Journal of the Korean Institute of Industrial Engineers
Vol. 48, No. 2, pp. 176-184, April 2022.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2022.48.2.176

데이터 증강 기법을 이용한 한글 개체명 인식

조경선 · 김성범†

고려대학교 산업경영공학과

Korean Named Entity Recognition Using Data Augmentation Techniques

Gyeong Seon Cho · Seoung Bum Kim

Department of Industrial and Management Engineering, Korea University

Named entity recognition (NER) is the task of natural language processing that recognizes a predefined entity name such as a person, place name, or organization into a token in a sentence. The NER is important because it significantly affects the performance of subsequent analyses including semantic search, question answering, and machine translation. The performance of NER for English has been greatly improved with an advent of deep learning techniques with a large dataset of English. However, only few studies have been conducted for languages spoken by ethnic minorities, such as Korean (i.e., Hangul), because an appropriate dataset for NER is difficult to obtain. In this study, we propose using various data augmentation techniques to improve the performance NER for Hangul datasets. Our methods can be applied without pre-trained models or external pre-building. We demonstrated the usefulness of the presented data augmentation techniques using a Changwon University-Naver Challenge dataset and found that even a small dataset can achieve a satisfactory performance for Hangul NER.

Keywords: Named Entity Recognition, Data Augmentation, Korean (Hangul), Natural Language Process

1. 서 론

개체명 인식(named entity recognition)은 사람, 지명, 기관, 날짜 등 미리 정의된 개체명을 문장 내 토큰(token)에 인식하는 자연어 처리(natural language processing)의 한 분야이다. 개체명 인식은 그 자체로도 의미가 있지만, 자연어 처리 분야에서 작업의 대상이나 의도 등 정보 추출을 위해 중요한 역할을 한다. <Figure 1>은 개체명 인식의 예제이다. (a) “서울시는 온맵시 나눔 바자회를 18일 시민청에서 연다고 밝혔다.”라는 문장에 개체명 인식을 수행하면, (b)의 “서울시는”-조직(ORG), “온맵시 나눔 바자회를”-사건(EVT), “18일”-날짜(DAT), “시민청에서”-장소(LOC)와 같이 개체명과 개체명에 연관된 정보를 인식할 수 있다. 개체명 인식이 활용되는 분야를 살펴보면, 질의응답에서 질문 대상을 추출하고, 뉴스 기사, 소설과 같은 긴 글에서 사람, 장소 등 중요한 정보를 감지하여 데이터베이스에 저장할 데이터를 수집하는데 사용된다. 또한 개체명 인식을 통해 정보검색에서 검색 대상을 정확하고 빠르게 인식할 수 있다. 이렇게 추출된 정보는 유튜브(Youtube)에서 시청한 콘텐츠와 비슷한 콘텐츠를 보여주는 추천 시스템이나, 고객의 피드백에서 반복되는 문제를 도출하여 빠르게 해결하도록 도와주는 고객 서비스 시스템 등 다양한 분야에 활용된다. 개체명 인식은 정보 추출에서 필수적인 작업이며(Nadeau and Sekine, 2007), 개체명 인식 시스템의 성능은 질의응답, 정보검색과 같은 상위 시스템 성능에 밀접한 영향을 끼친다. 이렇듯 개체명 인식은 대규모 데이터 셋 처리에서 작업의 대상, 의미 등 정보를 감지하는 매우 중요한 역할을 한다.

Collobert *et al.*(2011)을 시작으로 딥러닝 기법이 개체명 인식에 적용되었고 이후 지속적인 성능 향상을 보이고 있다. 이

† 연락저자 : 김성범 교수, 02841 서울특별시 성북구 안암로 145, 고려대학교 산업경영공학과, Tel : 02-3290-3397, Fax : 02-929-5888, E-mail : sbkim1@korea.ac.kr
2021년 11월 22일 접수; 2022년 1월 11일 수정본 접수; 2022년 1월 27일 게재 확정.

Figure 1. Example of Named Entity Recognition

는 CoNLL03이나 OntoNotes 데이터 셋과 같이 대규모 말뭉치를 포함한 영어 개체명 인식 데이터 셋이 있었기에 가능했다(Li *et al.*, 2020). 딥러닝 분야에서 좋은 성능을 도출하고 모델을 일반화하기 위해서는 양질의 학습데이터가 필요하기 때문이다. 그렇지 않으면 모델의 성능을 떨어뜨리는 과적합 또는 과소 적합이 발생할 가능성이 크다. 하지만 실제 존재하는 데이터 중에서 정확한 라벨이 붙은 양질의 데이터 수는 매우 적다. 영어와 비교하여 언어학적 특성이 다르고, 사용하는 인구가 적은 한글의 경우 개체명 인식에 필요한 라벨링된 학습데이터가 부족하다. 데이터 셋이 부족한 경우, 개체명 인식 성능이 떨어질 수 있기 때문에 학습데이터를 확보하거나 인위적으로 증가시키는 방법을 사용해야 하는데, 라벨링 된 데이터를 구축하기 위해서는 시간이 필요하고 많은 인력과 전문성이 요구된다(Van Engelen and Hoos, 2020). 데이터를 인위적으로 증가시키는 증강기법(augmentation)의 경우 영문 개체명 인식 향상을 위해서는 적용되었으나 한글에 대해서는 Lee *et al.*(2017)이 미리 구축한 사전을 이용하여 한글 개체명 인식 모델에 적용한 연구를 제외하고, 체계적인 연구가 거의 이루어지지 않았다. 따라서 본 연구에서는 학습데이터가 부족한 상황에서 한글 개체명 인식의 성능 향상을 목표로 하고 있다.

한글은 실질적인 의미를 가진 단어와 문법적인 기능을 가진 조사, 접사와 같은 요소가 결합한 교착어로써, 복잡한 언어구조를 가지고 있다(Lee *et al.*, 2020). 이로 인해 자연어 처리를 위한 자원이 부족하여 연구가 활발하게 진행되지 못하였다 (An and Kim, 2015). 한글과 영어의 대표적인 언어구조 차이는 토큰화이다. 영어는 띄어쓰기나 대소문자로 토큰화할 수 있고, 구분된 토큰은 의미가 있는 단어와 의미가 없는 단어로 쉽게 선별된다. 하지만 한글은 영어와 동일하게 토큰화를 진행하더라도, 조사나 접속사 같이 의미가 없는 품사를 제거해야 유의미한 단어를 추출할 수 있다. 하지만 딥러닝의 이점은 단어 임베딩(word embedding)을 사용한 벡터 표현(vector representation) 학습 및 비슷한 의미 체계를 구성을 할 수 있다는 점이므로 (Lecun *et al.*, 2015), 딥러닝 기반의 개체명 인식에서 토큰화가 가장 중요한 요소는 아니다. 따라서 본 연구에서는 기존 영어에서 사용한 띄어쓰기 토큰화 기반의 데이터 증강 기법을 한글 데이터 셋에 적용하여 한글 개체명 인식 분야에 부족한 데이터 증강 기법을 제시하고, 데이터가 부족한 상황에서 저비용, 저자원으로 한글 개체명 인식의 성능을 향상시키는 것을 목적으로 진행하였다.

2018년 네이버와 창원대가 함께하는 NLP Challenge에서 공개한 한글 개체명 인식 데이터 셋을 이용하여 학습데이터가 적은 상황을 만들고, 제안한 방법론을 적용하였을 때 전체 데이터 셋을 이용한 것과 비슷한 성능을 보임을 확인하였다. 본 논문의 중요 기여점 다음과 같다.

- 언어 구조적 한계를 극복하고, 영어 기반 데이터 증강 기법을 한글에 적용하여 한글 개체명 인식의 성능 향상을 보였다.
- 고성능이 필요한 사전 훈련 모델과 사전 구축 없이 활용할 수 있는 문장 기반의 데이터 증강 방법을 한글 개체명 인식에 적용하고 실험을 통해 성능 향상을 입증하였다.

본 논문은 다음과 같은 구조를 가진다. 제2장에서는 자연어 처리 분야에서 선행 연구된 데이터 증강 기법을 소개하고, 제3장에서는 본 연구에서 활용한 데이터와 한글 개체명 인식에 적용할 수 있는 데이터 증강 방법을 설명한다. 제4장에서는 실험을 통해 제안하는 방법론의 한글 개체명 인식에 대한 실험 결과를 보인다. 마지막으로 제5장에서는 결론 및 추후 연구과제를 제시한다.

2. 선행 연구

데이터 증강 기법은 데이터가 충분하지 않은 상황에서 인위적인 변화를 주어 데이터의 양을 증가시켜서, 학습에 필요한 새로운 데이터를 확보할 수 있는 방법론이다. 대표적으로 컴퓨터 비전에서는 이미지를 자르거나 뒤집기, 확대, 회전 등의 원본 데이터에 변형을 주어 데이터를 증가시킨다(Shorten and Khoshgoftaar, 2019). 이러한 과정을 통해서 원본 데이터와 크게 다르지 않은 실제 존재하는 것 같은 데이터를 생성할 수 있다. 이렇게 증강 기법을 이용하여 생성된 데이터는 모델의 성능 향상에 효과적인 결과를 내고 있다.

반면 자연어 처리에서는 문장에서 단어가 하나만 바뀌어도 전체 의미가 쉽게 달라질 수 있고, 문장 배열 순서의 변형으로 문법적인 오류가 발생할 수도 있기 때문에 데이터 변형 방법에 신중히 접근해야 한다. 이런 어려움에도 불구하고 자연어 처리 분야에서 다양한 데이터 증강 기법들이 연구되고 있다.

Edunov *et al.*(2018)은 역 번역(back translation)을 이용하여 새로운 데이터를 생성하는 방법을 제안하였다. 예를 들면 한국어를 영어로 번역한 후, 번역된 데이터를 다시 한국어로 번역하는 과정을 거친다. 이 과정에서 단어가 변형되거나 혹은 미세하게 문장 내 단어의 배열 순서가 변형되어 원본 데이터

Figure 2. Example of BIO Tagging (a) Original Data, (b) BIO Tagged Data

와 다른 데이터를 얻을 수 있다. 해당 연구에서는 탐욕 탐색(greedy search), 빔 탐색(beam search), 임의 샘플링(random sampling) 등의 방법을 사용하여 생성한 인공데이터를 학습데이터에 추가하였을 때 성능이 향상됨을 보였다. Wei and Zou(2019)는 특별한 언어 모델을 사용하지 않고, 단지 원본 데이터를 이용하는 증강기법을 사용하여 텍스트 분류 문제에서 효과를 입증하였다. Wei and Zou(2019)는 문장 내에서 특정 단어들을 선택하고, 그 단어를 유의어로 교체, 삽입하거나 삭제하여 문장을 변형하는 방법을 사용하였다. 직관적인 방법이지만 문장의 의미가 훼손되지 않으며, 원본 데이터의 50%만 사용하여도 100%를 사용할 때와 비슷한 결과를 얻을 수 있음을 보였다. 이는 데이터가 적은 상황에서도 해당 증강 기법을 이용하면 전체 데이터를 사용하는 것과 비슷한 성능을 얻을 수 있음을 의미한다. 사전 학습(pre-trained) 모델을 사용한 데이터 증강 모델도 많이 제안되고 있다. Wu *et al.*(2018)은 여러 개의 단어를 가린 후, 가린 단어를 사전 학습 모델인 bidirectional encoder representations from transformers(BERT)를 이용하여 예측하는 마스킹 된 언어 모델을 사용하여 데이터를 생성하는 방법을 제안하였다. 또한 Kumar *et al.*(2020)은 사전 학습 모델을 이용해서 데이터를 생성하여 학습 과정을 거치고, 이를 다시 사전 학습 모델로 포함해 파인 튜닝(fine tuning)하여 텍스트 분류의 성능을 향상하는 방법을 제안하였다.

개체명 인식은 토큰에 개체명을 할당하는 작업으로, 텍스트 분류, 감성 분석 등과 같이 문장 기반의 자연어 처리 분야와 차이가 있다. 이에 개체명 인식만을 위한 데이터 증강 기법이 연구되었다. Song *et al.*(2020)은 Wikidata를 이용하여 미리 구축된 사전(i.e., gazetteers)을 만들고, 개체명이 태깅된 단어를 기구축한 사전에서 임의로 선택한 단어와 교체하는 방법을 적용하였다. 이를 영어, 중국어, 러시아어 데이터 셋을 대상으로 연구를 진행하였고, 데이터 자원이 풍부한 영어와 중국어에서 성능향상을 확인하였다. Dai and Adel(2020)는 학습데이터에 있는 개체명을 사전으로 구축하고, 문장 내에서 태깅된 단어와 사전 내 동일한 개체명을 가진 단어나 복합어로 교체하는 증강기법을 사용하여 개체명 인식 모델의 성능을 향상하였다.

3. 실험 방법

3.1 데이터셋

실험은 2018년 네이버와 창원대가 함께하는 NLP Challenge (https://github.com/naver/nlp-challenge/tree/master/missions/ner)에서 제공한 데이터를 이용하여 진행하였다. 제공된 데이터는 총 90,000개의 문장으로 구성되어 있으며, 사람이름(PER), 지명(LOC), 기관명(ORG), 학문 분야(FLD), 인공조형물(AFW), 문명/문화(CVL), 특정 사고(EVT), 동물(ANM), 식물(PLT), 금속(MAT), 용어(TRM), 날짜(DAT), 시간(TIM), 숫자(NUM)를 의미하는 총 14개의 개체 범주로 태깅되어 있다. 사용한 데이터는 영어와 같이 띄어쓰기를 기반으로 토큰화 되었기 때문에 태깅된 개체명 토큰이 실제 개체명을 의미하는 단어 외에 접사, 조사 같이 의미를 포함하지 않는 품사를 포함한다. 본 연구의 목적이 데이터 셋이 부족한 상황에서 데이터 증강 기법을 사용하여 개체명 인식의 성능을 향상하는 데 있기 때문에 다양한 데이터 크기 상황을 설정하여 실험을 진행하였다. 데이터 셋은 전체 데이터 셋에서 무작위 추출된 500개 / 2,000개 / 5,000개 데이터를 가지고 있는 소(Small) / 중(Medium) / 대(Large)로 구성하였다.

3.2 데이터 증강 기법

본 연구에서는 BIO 태그를 사용하여 개체명을 예측하는 지도학습 기반의 개체명 인식 방법론을 사용하였다. BIO 태그란 각 토큰을 세부적으로 명명하는 방법으로, 개체명의 첫 번째 토큰에는 시작을 의미하는 B(begin), 개체명의 첫 번째가 아닌 토큰에는 내부를 의미하는 I(inside)를 표기한다. 지정된 개체명에 속하지 않는 토큰은 O(outside)로 표기한다. <Figure 2>는 BIO 태그의 예시로, (a)는 원본 데이터이고, (b)는 BIO 태그가 적용된 형태이다.

“서울시는”은 조직을 의미하는 개체명 ORG가 태그되고, 개체명의 첫 번째를 의미하는 B가 추가로 표기가 되어 B-ORG가 태그된다. “온맵시 나눔 바자회를”은 두 단어 이상 연결된 복합어이기 때문에, 각 단어로 분리한 토큰을 태그한다. 개체명은 사건을 의미하는 EVT가 태그되고, 개체명의 첫 번째 토큰인 “온맵시”에는 B-EVT, 개체명의 첫 번째 토큰이 아닌 “나눔”과 “바자회를”에는 각각 I-EVT로 태그된다. “밝혔다”등 지정할 수 있는 개체명이 없는 토큰에는 O가 태그된다.

본 방법론은 별도로 미리 생성하는 외부 사전이 없이, 학습 데이터셋이나 15만 개의 어의를 포함하는 한국어 어휘의미망(Korean Wordnet, http://korlex.pusan.ac.kr/)을 이용하여 쉽게 필요한 사전을 준비할 수 있다.

3.2.1 동일 태그 교체(Label-wise Tag Replacement)

동일 태그 교체는 학습데이터 내 태그를 활용하여 문장에서 태그를 선택하고, 선택된 태그와 일치하는 태그를 학습데이터 내에서 선택하여 교체하는 방법이다. 여기서 태그란 위에서 설명한 (BIO 태그)-(개체명)으로 구성된 토큰의 태그를 의미한다. 문장에서 선택되는 태그는 한 개 또는 그 이상을 임의로 선택하며, 전체 학습데이터에서 선택되는 태그도 임의로 선택한다. 예를 들면, “손흥민, 맨체스터 유나이티드 이적 확정”이라는 원본 데이터에서 “유나이티드”가 임의로 선택되면, 전체 학습데이터에서 무작위로 “유나이티드”와 같은 I-ORG 태그를 갖는“2TV”를 선택하여 교체한다. 이를 통해서 “손흥민, 맨체스터 2TV 이적 확정” 데이터가 생성된다(<Table 1> 참고).

3.2.2 동일 개체명 교체(Label-wise Mention Replacement)

동일 개체명 교체 방법은 동일 태그 교체처럼 학습데이터 내 태그를 활용한 방법이다. 먼저 문장에서 임의로 개체명을 선택한다. 선택된 개체명은 B 태그만 존재하는 한 개의 단어일 수도 있고, B-I 태그로 구성된 두 단어 이상의 복합어일 수도 있다. 이렇게 선택된 개체명을 학습데이터 내에서 같은 개체명을 갖는 복합어와 교체하는 방법이다. 문장 내에서 선택되는 개체명은 한 개 또는 그 이상이다. 예를 들면, “손흥민, 맨체스터 유나이티드 이적 확정”이라는 원본 데이터에서 조직을 의미하는 ORG 개체명이 임의로 선택되었다면, 전체 학습데이터에서 무작위로 ORG 개체명을 갖는 복합어 “KBS 2TV”를 선택한다. 이를 원본 데이터에 있는 “맨체스터 유나이티드”와 학습데이터에서 선택된“KBS 2TV”를 교체하여 “손흥민, KBS 2TV 이적 확정” 데이터를 생성한다(<Table 1> 참고).

3.2.3 유의어 교체(Synonym Replacement)

유의어 교체는 선택한 단어를 같은 의미가 있는 단어로 교체하는 방법으로 동의어를 찾기 위해 별도의 외부 자료 사전이 필요하다. 단, 사전을 미리 구축할 필요는 없으며, 한국어 어휘의미망을 이용하여 쉽게 동의어를 찾을 수 있다. 문장에서 임의로 토큰을 선택하고, 선택한 토큰에서 의미를 갖지 않는 품사를 제거한다. 실제 의미가 있는 단어만 추출한 후, 같은 의미의 다른 단어를 사전에서 선택하여 교체한다. 문장 내에서 선택되는 태그는 한 개 또는 그 이상이지만, 선택된 단어와 같은 의미를 가진 단어가 사전에 존재하지 않는다면, 교체되는 단어가 없을 수도 있다. 예를 들면, “손흥민, 맨체스터 유나이티드 이적 확정”이라는 원본 데이터에서 “이적”이라는 토큰을 임의로 선택한다. 선택된 토큰은 의미를 갖지 않는 품사를 포함하고 있지 않기 때문에 별도의 전처리 작업을 거치지 않고 동의어로 교체될 수 있다. 사전에 의해서 “이적”의 사전에 의한 동의어가 “입단”이라면, “손흥민, 맨체스터 유나이티드 입단 확정” 데이터가 생성된다(<Table 1> 참고).

3.2.4 임의 삽입(Random Insertion)

임의 삽입은 유의어 교체 방법과 같이 한국어 어휘의미망과 같은 사전이 필요한 방법론이다. 문장 내에서 임의로 토큰을 선택하고, 선택된 토큰에서 불용어를 제외하고 정제된 단어와 같은 의미를 가진 다른 단어를 사전에서 찾는다. 이렇게 선택된 동의어를 문장 내 임의의 자리에 삽입한다. 예를 들면, “손흥민, 맨체스터 유나이티드 이적 확정”이라는 원본 데이터에서 “이적”이라는 토큰이 선택되고, 그 동의어가 “입단”이라면 원본 데이터의 임의 위치에 동의어를 삽입하여 “입단 손흥민, 맨체스터 유나이티드 이적 확정” 데이터를 생성할 수 있다(<Table 1> 참고).

3.2.5 임의 교체(Random Swap)

임의 교체는 문장 내에서 임의의 두 개체명을 선택하고, 선택된 개체명들의 토큰 위치를 변경하는 방법이다. 문장 내 토큰만을 사용하여 변경하는 방법으로, 별도의 사전이나 학습데

Table 1. Example of Various Data Augmentation. Bold/underline Indicates the Result of Applying the Augmentation Methods

Model	Example				
Original Sentence	손흥민, B-PER	맨체스터 B-ORG	유나이티드 I-ORG	이적 O	확정 O
Label-wise Tag Replacement	손흥민, B-PER	맨체스터 B-ORG	**2TV** **I-ORG**	이적 O	확정 O
Label-wise Mention Replacement	손흥민, B-PER	**KBS** **B-ORG**	**2TV** **I-ORG**	이적 O	확정 O
Synonym Replacement	손흥민, B-PER	맨체스터 B-ORG	유나이티드 I-ORG	**입단** **O**	확정 O
Random Insertion	손흥민, B-PER	입단 **O**	맨체스트 B-ORG	유나이티드 I-ORG	확정 O
Random Swap	**맨체스트** **B-ORG**	**유나이티드** **I-ORG**	**손흥민**, **B-PER**	이적 O	확정 O
Random Deletion	손흥민, B-PER	맨체스터 B-ORG	유나이티드 I-ORG	이적 O	

이터를 필요로 하지 않는다. 예를 들면, “손흥민, 맨체스터 유나이티드 이적 확정”이라는 원본 데이터에서 사람을 의미하는 PER과 조직을 의미하는 ORG 개체명이 선택되었다면, 두 개체명을 갖는 토큰인 “손흥민”과 “맨체스터 유나이티드” 위치를 교체하여 “맨체스터 유나이티드 손흥민, 이적 확정” 데이터를 생성한다(<Table 1> 참고).

3.2.6 임의 삭제(Random Deletion)

임의 삭제는 문장 내에서 임의의 토큰을 선택하고, 선택된 토큰을 삭제하는 방법이다. 문장 내 토큰만을 사용하여 변경하는 방법으로, 별도의 사전이나 학습데이터를 참조하지 않는다. 예를 들면, “손흥민, 맨체스터 유나이티드 이적 확정”이라는 원본 데이터에서 “확정” 토큰이 선택되었다면, 선택된 토큰을 삭제하여 “손흥민, 맨체스터 유나이티드 이적”으로 데이터를 변형한다(<Table 1> 참고).

3.3 기본 모델 및 하이퍼파라미터

개체명 인식의 모델은 <Figure 3>처럼 (a)입력 데이터에 대한 분산 표현(distributed representation) 단계, (b)입력 데이터에 대한 양방향 문맥을 반영하는 컨텍스트 인코더(context encoder) 단계, 마지막으로 (c)출력 데이터의 양방향 문맥을 반영하는 태그 디코더(tag decoder) 단계로 분류된다.

본 연구에서는 사전 학습되지 않은 모델과 사전 학습된 모델에서 적용 방법론의 효과를 확인하기 위해, 컨텍스트 인코더 단계에서 BiLSTM과 KoBERT 모델을 사용한다. BiLSTM은 단어의 순서 집합인 문장처럼 순차 데이터 모델링에 적합한 모델이다. BERT는 자연어를 양방향으로 사전 학습하는 모델로써 자연어 처리에서 다양하게 활용할 수 있을 뿐 아니라 그 성능 또한 우수한 모델로 평가된다. 본 실험에서는 한국어 데이터셋을 사전 학습한 KoBERT를 이용하였다. (a)분산 표현 단계에서 BiLSTM 모델을 사용하기 전에 페이스북에서 개발한 FastText로 단어 임베딩을 진행하였다. 반면, KoBERT는 주변 단어들의 정보를 이용한 동적인 단어 임베딩을 모델 내부에서 생성하기 때문에 문맥에 관계없이 고정적으로 단어를 임베딩하는 분산 표현 단계를 생략하였다. (b)컨텍스트 인코더 단계에서 BiLSTM과 KoBERT 모델을 각각 실행한 후, (c)태그 디코더 단계에서는 많은 딥러닝 기반의 개체명 인식 모델에서 사용하고 있는 CRF 모델을 공통으로 적용하였다.

본 실험에서는 제안하는 6개 단일 방법론과 이를 조합하여

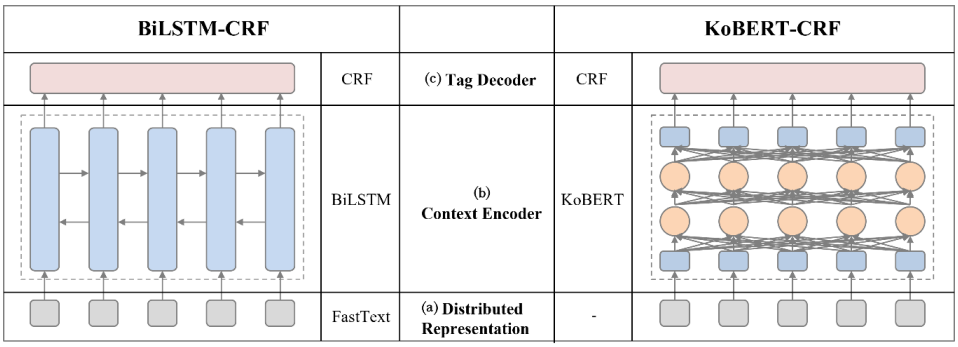

Figure 3. Overall Framework of the Backbone Model

Table 2. Number of Tags in the Training Dataset

Entity	Tag	Count	Entity	Tag	Count
Person	B-PER	31,610	Term	B-TRM	10,151
	I-PER	4,809		I-TRM	2,666
Date	B-DAT	5,121	Time	B-TIM	1,025
	I-DAT	1,499		I-TIM	436
Organization	B-ORG	23,695	Animal	B-ANM	3,052
	I-ORG	3,594		I-ANM	43
Civilization	B-CVL	39,597	Field	B-FLD	1,298
	I-CVL	2,995		I-FLD	40
Number	B-NUM	25,451	Material	B-MAT	191
	I-NUM	3,996		I-MAT	14
Location	B-LOC	12,060	Artifacts Works	B-AFW	3088
	I-LOV	191		I-AFW	1,485
Event	B-EVT	4,881	Plant	B-PLT	199
	B-EVT	3,288		I-PLT	3

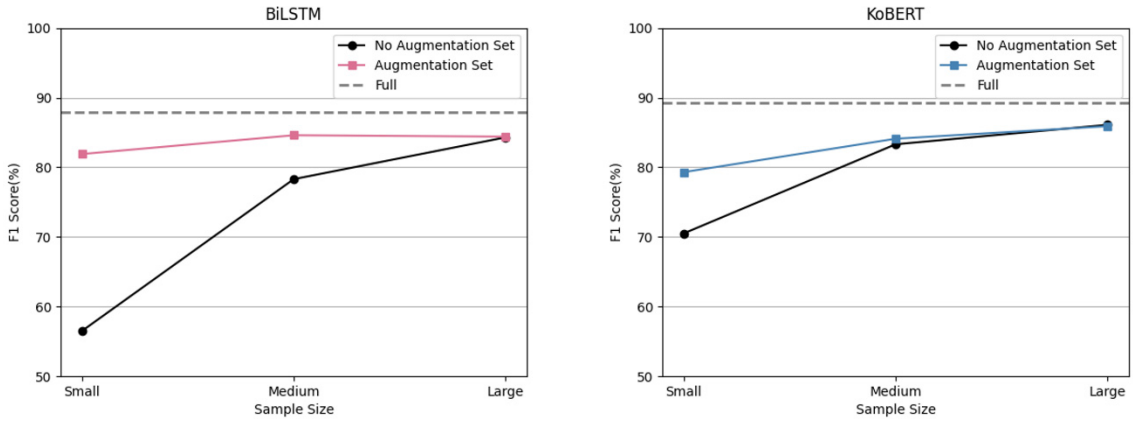

Figure 4. Average Performance of BiLSTM and KoBERT Models Applying Data Augmentation Methods over Three Different Data Sizes. We repeat all experiments five times with different random seeds.

사용한 복합 방법론을 BiLSTM과 KoBERT 모델에 각각 적용하고 그 성능을 확인하였다. 또한 적용 방법론이 최적의 성능을 도출하는데 필요한 데이터의 증강 갯수를 확인하기 위해서 원본 데이터 한 개당 {1, 2, 4, 8, 16}개의 데이터를 증가하여 그 효과를 비교하였다. 더불어 문장 내 토큰이 교체되는 비율이 성능에 영향을 미치는지 확인하기 위해서 {0.2, 0.4, 0.6, 0.8} 비율로 토큰을 교체하여 그 성능을 평가하였다.

다양한 개체명이 존재하는 개체명 인식에서는 모든 클래스, 즉 모든 태그의 성능을 평가해야 하기 때문에 일반적으로 F1 score를 평가지표로 사용한다. F1 score는 크게 micro, macro, weighted가 있다. 모든 클래스에 동일한 가중치를 부여하는 방법인 macro F1 score는 클래스가 균등하게 분포할 경우 사용하고, 샘플에 동일한 가중치를 주는 micro F1 score와 클래스에 가중치를 부여하는 weighted F1 score는 클래스가 불균등할 때 사용한다. <Table 2>는 본 실험에서 사용하는 전체 데이터셋의 태그 수이다. 표에서 보는 것과 같이 실험에 사용하는 데이터 셋의 태그 불균형이 심하여, 본 실험에서는 weighted F1 score 값으로 성능을 평가하였다.

4. 실험 결과

본 장에서는 네이버와 창원대가 함께하는 NLP Challenge 데이터에 적용한 증강기법의실험 결과를 포함하였다. 원본 데이터의 크기에 따른 성능을 확인하기 위해 데이터의 크기를 구분하고 (1)모델 (2)적용방법론 (3)데이터 증강 갯수 (4)데이터 교체 비율에 대한 개체명 인식의 성능 향상 정도를 확인하였다. 각 실험은 5회 반복하였다.

4.1 개체명 인식 모델에 따른 결과

<Figure 4>에서 보듯이, 사전 학습의 여부와 관계없이 컨텍스트 인코더에 적용한 BiLSTM과 KoBERT 모델 모두 본 연구에서 적용한 방법론으로 증강시킨 데이터를 학습데이터에 포함하였을 때 성능이 향상되었지만, 사전 학습 모델인 KoBERT보다 비사전 학습 모델인 BiLSTM을 사용하였을 때 더 큰 성능 향상 효과를 보였다. 데이터 셋이 작을 경우(N=500) BiLSTM 모델은 약 25%, KoBERT 모델은 약 7%의 성능 향상을 나타내었다. 데이터 셋이 작을수록 성능 향상의 폭은 더 컸다.

전체 데이터 셋(90,000개)에서 사용한 데이터의 비율은 작은 데이터 셋(500개)는 약 0.5%, 큰 데이터 셋(5,000개)는 약 5% 정도이다. 이는 전체 데이터셋과 비교하였을 때 매우 적은 양임에도, 전체 데이터 셋을 사용했을 때 대비 성능의 큰 차이가 없었다. BiLSTM 모델은 약 3%~5%, KoBERT 모델은 약 3%~12% 차이가 발생하였고, 특히 비사전 학습 모델인 BiLSTM 모델의 성능은 작은 데이터 셋에서도 전체 데이터 셋과 큰 차이가 나지 않음을 확인할 수 있었다. 이는 본 연구에서 제시한 증강 방법론을 적용하면, 적은 양의 데이터 셋에서도 큰 데이터 셋과 견줄만한 좋은 성능을 기대할 수 있다고 하겠다.

4.2 제안 방법론에 따른 결과

4.2.1 단일 방법론

본 연구에서 제시한 방법론을 각각 BiLSTM 기반 모델과 KoBERT 기반 모델에 적용하여 5회 반복 실험한 평균과 표준편차를 <Table 3>에서 보여 주었다. 밑줄 친 굵은 글씨는 제시한 방법론을 적용하지 않은 데이터 셋과 성능 차이를 의미한다. <Table 3>에서 확인할 수 있듯이, 적용한 증강 방법론 중 하나의 가장 우수한 방법론을 결정할 수는 없었다. 적용한 방법론은 한 문장 내에서 토큰을 일정한 비율로 교체, 삽입, 삭제, 순서 변경을 수행하였고, 학습데이터 또는 한국어 어휘의 미망에 있는 데이터를 이용하여 토큰을 변경하였다. 따라서 데이터 증강으로 인한 노이즈의 발생 비율과 변경되는 토큰이 각 방법론 별로 차이가 크지 않기 때문에 6가지 방법론 모두 데이터 셋의 크기, 모델 등 같은 조건에 대해서 비슷한 성능 향상을 보였다.

Table 3. Performance of BiLSTM and KoBERT Models by Using Data Augmentation Methods. N is the size of the dataset. We repeated all experiments five times with different random seeds. Mean values and standard deviations are reported. Bold/underline indicates the amount of performance changes compared to baseline model (without data augmentation)

Method	F1 Score (%)					
	BiLSTM			KoBERT		
	Small (N=500)	Medium (N=2000)	Large (N=5000)	Small (N=500)	Medium (N=2000)	Large (N=5000)
Without Data Augmentation	55.7 ± 3.2	78.0 ± 0.2	84.3 ± 0.2	70.5 ± 1.7	83.3 ± 0.3	86.1 ± 0.2
Label-wise Tag Replacement	81.0 ± 0.8 **(+25.3)**	83.7 ± 0.7 **(+5.7)**	84.4 ± 0.4 **(+0.1)**	79.5 ± 0.8 **(+9.0)**	84.4 ± 0.5 **(+1.1)**	86.3 ± 0.3 **(+0.2)**
Label-wise Mention Replacement	81.8 ± 0.5 **(+26.1)**	85.3 ± 0.9 **(+7.3)**	83.3 ± 0.3 **(-1.0)**	78.5 ± 0.5 **(+8.0)**	83.3 ± 0.3 **(-0.0)**	85.8 ± 0.2 **(-0.3)**
Synonym Replacement	81.9 ± 0.6 **(+26.2)**	85.0 ± 0.4 **(+7.0)**	84.5 ± 0.3 **(+0.2)**	80.2 ± 0.5 **(+9.7)**	83.9 ± 0.7 **(+0.6)**	85.3 ± 0.3 **(-0.8)**
Random Insertion	82.5 ± 0.6 **(+26.8)**	84.5 ± 0.4 **(+6.5)**	84.7 ± 0.2 **(+0.4)**	76.7 ± 0.4 **(+6.2)**	83.4 ± 0.5 **(+0.1)**	85.5 ± 0.2 **(-0.6)**
Random Swap	81.9 ± 1.0 **(+26.2)**	84.1 ± 0.4 **(+6.1)**	84.4 ± 0.3 **(+0.1)**	77.0 ± 0.8 **(+6.5)**	81.1 ± 0.4 **(-2.2)**	84.2 ± 0.2 **(-1.9)**
Random Deletion	81.5 ± 0.5 **(+25.8)**	84.7 ± 0.3 **(+6.7)**	84.8 ± 0.3 **(+0.5)**	76.5 ± 0.4 **(+6.0)**	82.5 ± 0.4 **(-0.8)**	85.3 ± 0.2 **(-0.8)**

4.2.2 복합 방법론

본 연구에서 제시한 단일 방법론은 토큰의 교체(동일 태그 교체, 동일 개체명 교체, 동일 유의어 교체), 삽입(임의 삽입), 삭제(임의 삭제), 순서 변경(임의 교체)로 구분된다. 이 구분에 따라 제시한 단일 방법론을 임의로 선택, 조합하여 성능을 확인하였다. 즉, {교체+삽입, 교체+삭제, 교체+순서변경, 삽입+삭제, 삽입+순서교체, 삭제+순서교체, 교체+삽입+삭제, 교체+삽입+순서변경, 교체+삭제+순서변경, 삽입+삭제+순서변경, 교체+삽입+삭제+순서교체}와 같이 방법론을 임의로 조합하여 성능을 측정하였다. <Figure 5>의 가로축은 조합한 단일 방법론 갯수를 의미하고, 세로축은 각 성능의 평균을 의미한다. 조합한 방법론을 사전 훈련되지 않은 BiLSTM 모델에 적용하면, 단일 방법론만 적용하였을 때와 비교하여 성능의 차이가 거의 없었다. 반면, 조합된 방법론을 사전 훈련된 모델인 KoBERT 모델에 적용하면, 성능이 약 10% 하강하였다. 방법론을 조합하여 사용할 경우 문장의 컨텍스트 정보를 훼손할 가능성이 크기 때문에 문맥에 의존하는 특징을 가지는 KoBERT가 문맥에 의존하지 않는 단어 임베딩을 사용한 BiLSTM 모델에 비해 효과가 많이 감소하였다.

4.3 데이터 증강 갯수에 따른 결과

데이터를 어느 정도 증강했을 때 최적의 성능이 측정되는지를 파악하기 위해서, 각 문장당 데이터를 {1, 2, 4, 8, 16}개 만큼 증가하였다. <Figure 6>에서 보듯이 원본 데이터만 사용하였을 때보다, 원본 데이터를 1배수 혹은 2배수로 증강하면, 성능이 크게 향상되었다. 이는 데이터 셋이 작을 경우 과적합 가능성이 크지만, 데이터 증강으로 발생한 노이즈가 이러한 현

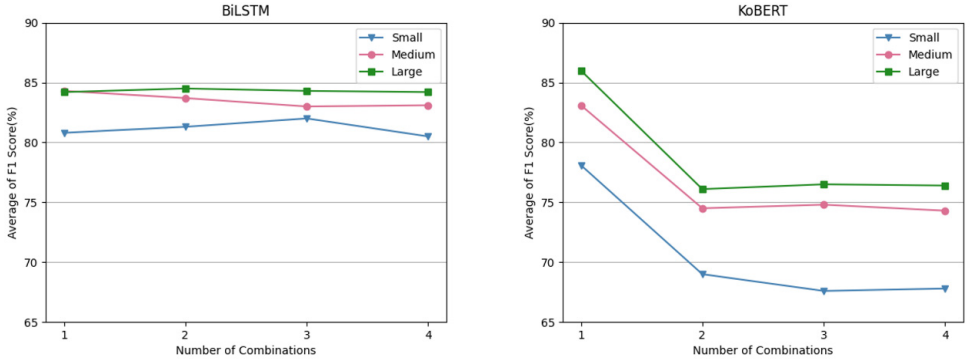

Figure 5. Performance of BiLSTM and KoBERT Models by Number of Method Combinations. The size of each dataset of small / medium / large is 500 / 2000 / 5000

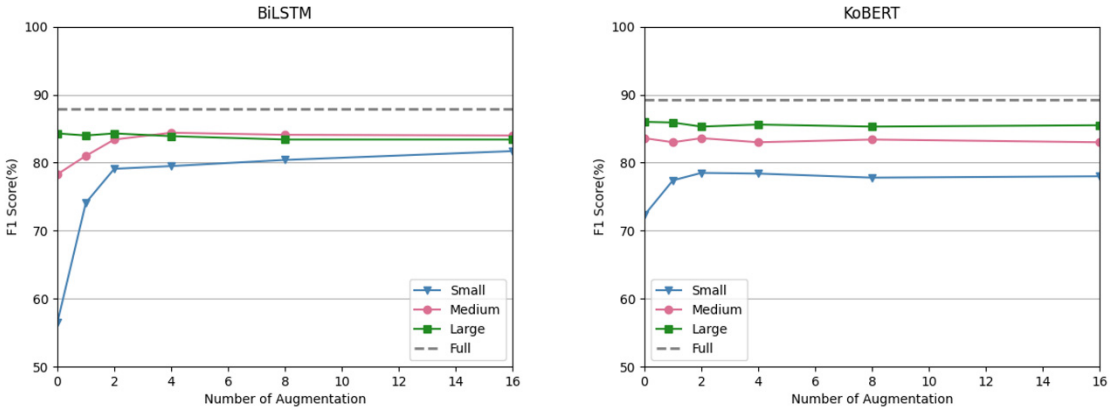

Figure 6. Performance of BiLSTM and KoBERT Models by Number of Data Augmentation. The size of each dataset of small / medium / large is 500 / 2000 / 5000. The dotted line represents the performance of the entire data set without data augmentation. We repeat all experiments five times with different random seeds

상을 보안하였기 때문이다. 반면 데이터가 많아질수록 노이즈의 영향이 줄어들고 모델이 일반화되기 때문에 성능 향상이 미비해졌다. 또한 훈련 데이터 크기가 BiLSTM 기반 모델의 경우 약 5,000개, KoBERT 기반 모델의 약 2,000개 이하일 때 데이터 증강 기법으로 인한 성능 향상을 기대할 수 있었다.

4.4 데이터 교체 갯수에 따른 결과

제시한 증강 방법론 중 동일 태그 교체, 동일 개체명 교체, 유의어 교체의 경우 한 개 이상의 태그 혹은 개체명을 갖은 토큰들이 변경되는 과정이다. 한 문장에서 변경되는 토큰 비율을 α라고 하고, 한 문장의 토큰의 갯수를 l이라고 한다면, 한 문장에서 변경되는 토큰의 개수 c는 아래와 같은 식으로 표현될 수 있다.

$$c = \max(1, \alpha l) \tag{1}$$

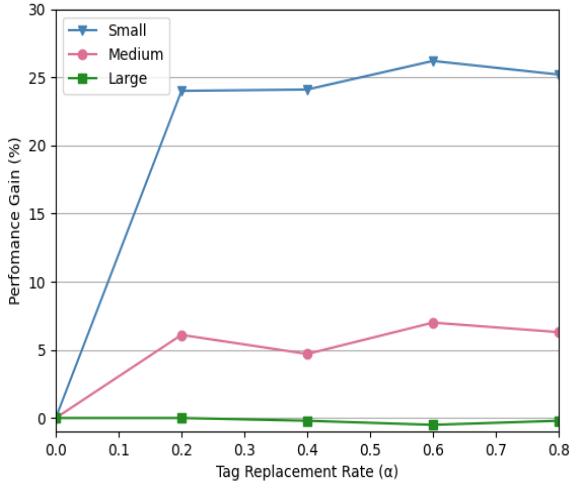

Figure 7. Performance According to the Tag Replacement Rate (α) of a Sentence

위 식에서 주목할 점은, 무조건 1개의 토큰을 교체한다는 점이다. BIO 태그 기반 개체명 인식 데이터에서 의미가 있는 B, I 태그 보다, 개체명이 지정되지 않은 O의 비중이 상당히 높다. B, I 태그가 적은 문장이 많은 데이터 셋이라면, 데이터 증강 기법을 적용하여도 원본 데이터와 같은 데이터가 발생하여 성능 향상이 미비할 수 있다. 따라서 다양한 데이터 증강을 위해 무조건 한 개 데이터를 교체하였다. 예를 들면, 한 문장의 토큰 수가 10개이고, O 태그를 제외한 토큰이 5개라고 할 경우 교체율이 0.2라면 전체 토큰 중 총 한 개의 토큰을 교체할 수 있다. 반면 O 태그를 제외한, B 또는 I 토큰이 한 문장에서 두 개일 경우 교체 토큰 수는 한 개 미만이지만, 무조건 한 개의 토큰을 교체하여 새로운 데이터를 생성한다. 토큰 변경 비율(α)을 {0.2, 0.4, 0.6, 0.8} 의 경우로 나누어 실험한 결과인 <Figure 7>에서 보듯이, 적은 양의 데이터 셋과 중간 크기 데이터 셋에서는 변경 비율이 0.6일 때 성능이 가장 높았고, 데이터 셋의 크기가 커질수록 한 문장에서 태그가 교체되는 비율이 성능에 의미 있는 영향을 미치지는 않음을 확인하였다. 전반적으로 토큰 변경 비율은 성능에 큰 영향을 미치지는 않았다.

5. 결 론

개체명 인식은 정보 추출 분야에서 필수적인 요소이고, 개체명 인식의 성능이 정보검색, 질의응답 같은 상위 시스템 성능에 큰 영향을 미친다는 점에서 매우 중요하다. 딥러닝을 기반으로 하는 개체명 인식 시스템에서 양질의 데이터 확보는 모델의 일반화와 최고의 성능을 얻기 위한 기본 조건이다. 하지만 한글 데이터는 시간과 비용 등의 제약사항으로 데이터를 확보하는 데 어려움이 있다. 데이터 증강 기법을 적용하여 데이터 수를 늘려서, 성능을 향상하는 연구가 다양하게 진행되고 있지만, 한글 데이터셋을 이용한 개체명 인식 분야에서는 거의 연구가 진행되지 않았다. 본 연구에서는 한글 개체명 인

식 데이터 셋에 적용할 수 있는 다양한 데이터 증강 기법의 사용을 제안하였다. 특히 본 연구를 통해 적은 양의 데이터 셋에서 제안 증강 기법을 적용한 데이터로 학습하였을 때, 큰 데이터셋을 사용한 것만큼 성능이 향상될 수 있음을 확인하였다.

다만 연구의 몇 가지 한계점도 존재한다. 먼저 데이터 셋이 충분할 경우 성능 향상이 미비하였으며, 데이터를 증강하더라도 일정 크기 이상의 훈련데이터에서는 유의미한 성능 향상이 없었다. 또한 사전 훈련된 모델을 사용할 경우에 사전 훈련되지 않은 모델보다 큰 성능 향상을 보이지 않았다. 실험에서 사전 훈련된 모델은 작은 데이터 셋에서만 약간의 성능 향상을 보였고, 2,000개 이상의 데이터 셋에서는 데이터 증강에 대한 효과가 나타나지 않았다.

본 연구는 한글 개체명 인식 분야에서 고성능이 필요한 사전 학습 모델이나 외부 사전을 구축하는 작업 없이 한글 개체명 인식 분야에 데이터 증강 기법을 도입한 연구라는 점에서 의의가 있다. 또한, 영어와 언어 구조적 차이를 극복하고 띄어쓰기 기반의 토큰화를 이용하여, 영어 데이터 셋 기반의 데이터 증강 기법을 한글에 적용한 연구라는 점에서 의의를 찾을 수 있다. 본 연구에서 제안하는 방법론은 의료, 금융과 같이 전문성이 필요하고, 충분한 데이터 확보가 어려운 특정 도메인에 대해서도 제약 없이 적용할 수 있는 방법론이라는 점에서 시사하는 바가 크다. 이번 연구를 통해서 한글 개체명 인식이 도메인에 상관없이 일반적으로 성능을 확보하는 데 큰 도움이 될 수 있을 것으로 본다. 또한 데이터 확보의 어려움 때문에 개체명 인식 연구가 잘 이루어지지 않은 다양한 산업 분야에서도 이바지할 수 있을 것으로 기대된다.

참고문헌

An, J. and Kim, H.-W. (2015), Building a Korean Sentiment Lexicon Using Collective Intelligence, *Journal of Intelligence and Information Systems*, **21**(2), 49-67.

Collobert, R., Weston, J., Com, J., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011), Natural Language Processing (Almost) from Scratch, In *Journal of Machine Learning Research*, **12**, 2493-2537.

Dai, X. and Adel, H. (2020), An analysis of simple data augmentation for named entity recognition, ArXiv Preprint ArXiv:2010.11683.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018), Understanding Back-Translation at Scale, ArXiv Preprint ArXiv:1808.09381.

Lecun, Y., Bengio, Y., and Hinton, G. (2015), Deep learning, In *Nature*, **521**(7553), 436-444, Nature Publishing Group.

Lee, D., Yu, W., and Lim, H. (2017), Bi-directional LSTM-CNN-CRF for Korean Named Entity Recognition System with Feature Augmentation, *Journal of the Korea Convergence Society*, **8**(12), 55-62.

Lee, S. H., Jang, D. P., and Sung, K. K. (2020), Donguibogam-based pattern diagnosis using natural language processing and machine learning, *Journal of Korean Medicine*, **41**(3), 1-8.

Li, J., Sun, A., Han, J., & Li, C. (2020), A survey on deep learning for named entity recognition, *IEEE Transactions on Knowledge and Data Engineering*, 1-1.

Nadeau, D. and Sekine, S. (2007), A survey of named entity recognition and classification, *Lingvisticae Investigationes*, **30**(1), 3-26.

Shorten, C. and Khoshgoftaar, T. M. (2019), A survey on image data augmentation for deep learning, *Journal of Big Data*, **6**(1).

Song, C. H., Hltcoe, D. L., Finin, T., and Mayfield, J. (2020), Improving neural named entity recognition with gazetteers, ArXiv Preprint ArXiv:2003.03072.

Van Engelen, J. E. and Hoos, H. H. (2020), A survey on semi-supervised learning, *Machine Learning*, **109**(2), 373-440.

Wei, J. and Zou, K. (2019), EDA: Easy data augmentation techniques for boosting performance on text classification tasks, ArXiv Preprint ArXiv:1901.11196.

Wu, X., Lv, S., Zang, L., Han, J., and Hu, S. (2018), Conditional BERT contextual augmentation, In *International Conference on Computational Science*, Springer, Cham, 84-95.

저자소개

조경선 : 고려대학교 산업경영공학과에서 2012년 학사 학위를 취득하고 동 대학원에서 석사 과정에 재학 중이다. 연구분야는 Deep Learning for Natural Language Processing, Anomaly Detection on Multivariate Time Series이다.

김성범 : 고려대학교 산업경영공학부 교수로 2009년부터 재직하고 있으며, 인공지능공학연구소 소장 및 기업산학연협력센터 센터장을 맡고 있다. 미국 텍사스주립대학교 산업공학과 교수를 역임하였으며, 한양대학교 산업공학과에서 학사학위를 미국 Georgia Institute of Technology에서 산업공학 석사 및 박사학위를 취득하였다. 연구 분야는 인공지능, 머신러닝, 최적화이다.