ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 46, No. 6, pp. 544-553, 2019. 6
https://doi.org/10.5626/JOK.2019.46.6.544

멀티헤드 주의집중 기법과 하이웨이 네트워크를 활용한 생물학 개체명 인식

(Biomedical Named Entity Recognition using Multi-head Attention with Highway Network)

조 민 수 † (Minsoo Cho)　박 진 욱 † (Jinuk Park)　하 지 환 † (Jihwan Ha)　박 찬 희 †† (Chanhee Park)　박 상 현 ††† (Sanghyun Park)

요 약 생물학 개체명 인식이란 생물학 문헌으로부터 질병, 유전자, 단백질과 같은 생물학 개체명을 추출하고 그 종류를 분류하는 작업으로, 생물학 데이터로부터 유의미한 정보를 추출하는데 중요한 역할을 한다. 본 연구에서는 입력 단어의 자질을 자동으로 추출할 수 있는 딥러닝 기반의 Bi-LSTM-CRF 모델을 활용한 개체명 인식 연구를 진행하였다. Multi-head 주의 기제 기법을 적용하여 입력 단어들 간의 관계를 포착하고 관련성이 높은 단어에 주목하여 예측의 성능을 높였다. 또한, 단어 단위 임베딩 벡터 외 문자 단위 임베딩 벡터를 결합하여 입력 임베딩의 표상을 확장하고, 각 표상의 정보 흐름을 학습하기 위해 Highway 네트워크에 적용하였다. 제안하는 모델의 성능을 평가하기 위해 두 개의 영어 생물학 데이터셋으로 비교 실험을 진행하였으며, 그 결과 기존 연구의 모델들보다 향상된 성능을 보였다. 이를 통해 제안하는 방법론이 생물학 개체명 인식 연구에서 효과적인 방법론임을 입증하였다.

키워드: 정보 추출, 자연어 처리, 개체명 인식, Multi-head 주의 기제 기법, Highway 네트워크, 단어 임베딩

Abstract Biomedical named entity recognition(BioNER) is the process of extracting biomedical entities such as diseases, genes, proteins, and chemicals from biomedical literature. BioNER is an indispensable technique for the extraction of meaningful data from biomedical domains. The proposed model employs deep learning based Bi-LSTM-CRF model which eliminates the need for hand-crafted feature engineering. Additionally, the model contains multi-head attention to capture the relevance between words, which is used when predicting the label of each input token. Also, in the input embedding layer, the model integrates character-level embedding with word-level embedding and applies the combined word embedding into the highway network to adaptively carry each embedding to the input of the Bi-LSTM model. Two English biomedical benchmark datasets were employed in the present research to evaluate the level of performance. The proposed model resulted in higher f1-score compared to other previously studied models. The results demonstrate the effectiveness of the proposed methods in biomedical named entity recognition study.

Keywords: information retrieval, natural language processing, named entity recognition, multi-head attention mechanism, highway network, word embedding

· 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(IITP-2017-0-00477, (SW스타랩) IoT 환경을 위한 고성능 플래시 메모리 스토리지 기반 인메모리 분산 DBMS 연구개발)

† 비 회 원 : 연세대학교 컴퓨터과학과
minsoo0104@yonsei.ac.kr
parkju536@yonsei.ac.kr
jihwanha@yonsei.ac.kr
†† 학생회원 : 연세대학교 컴퓨터과학과
channy_12@yonsei.ac.kr
††† 종신회원 : 연세대학교 컴퓨터과학과 교수(Yonsei Univ.)
sanghyun@yonsei.ac.kr
(Corresponding author임)

논문접수 : 2018년 12월 5일
(Received 5 December 2018)
논문수정 : 2019년 3월 8일
(Revised 8 March 2019)
심사완료 : 2019년 3월 18일
(Accepted 18 March 2019)

정보과학회논문지 제46권 제6호(2019. 6)

1. 서 론

개체명 인식(Named entity recognition)이란 문서에서 인명, 지명, 기관명 등과 같은 개체명을 인식하여 해당 개체를 추출하고, 추출된 개체명의 종류를 분류하는 작업을 말한다. 개체명 인식은 자연어처리 분야의 정보검색, 정보 추출, 질의응답 등에 핵심 요소로 활용되고 있으며 관련 연구 또한 활발하게 진행되고 있다.

생물학 분야에서도 생물 의학 정보처리의 기초 단계로 개체명 인식이 사용되고 있다. 생물 의학 분야의 급속한 발전으로 생물 의학 데이터의 수가 기하급수적으로 증가하고 있다. 이러한 방대한 양의 생물 의학 데이터에서 보다 의미 있는 정보를 효율적으로 추출하기 위해 생물정보학의 역할이 더욱 중요해지고 있으며, 그 방법론 중 하나로 개체명 인식이 활용되고 있다. 생물학 개체명 인식은 생물학 문헌으로부터 화학 물질, 질병, 유전자, 단백질과 같은 생물학 개체명을 추출하는 것으로, 생물학 용어 자체의 불규칙한 표기법, 문자와 숫자로 구성된 접두사/접미사, 약어간의 모호성과 같은 몇 가지 까다로운 특징으로 인해 일반 개체명 인식보다 낮은 성능을 보이고 있다. 영어권에서는 이러한 생물학 개체명 인식의 낮은 성능을 향상시키고자 다양한 영어 데이터셋을 활용한 연구가 발표되고 있다.

본 연구에서는 딥러닝(Deep learning)을 활용한 Bi-LSTM(Bidirectional long short term memory) 인공신경망 모델[1]과 전통적으로 사용되었던 CRF(Conditional random fields)[2] 모델을 결합한 Bi-LSTM-CRF 모델을 활용한다[3]. RNN(Recurrent neural network), LSTM과 같은 순환 신경망의 경우, 입력 문장의 길이가 길어지면 앞에서 학습한 정보가 희석되어 단어들 사이의 관계 정보가 손실되는 문제점이 발생할 수 있다. 이와 같은 한계점을 극복하기 위해, 본 연구에서는 거리가 먼 단어에 대해서도 단어 간의 관계를 포착할 수 있는 Multi-head 주의 기제(Attention) 기법[4]을 적용한다. 주의 기제 기법을 적용함으로써, 중요도가 높은 단어에 대해서는 중요도를 강조하고, 중요도가 낮은 단어에 대해서는 억제하여 개체명 예측의 성능을 향상시켰다.

일반적으로 Bi-LSTM 모델에서는 모델의 입력으로 단어 단위의 임베딩 벡터(Word-level embedding vector)를 사용하는데, 단어 단위 임베딩 방법의 대표적인 문제점은 임베딩 사전에 없는 미등록 단어(Out-of-vocabulary)가 등장하게 되면 이를 모두 <UNK> 토큰으로 처리하여 단어의 의미를 제대로 반영할 수 없다는 것이다. 본 연구에서는 이러한 단어 단위 임베딩의 단점을 보완하기 위해 단어 단위(Word-level)와 문자 단위(Character-level) 임베딩 벡터를 모두 사용하는 방법론[5,6]을 모델에 적용하였다. 문자 단위의 표상 방법으로부터 단어 벡터를 형성하기 위해, CNN(Convolutional neural network)[7]과 LSTM 방법을 모두 사용하여, 단어의 길이에 제약 없이 지역적 자질(feature)과 전역적 자질을 모두 추출할 수 있도록 하였다. 또한, 제안하는 모델에서는 단어 단위와 문자 단위를 모두 결합한 임베딩 벡터를 Highway 네트워크[8]에 적용하였다. 이를 통해 활성화 연산을 수행하는 정보와 수행하지 않는 정보를 모두 활용하여, 단어 정보의 흐름을 효율적으로 학습시켰으며, 단어의 의미를 보다 정확하게 내포하는 임베딩 벡터의 생성으로 모델의 성능 향상을 도모하였다.

본 연구에서는 생물학 개체명 인식 모델의 성능을 개선하고자 다음과 같은 방법론을 제안한다. 첫째, 단어 표상을 확장하는 방법으로 CNN과 LSTM으로 생성된 문자 단위 임베딩 벡터를 단어 단위 임베딩 벡터와 결합하여 Highway 네트워크에 적용한다. 둘째, Multi-head 주의 기제 기법을 적용하여 단어들 간의 관련성 정보를 구하고, 이를 각 단어의 레이블을 예측하는데 함께 사용한다.

본 논문의 구성은 다음과 같다. 2장에서는 개체명 인식의 방법과 선행연구에 대해 소개하고, 3장에서는 논문에서 사용하는 기본 구조와 제안하는 모델의 전반적인 방법론을 소개한다. 4장에서는 실험결과를 설명하고, 마지막 5장에서는 결론 및 향후 계획을 기술한다.

2. 관련 연구

생물학 개체명 인식 연구는 영어권에서 처음 시작되어, 현재까지도 다양한 기법으로 연구되고 있다. 그 기법으로 크게 사전 기반[9], 규칙 기반[10], 기계학습 기반[11] 방법이 있으며, 최근에는 딥러닝을 활용한 연구가 활발하게 이루어지고 있다.

사전 기반 방법은 개체명 사전에 등록된 특정 개체명이 문서에 포함되어 있는 경우에 개체를 추출하는 방식으로 가장 간단하게 적용할 수 있는 방법이다. 하지만, 생물학 용어의 특성상 존재하는 개체명의 불규칙성과 까다로운 표기법으로 인해 개체명 사전만으로 개체를 추출하는데 한계가 존재한다. 규칙 기반 방법은 사전에 정의해놓은 정규 표현식과 사전 탐색 방법을 이용하여 개체를 추출하는 방식으로 좋은 패턴을 찾는 것이 성능을 크게 좌우한다. 기계 학습 방법으로는 HMM(Hidden markov model)[12], MEM(Maximum entropy model)[13], SVM(Support vector machine)[14] 그리고 CRF을 이용한 방법이 존재하며, 그 중 예측 레이블 간 인접성 정보를 활용하는 CRF가 가장 좋은 성능을 보여 딥러닝 모델과 결합되어 사용되고 있다. 딥러닝 방법은 이전 방법들과 다르게 좋은 자질을 선택하기 위한 전문가의 지식이 요구되지 않으며, 학습에 의해 좋은 자질이

자동으로 선택 및 추출되는 장점을 가지고 있다. 최근 생물학 개체명 인식에서 가장 좋은 성능을 보이고 있는 선행 연구 모두 딥러닝 모델을 활용한 연구로 LSTM과 CNN 구조를 확장한 모델을 적용하고 있다.

[3]은 양방향 LSTM과 CRF를 결합한 최초의 연구로, 전통적인 기계학습 기법들과 대비하여 우수한 성능을 보여, 이를 확장한 후속 연구가 계속해서 진행되고 있다. 이후 연구[5,6]에서는 Bi-LSTM-CRF 모델에 단어 임베딩에서 미등록 문제의 한계를 극복하기 위해 문자 단위 임베딩 벡터를 사용하여 단어 표상을 확장시켰으며, 또 다른 연구[15]에서는 단어 표상 확장 방법으로 품사 정보와 약어 여부 정보를 포함시켰다. 한국어 개체명 인식 연구에서도 단어 표상을 확장시키기 위한 방법으로, 음절 단위 임베딩 벡터와 품사 임베딩 벡터를 사용하여 성능 향상을 보인 연구가 있다[16]. CNN 구조를 사용한 선행연구로는 [17]이 있으며, 여러 커널 사이즈를 사용하여 단어의 지역 정보를 추출함으로써 기존 양방향 LSTM 구조와 차별점을 두었다. 이 외에도, 본 연구에서 적용하는 주의 기제 기법에 관한 다양한 선행 연구 [18,19]가 있으며, 학습 데이터가 부족할 때 다른 데이터로 추출된 지식과 모델을 활용한 전이학습(Transfer Learning) 기반의 생물학 개체명 인식 선행 연구가 있다[20].

본 연구에서는 미등록 단어 문제를 해결하고자, 연구[5,6]의 방법론을 확장하여 각 단어를 단어 단위 임베딩 벡터와 문자 단위 임베딩 벡터로 구성하였으며, 문자 단위 임베딩으로는 LSTM과 CNN으로 생성된 두 개의 벡터를 모두 사용하였다. 또한, 각 단어의 임베딩 벡터를 Highway 네트워크에 적용하여 단어의 최종 표상을 나타내었다. 추가적으로, 본 연구에서는 입력 문장의 길이가 길어질 때 발생하는 정보 희석 문제를 해결하고자, 이전 다국어 개체명 인식 연구[19]에서 효용성이 입증된 Multi-head 주의 기제 기법을 Bi-LSTM의 은닉 계층(Hidden states)에 적용한다.

3. 개체명 인식 모델

본 장에서는 모델의 기본 구조인 Bi-LSTM-CRF 모델을 설명하고, 모델에 적용되는 단어의 임베딩 방법, Highway 네트워크, Multi-head 주의 기제 기법에 대한 방법론을 각 절에서 자세하게 기술한다.

3.1 Bi-LSTM-CRF 모델

LSTM은 RNN의 변형 구조를 지닌 신경망 모델로 기계번역, 이미지 캡션, 언어 모델링과 같은 다양한 자연언어처리 분야에서 활용되고 있다. RNN은 그림 1과 같이 과거의 정보를 기억하고, 그 정보를 다음 단계의 학습에 반영할 수 있는 순환적인 구조를 가지고 있어, 순차적인 데이터를 학습하는데 유용하다. 하지만 RNN

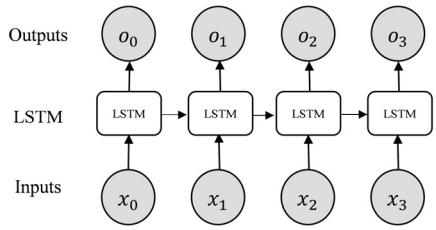


그림 1 Long Short Term Memory 모델
Fig 1. Long Short Term Memory Model

의 입력정보가 길어질 경우, 관련 정보 간의 거리가 멀어져 장기 의존성(Long term dependency)이 떨어지는 문제가 발생하게 된다. 이러한 RNN의 장기 의존성 문제를 완화하고자 제안된 모델이 LSTM이다. LSTM은 RNN의 은닉 계층에 셀 스테이트(Cell state)와 입력(Input), 망각(Forget), 출력(Output) 게이트(Gate) 요소를 활용하여 정보들이 선택적으로 흘러갈 수 있도록 구성해 정보를 더 오래 기억하도록 하였다. 새로운 정보가 들어왔을 때, 망각 게이트는 이전 상태의 어떠한 정보를 저장할지 결정하며, 입력 게이트는 새로운 정보를 얼마나 반영할지 결정한다. 두 개의 게이트는 셀 스테이트를 업데이트 하는데 사용되며, 출력 게이트는 업데이트된 셀의 출력 값을 제어하여 어떤 값을 출력할지 결정한다.

위에서 언급한 그림 1의 순방향 LSTM 구조는 순차적으로 학습하는 선행학습만 가능하지만, 양방향 LSTM의 경우 선행 학습과 역순으로 학습하는 후행 학습을 병행할 수 있다. 따라서, 본 모델에서는 그림 2와 같이 입력 $X=\{x_1,x_2,x_3,...x_t\}$에 대하여 순방향의 은닉 계층 $\overrightarrow{h_t}$와 역방향의 은닉 계층 $\overleftarrow{h_t}$을 연결한 은닉 층 $h_t=[\overrightarrow{h_t};\overleftarrow{h_t}]$를 각 단어의 데이터 표현형으로 사용하여 양방향의 은닉 정보를 모두 반영하였다.

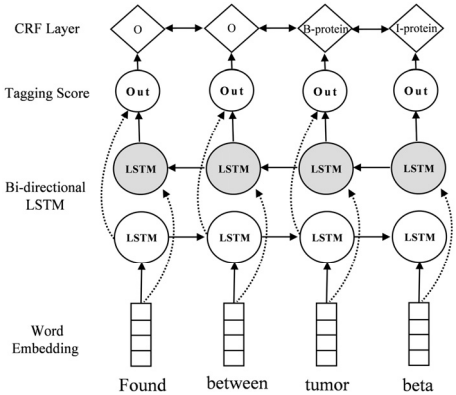


그림 2 Bi-LSTM-CRF 모델
Fig. 2 Bi-LSTM-CRF Model

그 위에 CRF 모델을 결합하여 예측 레이블 간의 인접성 정보 $T_{y_i y_{i+1}}$을 포착하고 이를 바탕으로 최적의 상태열을 추측하도록 하였다. 입력 $X=\{x_1,x_2,x_3,...x_t\}$와 예측 레이블 $y=\{y_1,y_2,y_3,...y_t\}$에 대한 스코어(Score) 함수는 식 (1)과 같이 계산된다.

$$s(X,y)=\sum_{i=0}^{t}(T_{y_i y_{i+1}}+M(S)_{i,y_i}) \quad (1)$$

식 (1)에서 각 문장의 최종 스코어 $s(X,y)$는 레이블 y_i에서 y_{i+1}로의 전이 행렬(Transition matrix) $T_{y_i y_{i+1}}$와 Bi-LSTM의 은닉 계층을 완전연결계층에 통과시켜 나온 i번째 단어에 대한 레이블 y_i의 태깅 스코어(Tagging score) $M(S)_{i,y_i}$을 합하여 계산된다.

$$p(y|X)=\frac{\exp(s(X,y))}{\sum_{y'\in Y(x)}\exp(s(X,y'))} \quad (2)$$

$$loss=-\log p(y|X) \quad (3)$$

이후, 식 (2)에서는 모든 가능한 상태열 $Y(x)$의 최종 스코어에 소프트맥스(softmax) 함수를 적용하여 각 상태열의 확률 분포 $p(y|X)$를 구한다. 식 (3)에서는 $p(y|X)$에 음의 로그 함수(negative log-likelihood)를 취해 로그 값을 최소화하는 방향으로 모델을 학습시킨다.

$$\hat{y}=\arg max_{y'\in Y(x)}s(X,y') \quad (4)$$

모델을 학습한 후, 디코딩을 할 때는 식 (4)와 같이 입력 X에 대하여 $s(X,y)$의 확률 값을 최대로 하는 $\hat{y}$를 선택한다.

3.2 입력 임베딩

3.2.1 단어 단위 임베딩

하나의 단어를 특정 차원의 벡터로 수치화하는 방법을 단어 임베딩(Word embedding) 또는 분산 표현(Distributed representation)이라하며, 이는 방대한 양의 문자 데이터를 다양한 신경망 모델에서 학습하기 위해 사용된다[21]. 단어 임베딩을 통해, 단어의 의미가 비슷할수록 높은 유사도를 나타내는 벡터로 표현할 수 있다. 대표적인 임베딩 모델링 방법으로는 Word2vec, Glove, Fasttext가 있다. 본 연구에서는 생물학 데이터에서만 등장하는 생소한 용어의 의미를 보다 정확하게 표현하기 위해 BioASQ[22]에서 제공하는 대량의 생물학 말뭉치(Corpus)를 Word2vec방법으로 학습한 단어 임베딩 벡터(Pre-trained word embedding vector)를 모델의 입력으로 사용하였다.

3.2.2 문자 단위 임베딩

문자 단위 임베딩은 미등록 단어 또는 자주 등장하지 않는 단어에 대해서도 단어 구조 정보를 활용할 수 있어 다양한 신경망 기반의 연구에서 모델의 성능 향상을 보였다[5,6]. 본 연구에서는 문자 단위의 단어 표상 방법으로 CNN과 LSTM을 사용하여, CNN으로 인접 문자열 간의 지역적인 특성을 추출하고, LSTM을 통한 문자열의 순서정보와 전역적인 특성을 학습하도록 하였다. 그림 3과 같이 각 모델에서 생성된 두 종류의 문자 단위 임베딩 벡터와 단어 단위 임베딩 벡터를 결합(Concatenate)하였으며, 이를 완전연결계층(Fully-connected layer)에 통과시켜준 임베딩 벡터를 Highway 네트워크의 입력으로 사용하였다.

3.3 Highway Network

Highway 네트워크는 심층 모델에서 신경망의 깊이가 깊어질수록 발생하기 쉬운 그래디언트 소실(Gradient vanishing) 문제를 해결하기 위해 제시된 방법[8]으로, 해당 층에서의 선형 연산과 활성화 연산을 수행하지 않는 우회로를 제공하여 심층 신경망에서도 빠르게 학습할 수 있도록 하였다. 변경 게이트(Transform gate)와 캐리 게이트(Carry gate), 이 두 개의 게이트를 이용하여 모델의 층을 얼마나 거칠지 또는 얼마나 연산 없이 통과할지를 학습한다.

실제로 한국어 개체명 인식 연구 중 Highway 네트워크를 그래디언트 소실 문제를 해결하기 위해 Bi-LSTM 레이어에 적용한 연구[23]가 있다. 본 연구에서는 Highway 네트워크를 Bi-LSTM 레이어가 아닌 입력 레이에 적용하여 입력 단어의 표상을 학습하는데 활용한다. 그림 4와 같이 입력 임베딩을 Highway 네트워크에 적용함으로써,

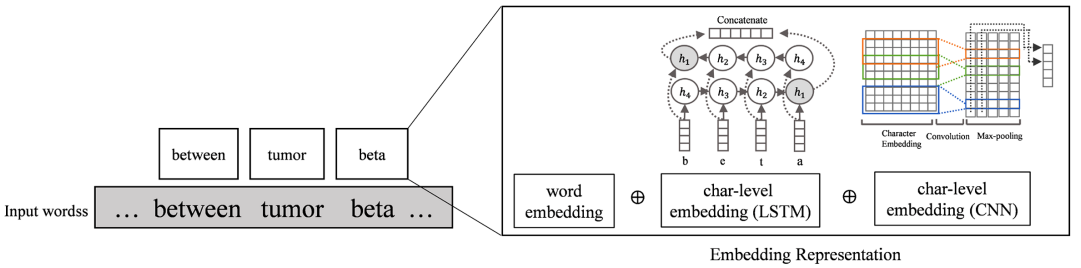


그림 3 문자 단위와 단어 단위 임베딩 벡터의 결합

Fig. 3 Integration of Word and Character level Embedding Vector

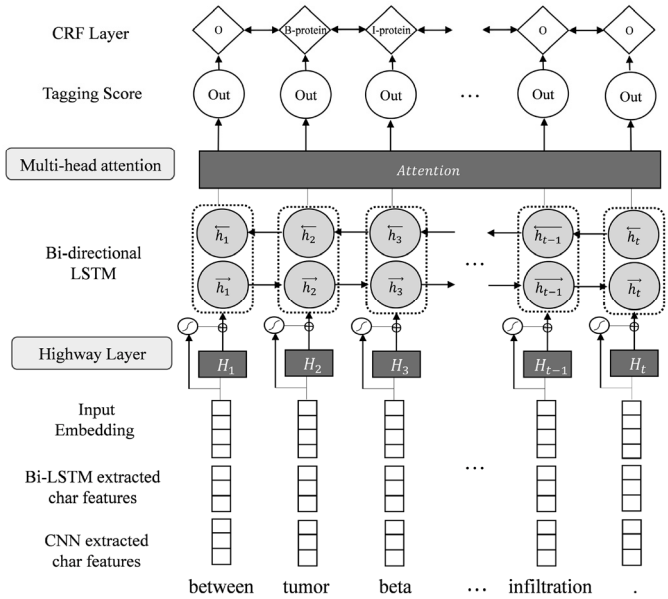


그림 4 주의 주제 기반의 Highway Bi-LSTM-CRF
Fig. 4 Attention-based Highway Bi-LSTM-CRF

문자와 단어 단위로 구성된 임베딩 벡터를 그대로 통과시킬지 혹은 활성화 연산을 거쳐 변환할지를 학습하도록 한다. 정보의 흐름을 학습함으로써, 단어의 의미를 보다 정확하게 내포하는 임베딩 벡터를 추출할 수 있도록 한다.

$$z = t \odot g(W_H y + b_H) + (1-t) \odot y \quad (5)$$

$$t = \sigma(W_T y + b_T) \quad (6)$$

식 (5)에서 g는 비선형 연산(ReLU), t는 변형 게이트, $(1-t)$는 케리 게이트를 나타내며, y는 문자 단위와 단어 단위 임베딩 벡터를 결합하여 완전연결계층을 통과한 입력 벡터를 나타낸다. W_H, W_T 와 b_H, b_T 는 각각 가중치 행렬(Weight)과 편향치(Bias)를 의미한다.

3.4 Multi-head 주의 기제 기법

주의 기제란 모델이 주어진 작업을 수행할 때, 관련성 높은 특정 벡터에 주목하게 하여 모델의 성능을 높이는 방법이다. 주의 기제 기법은 기계번역에서 처음 도입되어 텍스트 요약, 캡션 생성 등 다양한 연구에서 적용되어 효과적인 성능을 보이고 있다. 본 연구는 [4]에서 제안한 Transformer 구조의 Multi-head 주의 기제 기법을 적용하여 Scaled dot-product으로 문장 내 단어 간의 유사도를 계산하고, 유사도를 각 단어에 반영해 단어의 레이블을 예측할 때 유사성이 높은 단어에 주목하도록 하였다. 그림 5는 Multi-head 주의 기제의 과정을 도식화한 그림이다.

주의 기제 함수의 수식은 (7)과 같다. 함수의 입력 값인 query(Q)와 Key(K)의 행렬 곱을 계산한 다음, K의

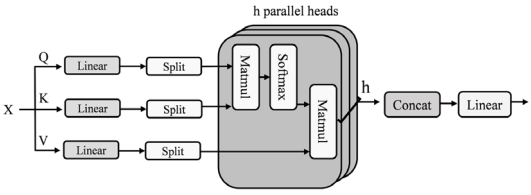


그림 5 Multi-head 주의 기제 구조
Fig. 5 Multi-head Attention Structure

차원인 d_k의 제곱근으로 축소(Scaling)하고, 소프트맥스를 계산하여 모든 단어 간의 유사도를 나타내는 가중치를 구한다. 계산된 가중치를 V와 행렬 곱하여, V에 가중치를 반영한다. 본 모델에서 주의 기제 함수의 입력 값인 Q, K, V는 문장을 구성하는 단어의 은닉 계층으로 모두 동일한 값이다.

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

Multi-head 주의 기제 기법은 그림 4와 같이 Q, K, V에 대해 식 (7)의 주의 기제를 병렬적으로 여러 번 수행하는 기법이다. 이 기법은 여러 부분 공간(Subspace)에 나타내어진 h개의 Q, K, V 선형 값에 주의 기제를 h번 수행함으로써, 각 입력 단어에 대해 더 다양한 주의 기제 정보를 활용할 수 있는 장점을 가지고 있다. 주의 기제는 식 (8)과 같이 Q, K, V에 W^Q, W^K, W^V을 행렬 곱하여 각각 d_q, d_k, d_v 차원으로 h번 선형 투영(Linear projection)

시키고, h개의 QW_i^Q, KW_i^K, VW_i^V에 대해 주의 기제를 수행한다. 식 (9)에서는 h번의 주의 기제 결과 값을 모두 결합(Concatenate)한 다음, 다시 투영시켜 최종 값을 계산한다.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (8)$$

$$MltiHead(Q, K, V) = Concat(head_1, \dots head_h)W^O \quad (9)$$

계산된 Multi-head 주의 기제의 결과 값은 Bi-LSTM을 통과한 은닉 계층과 결합하여 각 단어의 최종 표상으로 사용된다. 이후 과정은 3.1절의 최종 스코어를 구하는 방식과 동일하다. 입력 X에 대한 최종 스코어는 CRF의 전이 행렬과 최종 표상을 완전열결계층에 통과시켜 나온 태깅 스코어를 합산하여 나타낸다.

4. 실험 및 토의

4.1 평가 데이터셋

본 연구에서는 생물학 개체명 인식 모델의 학습 및 성능 평가를 위하여 JNLPBA[24]와 NCBI-Disease[25] 영어 데이터셋을 사용하였다. JNLPBA 데이터셋은 MEDLINE 데이터베이스에서 제공하는 생물학 논문 요약문으로부터 발췌한 22,402개의 문장으로 구성되어 있으며, 이 중 16,690개 문장은 훈련(Train) 데이터, 1,856개 문장은 검증(Validation) 데이터, 3,856개 문장은 테스트(Test) 데이터로 사용하였다. NCBI-Disease 데이터셋은 793개의 질병 관련 논문 요약문으로부터 발췌한 6,982개의 문장으로 구성된 데이터셋으로, 이 중 5,145개 문장은 훈련 데이터, 787개 문장은 검증 데이터, 960개 문장은 테스트 데이터로 사용하였다. JNLPBA 데이터셋은 Protein, DNA, RNA, Cell-line 그리고 Cell-type 총 5개의 생물학 개체명 유형을 가지고 있으며, NCBI-Disease 데이터셋은 Disease 개체명 유형 하나만을 포함하고 있다. 표 1은 각 데이터셋을 요약한 표이다.

4.2 실험 환경 및 모수 설정

본 모델은 python 3.5 기반 텐서플로우(Tensorflow)[30] 프레임워크를 활용하여 구현하였으며, 실험은 리눅스 Centos 운영체제 환경에서 NVIDIA GeForce GTX 1070 GPU로 진행되었다.

단어 임베딩 벡터는 BioASQ에서 제공하고 있는 200차원의 사전 학습된 생물학 단어 벡터를 사용하였으며, 문자 임베딩 벡터는 100차원의 벡터로 무작위로 초기화한 후 학습되도록 설정하였다.

문자 단위의 단어 표상 방법에서 사용되는 CNN에서는 커널 크기를 [3, 5, 7]로, 비선형 함수의 종류는 ReLU를 사용하였으며, LSTM 방법에서는 네트워크의 크기를 100차원으로 설정하였다. Multi-head 주의 기제 기법의 head의 개수는 Transformer 구조에서와 동일하게 8개로 설정하였다.

표 1 데이터셋 설명
Table 1 Dataset details

Dataset	JNLBPA	NCBI-Disease
Target entity	Protein, DNA, RNA, Cell-line, Cell-type	Disease
Type	Sentences	Sentences
Train	16,690	5,145
Development	1,856	787
Test	3,856	960

총 40번의 에폭(Epoch)으로 모델을 학습하였으며, 과적합(Overfitting)을 막기 위해 5번의 에폭 동안 검증 데이터의 성능 개선이 없을 경우 초기 종료(Early stopping) 하도록 하였다. 과적합을 막기 위한 또 다른 방법으로 드랍 아웃(Dropout)과 L2정규화를 적용하였으며, 각각 0.5와 5.0e-4로 모수를 설정하였다. 최적화 알고리즘은 Adam optimizer[26]를 사용하였으며, 학습 속도는 0.001로 초기화하였다.

4.3 평가 방법

모델의 성능을 평가하기 위한 방법으로 대부분의 개체명 인식 모델에서 사용하는 precision(정밀도), recall(재현율), f1-score 평가 지표를 사용하였다. F1-score는 정밀도와 재현율의 조화 평균으로, 데이터를 구성하는 각 유형의 개수가 고르지 않은 불균형 데이터일 경우 자주 사용하는 지표이다. 생물학 개체명 인식에서 사용하는 데이터셋의 경우, 생물학 개체에 해당하지 않는 유형의 비중이 크다. 따라서, 이러한 편중된 유형 분포를 고려하여 본 연구에서는 f1-score을 사용하여 모델간의 성능을 비교 한다. 각 평가 지표의 수식은 아래와 같다.

$$precision = \frac{TP}{TP + FP} \quad (10)$$

$$recall = \frac{TP}{TP + FN} \quad (11)$$

$$f1 - score = \frac{2 * precision * recall}{precision + recall} \quad (12)$$

각 수식에 있는 TP, FP, FN은 다음과 같이 정의된다. TP(True positive)는 실제 정답인데 정답이라고 분류한 경우, FP(False positive)는 오답인데 정답으로 분류한 경우 그리고 FN(False negative) 정답인데 오답으로 분류한 경우이다. 이를 종합적으로 해석하면, 정밀도는 모델이 정답이라고 분류한 것 중 정답일 확률, 재현율은 실제 정답 중에 모델이 정답으로 분류한 확률로 해석 가능하다.

4.4 실험 결과 및 토의

4.4.1 기존 연구와 성능 비교

표 2는 JNLBPA와 NCBI-Disease 데이터셋으로 기존 생물학 개체명 인식 모델과 본 연구에서 제안하는

표 2 기존 연구와 성능 비교 결과
Table 2 Performance comparison of previously studied models

	JNLBPA			NCBI-Disease		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Zhu et al. [17]	69.73	75.64	72.57	86.46	**88.07**	87.26
Wang et al. [27]	**72.72**	77.83	75.19	85.00	87.80	86.37
Habibi et al. [5]	71.35	75.74	73.48	86.11	85.49	85.80
Dang et al. [15]	-	-	-	85.03	83.80	84.41
Proposed model	72.32	**78.71**	**75.38**	**88.58**	86.10	**87.32**

표 3 다양한 단어 단위 임베딩 벡터에 따른 실험 결과
Table 3 Experimental results on various types of word embedding

	JNLBPA			NCBI-Disease		
word embedding	Precision	Recall	F1-score	Precision	Recall	F1-score
Random	65.13	71.19	68.03	74.48	74.28	74.37
GloVe	66.48	72.12	69.18	78.87	77.00	77.92
Pubmed-PMC	**69.38**	**74.83**	**72.00**	**83.11**	**78.04**	**80.49**

모델의 성능을 비교한 것이다. 표 2에 있는 기존 연구 모두 딥러닝 기반의 모델로, CNN 구조[17] 또는 Bi-LSTM-CRF 구조[5,15,27]를 활용한다. 표 2의 모델 간의 성능을 비교한 결과, 본 연구에서 제안하는 모델이 f1-score 기준 가장 높은 성능을 보였다. 이를 통해 Bi-LSTM-CRF 구조에 Highway 네트워크와 Multi-head 주의 기제 기법을 적용한 본 연구의 방법론이 우수한 방법론임을 입증하였다. 또한 JNLPBA 데이터셋의 경우, 기존 연구에 비해 f1-score 뿐만 아니라 재현율의 성능이 향상된 것을 확인할 수 있다. 다중 클래스 분류에 해당하는 데이터셋의 경우, 모델이 전체 정답 개체명 중 실제 정답으로 예측하는 성능이 기존 연구보다 효과적인 것으로 분석된다. 이진 분류에 해당하는 NCBI-Disease 데이터셋의 경우, 기존 연구보다 정밀도의 성능이 향상되었다. 이를 통해 모델이 정답이라고 예측한 것 중 실제 정답의 비율이 기존 연구에 비해 향상된 것을 확인할 수 있다. 이후 4.4.2절과 4.4.3절 실험에서는 단어의 표상에 대한 실험을 진행하고, 4.4.4절과 4.4.5절에서는 모델의 구성요소에 대한 실험을 진행하여 본 연구에서 제안된 각 방법론의 역할과 효용성을 확인한다.

4.4.2 단어 단위 임베딩 실험

본 실험에서는 다양한 단어 임베딩 벡터에 따른 모델의 성능을 비교 분석하였다. 표 3은 Bi-LSTM-CRF 모델에 세 가지의 다른 단어 임베딩 벡터를 사용하여 실험한 결과이다. 두 개의 데이터셋 모두 임베딩 벡터를 랜덤 값으로 초기화한 경우, 가장 낮은 성능을 보였다. 가장 높은 성능을 보인 실험은 BioASQ에서 제공하는 Pubmed-PMC 말뭉치로 사전 학습된 임베딩 벡터를 사용한 것으로, 임베딩 벡터를 랜덤하게 설정한 실험보다 각각의 데이터셋에서 f1-score 기준 3.97%p, 6.12%p 성능 향상을 보였다. 따라서 이후의 모든 실험에서는 가장 좋은 성능을 보인 Pubmed-PMC 단어 임베딩 벡터를 사용하여 진행하였다.

4.4.3 단어와 문자 단위의 혼합 임베딩 실험

본 실험은 문자 단위 임베딩과 단어 단위 임베딩의 조합에 따라, 본 연구에서 제안하는 모델의 성능을 관찰하기 위해 진행하였다. 표 4에 있는 (1)~(5) 실험 모두 제안하는 방법론인 Highway 네트워크와 Multi-head 주의 기제를 추가한 모델을 기준으로 진행되었다. 실험 결과, 문자 단위 임베딩 벡터를 사용한 (2)~(5) 실험 모두 단어 단위 임베딩만 사용한 (1) 실험보다 각 데이터셋에서 f1-score 기준 최대 2.93%p, 3.55%p의 성능 향상이 있었다. 또한, 두 데이터셋에서 모두 CNN과 LSTM 문자 단위 임베딩을 혼합하여 사용한 (4) 실험이 단어 단위 임베딩 하나만 사용한 (2),(3) 실험 보다 성능이 우수했다. 추가적으로 (5) 실험은 혼합된 임베딩 벡터를 200차원의 벡터로 출력하는 완전연결계층에 적용한 모델의 실험 결과로, 두개의 데이터셋에서 각각 f1-score 기준 75.38%, 87.32%으로 가장 좋은 성능을 보였다. 이를 통해, 벡터를 완전연결계층에 적용시킬 때, CNN과 LSTM을 모두 사용한 문자 단위 임베딩 벡터의 성능이 더욱 극대화되는 것을 확인 할 수 있다.

4.4.4 Highway 네트워크 적용에 따른 성능 평가

표 5는 Highway 네트워크와 Multi-head 주의 기제 기법 방법론이 본 연구에서 제안하는 모델에 미치는 영향을 비교 분석하기 위해, 각 구성 요소를 제거하고 진행한 실험 결과이다. 표 5의 결과, Highway 네트워크를 제거했을 때 각각의 데이터셋에서 f1-score 기준 0.58%p, 1.75%p 하락했으며, 이를 통해 Highway 네트워크가 제안하는 모델의 성능 향상에 기여함을 확인하였다. 이는

표 4 단어 단위 임베딩 벡터와 문자 단위 임베딩 벡터의 조합에 따른 실험 결과
Table 4 Experimental results on a combination of character-level embedding and word-level embedding

	JNLBPA			NCBI-Disease		
	Precision	Recall	F1-score	Precision	Recall	F1-score
(1) word	69.36	75.83	72.45	84.36	83.20	83.77
(2) word + character(CNN)	70.97	78.48	74.54	85.92	83.71	84.80
(3) word + character(LSTM)	70.68	78.51	74.39	85.04	85.89	85.46
(4) word + character(CNN, LSTM)	71.12	78.65	74.70	86.17	85.44	85.80
(5) word + character(CNN, LSTM) + fc	**72.32**	**78.71**	**75.38**	**88.58**	**86.10**	**87.32**

표 5 Multi-head 주의 기제와 Highway 네트워크 구성 요소 제거에 따른 실험 결과
Table 5 Ablation experimental results of multi-head attention and highway layer

	JNLBPA			NCBI-Disease		
Ablation	Precision	Recall	F1-score	Precision	Recall	F1-score
Multi-head attention	71.22	78.74	74.79	85.08	84.82	84.95
Highway layer	71.40	78.59	74.80	86.43	84.79	85.57
Multi-head attention, Highway layer	70.86	78.62	74.54	84.09	84.96	84.52
Proposed model	**72.32**	**78.71**	**75.38**	**88.58**	**86.10**	**87.32**

표 6 다양한 종류의 주의 기제 기법에 따른 실험 결과
Table 6 Experimental results on different types of attention mechanism

	JNLPBA			NCBI-Disease		
Attention	Precision	Recall	F1-score	Precision	Recall	F1-score
Bahadanau attention	71.15	78.79	74.78	85.13	84.92	85.03
Luong attention	71.06	78.51	74.60	85.93	85.34	85.63
Multi-head attention	**72.32**	**78.71**	**75.38**	**88.58**	**86.10**	**87.32**

Highway 네트워크를 적용함으로써, 활성화 연산을 수행하는 정보와 수행하지 않는 정보가 학습되어 보다 함축적인 단어 임베딩 벡터가 생성으로 모델의 성능이 향상된 것으로 판단된다.

4.4.5 주의 기제 기법 적용에 따른 성능 평가

마찬가지로 Multi-head 주의 기제 기법을 제거했을 때 각각의 데이터셋에서 f1-score 기준 0.59%p, 2.37%p의 성능저하가 있음을 표 5에서 확인할 수 있다. 이는 Multi-head 주의 기제를 통해 입력 단어들 간의 상관성이 학습되어, 레이블링의 정확도가 향상된 것으로 해석할 수 있다. 본 연구에서는 Multi-head 주의 기제 기법 외에도 두 가지 다른 주의 기제 기법인 Bahadanau 주의 기제 기법[28]과 Luong 주의 기제 기법[29]을 적용하여 비교 실험을 진행하였으며, 그 결과는 표 6과 같다. 세 가지의 주의 기제 기법 중 Multi-head 주의 기제 기법이 각각의 데이터셋에서 가장 효과적인 주의 기제 방법임을 확인하였다.

4.4.6 오류 분석

표 7은 본 연구에서 제안하는 모델의 개체명 인식 정답과 오류 예시이다. 위 두 예시는 본 연구에서 제안하는 모델이 기존 GRAM-CNN 모델[17]보다 정확하게 인식한 예시이다. GRAM-CNN 모델의 경우, 개체명 자체를 인식하지 못하거나 일부만을 인식하였지만, 본 연구에서 제안하는 모델의 경우, 개체명 전체를 정확하게 인식하였다.

아래 두 예시는 본 모델에서 발생한 인식 오류를 나타낸다. 첫 번째와 두 번째 오류 예시 모두 개체명의 일부만 인식하여 오류가 발생하였다. 첫 번째의 경우 개체명의 일부인 'multiple'을 개체명을 수식하는 형용사, 즉 일반 단어로 인식하여 오류가 발생한 것으로 추측된다. 두 번째 예시의 경우, 개체명내의 접속사 사용으로 인하여 모델의 인식 혼동이 발생한 것으로 추측된다. 결과적으로, 두 예시 모두 문장 내에서 생물학 개체명이 존재하는지에 대한 여부보다 개체명이 존재하는 경우, 일반 단어와 분리하여 생물학 개체명만을 정확하게 추출하는 것에 대한 성능 개선이 필요한 것으로 분석된다. 향후에는 이와 같은 오류를 범하지 않기 위해 생물학 개체명의 불규칙한 표기법과 명명 규칙을 고려하여 개체명 전체를 정확하게 추출할 수 있는 방법론을 모색해야 할 것이다.

5. 결론 및 향후 연구

본 연구에서는 Highway 네트워크와 Multi-head 주의 기제 기법을 적용한 생물학 개체명 인식 모델을 제시한다. 또한, CNN과 LSTM으로 생성된 문자 단위 임베딩 벡터를 단어 단위 임베딩 벡터와 결합하여 입력

표 7 JNLPBA 데이터셋에 대한 개체명 인식 결과의 정답과 오류 예시
Table 7 Examples of correct and incorrect test results for JNLPBA dataset

	Correctly predicted	Answer of test set
Our model GRAM-CNN	the authors studied specimens of breast carcinomas from 60 consecutive female ... the authors studied specimens of breast carcinomas from 60 consecutive female ...	breast carcinomas : Protein
Our model GRAM-CNN	Analysis of the region 3' to the CD4+ T-cell gene RPT-1 ... Analysis of the region 3' to the CD4+ T-cell gene RPT-1 ...	CD4+ T-cell gene RPT-1 : Protein
	Incorrectly predicted	Answer of test set
Our model	Induction of early B cell factor (EBF) and multiple B lineage genes ... DNA	multiple B lineage genes : DNA
Our model	DNA binding of the 90- to 100-kDa proteins was not inhibited ... Protein	90- to 100-kDa proteins : Protein

임베딩 벡터로 사용하였다. 제안하는 모델을 검증하기 위해 두 개의 생물학 데이터셋에 대해 다양한 실험을 수행하고, 그 결과를 비교 및 분석하였다. 그 결과, 본 연구의 최종 모델이 제안한 방법론을 적용하지 않은 기본 Bi-LSTM-CRF 모델보다 각각의 데이터셋에서 f1-score 기준 3.38%p, 6.83%p 성능 향상을 확인할 수 있었다. 또한, 각 방법론을 개별적으로 적용할 때보다 모두 포함하여 적용할 때, 성능의 향상 폭이 가장 큰 것을 알 수 있었다. 표 2에 언급된 기존 연구의 모델과 성능 비교에서도 가장 우수한 성능을 보였으며, 이와 같은 결과를 통해 본 연구에서 제안하는 방법론이 생물학 개체명 인식 연구에서 효과적인 방법론임을 입증하였다.

향후에는 더 다양한 생물학 데이터셋에 대해 실험을 수행할 계획이다. 또한, CNN, LSTM으로 생성된 문자 단위 임베딩 벡터와 단어 단위 임베딩 벡터의 결합 방법으로 완전연결계층 외에 각 결합의 중요도를 학습하기 위한 임베딩 레이어에서의 주의 기제 기법 적용에 대해 모색하고 향후 연구를 진행할 것이다.

References

[1] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," *Neural computation*, Vol. 9, No. 8, pp. 1735-1780, Nov. 1997.

[2] J. Lafferty, A. McCallum and FCN. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proc. of the 18th International Conference on Machine Learning (ICML)*, pp. 282-289, 2001.

[3] G. Lample et al., "Neural architectures for named entity recognition," *Proc. of NAACL*, arXiv preprint arXiv:1603.01360, 2016.

[4] R.K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *Deep Learning Workshop, International Conference on Machine Learning*, arXiv preprint arXiv:1505.00387, 2015.

[5] M. Habibi et al., "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, Vol. 33, No. 14, pp. i37-i48, Jul. 2017.

[6] J.P.C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 357-370, 2016.

[7] X. Zhang, J. Zhao and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, pp. 649-657, 2015.

[8] R.K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *Deep Learning Workshop, International Conference on Machine Learning*, arXiv preprint arXiv:1505.00387, 2015.

[9] Y. Tsuruoka and J. Tsujii, "Improving the performance of dictionary-based approaches in protein name recognition," *Journal of biomedical informatics*, Vol. 37, No. 6, pp. 461-470, Dec. 2004.

[10] Tsai, R.T.H et al., "NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition," *BMC bioinformatics*, Vol. 7, No. 5, BioMed Central, Dec. 2006.

[11] N. Ponomareva et al., "Conditional random fields vs. hidden markov models in a biomedical named entity recognition task," *Proc. of Int. Conf. Recent Advances in Natural Language Processing*, p. 483, 2007.

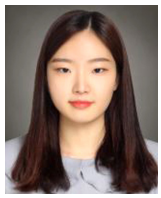
[12] K. Seymore, A. McCallum, and R. Rosenfeld, "Learning hidden Markov model structure for information extraction," *AAAI-99 workshop on machine learning for information extraction*, pp. 37-42, 1999.

[13] O. Bender, F.J. Och, and H. Ney, "Maximum entropy models for named entity recognition," *Proc. of the seventh conference on Natural language learning at HLT-NAACL, Association for Computational Linguistics*, Vol. 4, pp. 148-151, 2003.

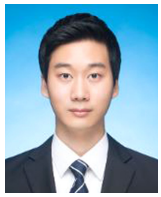
[14] H. Isozaki and H. Kazawa, "Efficient support vector classifiers for named entity recognition," *Proc. of the 19th international conference on Computational linguistics, Association for Computational Linguistics*, Vol. 1, pp. 1-7, 2002.

[15] T.H. Dang et al., "D3NER: Biomedical named entity recognition using CRF-biLSTM improved with fine-

tuned embeddings of various linguistic information," *Bioinformatics*, Vol. 34, No. 20, pp. 3539-3546, Oct. 2018.
[16] H. Yu, Y, Ko, "Expansion of Word Representation for Named Entity Recognition Based on Bidirectional LSTM CRFs," *Journal of KIISE*, Vol. 44, No. 3, pp. 306-313, 2017. (in Korean)
[17] Q. Zhu et al., "GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text," *Bioinformatics*, Vol. 34, No. 9, pp. 1547-1554, 2017.
[18] L. Luo et al., "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition," *Bioinformatics*, Vol. 34, No. 8, pp. 1381-1388, 2017.
[19] M. Cheon et al., "Character-Aware Neural Networks with Multi-Head Attention Mechanism for Multilingual Named Entity Recognition," *Proc. of the 30th Annual Conference on Human and Cognitive Language Technology*, pp. 167-171, 2018.
[20] J.M. Giorgi. and G.D. Bader, "Transfer learning for biomedical named entity recognition with neural networks," *Bioinformatics*, Vol. 34, No. 23, pp. 4087-4094, Dec. 2018.
[21] T. Mikolov et al., "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, pp. 3111-3119, 2013.
[22] Y. Mao, C.H. Wei, and Z. Lu, "NCBI at the 2014 BioASQ Challenge Task: Large-scale Biomedical Semantic Indexing and Question Answering," *CLEF (Working Notes)*, pp. 1319-1327, 2014.
[23] C.M. Park, B.J. Kim, and J.Y. Seo, "Multi-Task based Korean Named Entity Recognition with Highway Bi-LSTM-CRFs," *Proc. of HCI Korea 2018*, pp. 432-435, Jan. 2018. (in Korean)
[24] J.D. Kim et al., "Introduction to the bio-entity recognition task at JNLPBA," *Proc. of the international joint workshop on natural language processing in biomedicine and its applications. Association for Computational Linguistics*, pp. 70-75, 2004.
[25] R.I. Doğan, R Leaman, and Z. Lu, "NCBI disease corpus: a resource for disease name recognition and concept normalization," *Journal of biomedical informatics*, Vol. 47, pp. 1-10, 2014.
[26] D.P.Kingma, and J. Ba, "Adam: A method for stochastic optimization," *ICLR*, arXiv preprint arXiv: 1412.6980, 2015.
[27] X. Wang et al., "Cross-type Biomedical Named Entity Recognition with Deep Multi-Task Learning," *Bioinformatics*, arXiv preprint arXiv:1801.09851, 2018.
[28] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *ICLR*, arXiv preprint arXiv:1409.0473, 2015.
[29] M.T. Luong, H. Pham and C. D. Manning, "Effective approaches to attention-based neural machine translation," *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412-1421, Sep. 2015.
[30] Tensorflow, https://www.tensorflow.org/



조 민 수
2017년 동국대학교 정보통신공학과(학사)
2017년~현재 연세대학교 컴퓨터과학과 석사과정. 관심분야는 빅데이터마이닝 & 기계 학습



박 진 욱
2016년 서울시립대학교 통계학과(학사)
2017년~현재 연세대학교 컴퓨터과학과 석박사통합과정. 관심분야는 빅데이터마이닝 & 기계 학습

박 찬 희
2018년 서울여자대학교 컴퓨터학과(학사)
2018년~현재 연세대학교 컴퓨터과학과 석사과정. 관심분야는 빅데이터마이닝 & 기계 학습

하 지 환
2013년 부산대학교 바이오정보전자과(학사). 2013년~현재 연세대학교 컴퓨터과학과 박사과정. 관심분야는 바이오인포맥틱스, 기계 학습, 데이터마이닝, 데이터베이스

박 상 현
1989년 서울대학교 컴퓨터공학과 졸업(학사). 1991년 서울대학교 대학원 컴퓨터공학과(공학석사). 2001년 UCLA 대학원 컴퓨터과학과(공학박사). 1991년~1996년 대우통신 연구원, 2001년~2002년 IBM T. J. Watson Research Center Post-Doctoral Fellow. 2002년~2003년 포항공과대학교 컴퓨터공학과 조교수. 2003년~2006년 연세대학교 컴퓨터과학과 조교수. 2006년~2011년 연세대학교 컴퓨터과학과 부교수 2011년~현재 연세대학교 컴퓨터과학과 교수. 관심분야는 데이터베이스, 데이터마이닝, 바이오인포매틱스, 빅데이터마이닝 & 기계 학습