Xiaodan Zhu, H. Jiang, Distraction-based neural networks for modeling documents, in : Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI, pp. 2091-2100, 2016.

[7] Kaggle[Internet]. Available:https://www.kaggle.com/snap/amazon-fine-food-reviews.

[8] AI-Hub [Internet]. Available: https://aihub.or.kr/aidata/8054.

[9] Loper, Edward, and Steven Bird. "Nltk: The natural language toolkit." arXiv preprint cs/0205028, 2002.

[10] Hochreiter, S., &Schmidhuber, J. Long short-term memory. Neural computation , vol. 9, no. 8, pp. 1735-1780, 1997.

[11] Jang, Eric, Shixiang Gu, and Ben Poole. "Categorical reparameterization with gumbel-softmax." arXiv preprint arXiv:1611.01144, 2016.

[12] Zou, Fangyu, et al. "A sufficient condition for convergences of adam and rmsprop." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[13] Park, Eunjeong L., and Sungzoon Cho. "KoNLPy: Korean natural language processing in Python." Annual Conference on Human and Language Technology . Human and Language Technology, 2014.

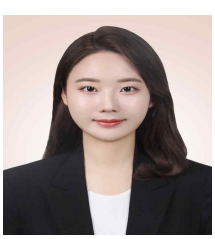
[14] Hankook Ilbo, “8 out of 10 public institutions, money and manpower compared to government evaluations”, 2020 [Internet]. Available: https://www.hankookilbo.com/News/Read /201502110420245700

[15] Maeil Economic Daily, "Big data is essential, but there is no way to obtain it". 2020 [Internet]. Available: https://www.mk.co.kr/news/economy /view/2019/08/654130/

이소연(Soyeon Lee)

2021년 : 전남대학교 (공학석사)

※관심분야 : 생명정보학(bioinformatics), 인공지능(artificial intelligence)



최지은(Ji-Eun Choi)

2021년 : 전남대학교 (공학사)

20018년~2021년: 전남대학교 컴퓨터정보통신공학과 공학사
※관심분야 : 인공지능(AI), 자연어처리(NLP) 등

유선용(Sunyong Yoo)

2012년 : 한국항공대학교 정보통신공학과 (공학석사)
2018년 : 한국과학기술원 바이오및뇌공학과 (공학박사)

2018년~2019년: 국민건강보험공단 빅데이터실 부연구위원
2019년~현 재: 전남대학교 ICT융합시스템공학과 조교수
※관심분야 : 생명정보학(bioinformatics), 인공지능(artificial intelligence), 빅데이터(bigdata) 등