

# 생의학 분야 학술 문헌에서의 이벤트 추출을 위한 심층 학습 모델 구조 비교 분석 연구\*

## A Comparative Study on Deep Learning Topology for Event Extraction from Biomedical Literature

김 선 우 (Seon-Wu Kim)\*\*, 유 석 종 (Seok Jong Yu)\*\*\*  
이 민 호 (Min-Ho Lee)\*\*\*\*, 최 성 필 (Sung-Pil Choi)\*\*\*\*\*

### 목 차

- |                                   |               |
|-----------------------------------|---------------|
| 1. 서 론                            | 4. 실험 방법 및 환경 |
| 2. 관련 연구                          | 5. 실험 및 결과 분석 |
| 3. 생의학 이벤트 추출을 위한 컨볼루션<br>네트워크 모델 | 6. 결론 및 향후 연구 |

### 초 록

최근 생의학 분야의 학술 문헌이 기하급수적으로 급증함에 따라 관련 분야 연구자들은 선행 연구 및 연구 동향 파악에 어려움을 겪고 있다. 이에 효율적인 선행 연구 및 연구 동향 파악을 위한 정보 추출 기술이 요구되며, 학술 문헌의 정보 추출을 위한 개체 인식 및 개체 간의 생의학 이벤트 추출 연구가 활발히 진행되고 있다. 본 연구는 이에 심층 학습(Deep Learning)의 기법 중 하나인 컨볼루션 네트워크(Convolutional Neural Networks, CNN) 모델을 기반으로 이벤트 내의 개체 유형 정보의 적용 위치와 함께, 이벤트 식별 및 분류를 고려하여 총 8가지의 모델을 구성하여 실험하였다. 실험 결과, 본 연구에서 제안하는 모델 중 최고 성능을 보인 개체 유형 완전연결 모델이 이벤트 분류 실험에서 F-점수 72.09%의 높은 성능을 보였으나, 이벤트 추출 실험에서는 학습 컬렉션의 불균형 문제 및 이벤트 식별 모델의 성능 저조 등으로 인하여 F-점수 21.81%의 비교적 저조한 성능을 보였다.

### ABSTRACT

A recent sharp increase of the biomedical literature causes researchers to struggle to grasp the current research trends and conduct creative studies based on the previous results. In order to alleviate their difficulties in keeping up with the latest scholarly trends, numerous attempts have been made to develop specialized analytic services that can provide direct, intuitive and formalized scholarly information by using various text mining technologies such as information extraction and event detection. This paper introduces and evaluates total 8 Convolutional Neural Network (CNN) models for extracting biomedical events from academic abstracts by applying various feature utilization approaches. Also, this paper conducts performance comparison evaluation for the proposed models. As a result of the comparison, we confirmed that the Entity-Type-Fully-Connected model, one of the introduced models in the paper, showed the most promising performance (72.09% in F-score) in the event classification task while it achieved a relatively low but comparable result (21.81%) in the entire event extraction process due to the imbalance problem of the training collections and event identify model's low performance.

키워드: 생의학 이벤트, 이벤트 추출, 정보 추출, 자연어 처리, 심층 학습

Biomedical Event, Event Extraction, Information Extraction, Natural Language Processing(NLP), Deep-Learning

\* 본 연구는 한국과학기술정보연구원 주요사업 "초고성능컴퓨팅 기반 건강한 고령사회 대응 빅데이터 기술 개발" 과제의 연구비 지원으로 수행되었음(K-17-L03-C02).

\*\* 경기대학교 일반대학원 문헌정보학과 석사과정(kimsw@kyonggi.ac.kr) (제1저자)

\*\*\* 한국과학기술정보연구원 생명의료융합기술연구실 실장(codegen@kisti.re.kr) (공동저자)

\*\*\*\* 한국과학기술정보연구원 생명의료HPC연구센터 센터장(cokeman@kisti.re.kr) (공동저자)

\*\*\*\*\* 경기대학교 문헌정보학과 조교수(spchoi@kgu.ac.kr) (교신저자)

논문접수일자: 2017년 10월 16일 최초심사일자: 2017년 10월 25일 게재확정일자: 2017년 11월 13일

한국문헌정보학회지, 51(4): 77-97, 2017. [http://dx.doi.org/10.4275/KSLIS.2017.51.4.077]

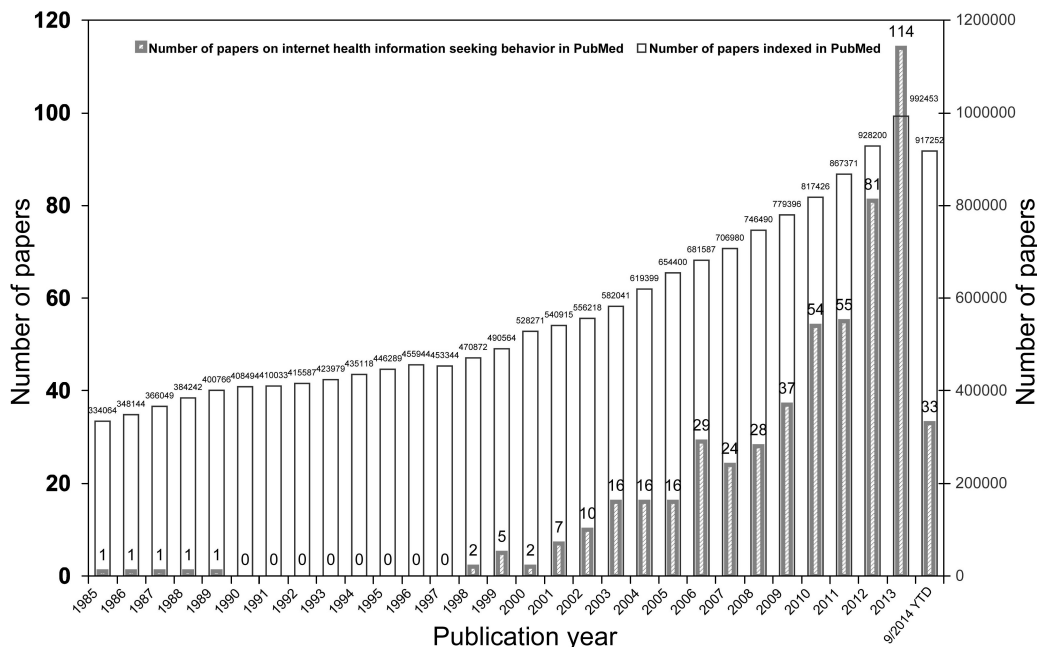
### 1. 서론

최근 생의학 분야 연구의 급속한 발전으로 세계적으로 생산되는 학술 논문의 규모도 매년 빠른 속도로 증가하고 있다. 세계 최대 규모의 생의학 분야 학술 데이터베이스인 PubMed의 경우, 최근 몇 년 사이에 그 규모가 배로 증가하였으며, 현재도 분당 2건 이상씩 증가하고 있다. <그림 1> (Li, F. et al, 2015, fig.1)은 PubMed의 최근 몇 년 간의 학술 논문 규모 및 검색형태에 대한 그래프이다.

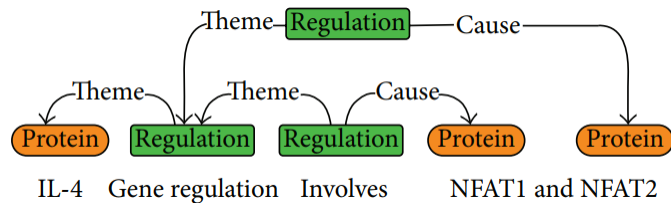
이러한 학술 문헌의 급증으로 인해, 유관 연구자들은 빠른 속도로 발전하고 있는 학술적 성과 및 유관 기술을 이해하고 활용하는 것에 부담을 느끼는 등 어려움을 겪고 있다. 이에 따라, 텍스트 마이닝(Text Mining)이나 학술 지

식 추출(Science Knowledge Extraction) 기술 등을 활용하여 학술 문헌을 정형화하고 이를 효과적으로 서비스하는 학술 정보 서비스의 개발이 중요하게 대두되고 있다(김정균, 조혜진, 이현주 2015).

최근 정보 추출 기술을 통한 이벤트 기반 텍스트 마이닝 기술이 주목을 받고 있다. 이는 학술 문헌 내에서 정보 추출의 결과로 추출되는 개체(Entity)와 개체 간의 관계를 찾아내고, 핵심 이벤트에 대한 정보를 추출해내는 기술이다(김태현 외 2002). 하나의 이벤트는 개체 간의 관계를 표현할 수 있다. 특히 생의학 분야 내에서 발생하는 생의학 이벤트(Biomedical event)는 개체 간에서 발생하는 하나 이상의 동적인 관계를 표현할 수 있어 더욱 큰 의미를 가지게 된다. <그림 2> (Xia, Fang and Zhang



<그림 1> PubMed 데이터베이스 내의 학술 논문 규모 및 검색 형태



〈그림 2〉 생의학 문헌 내 문장에서의 생의학 이벤트의 예시

2014, fig.1)는 생의학 이벤트의 예시이다.

〈그림 2〉와 같이, 하나의 문장에서 3개의 생의학 이벤트를 추출할 수 있으며, 각 이벤트들은 이벤트 트리거(“Gene regulation”, “Involves” 등), 이벤트 유형(“Regulation”), 핵심 참여 객체(Theme), 원인 참여 객체(Cause) 등이 존재한다. “IL-4”와 같이 한 개체가 다수의 개체와 이벤트를 갖거나, “Gene regulation”과 같이 이벤트가 다른 개체와 이벤트를 갖기도 한다.

이렇게 생의학 분야 학술 문헌 내에서 이들 생의학 이벤트를 인식하고 추출하는 기술은 고도화된 학술 지식베이스 구축에 있어서 핵심적인 역할을 수행할 수 있으며 그 활용 범위도 매우 크다. 생의학 이벤트를 추출하기 위해서는 학술 문헌 내에서 개체(단백질, 유전자, 질병명, 약물명, 기타 용어 등)를 자동으로 추출하고 이들 간의 상호 연관 관계를 추출하는 생의학 텍스트 마이닝(Biomedical text mining) 기술이 필요하다(이기현, 허고은, 송민 2015). 본 논문에서는 생의학 이벤트 추출을 위해 심층 학습 기반의 다양한 토폴로지를 구성하고 비교 분석하여, 적절한 토폴로지를 제시하는 것이다. 2장에서는 현재까지 연구된 다양한 생의학 정보 추출 기술을 소개하고, 3장에서는 그 중에서도 생의학 이벤트를 추출하기 위한 심층학습 기반의 모델 구조를 소개한다. 4장에서는 실험 결과

를 비교분석하여 각 토폴로지를 분석한다.

## 2. 관련 연구

Li et al.(2015)는 각 단어의 대표성을 벡터 값으로 표현하는 워드 임베딩(Word Embedding)을 적용하고 각 문장의 자질들을 분석하는 과정에서 개체와 개체의 유형을 주요 정보로 삼았다. 생의학 이벤트 추출의 중요 단서가 되는 문헌 내의 표현인 트리거를 활용한 SVM(Support Vector Machine) 모델을 제시하여 우수한 성능을 보였다. 또한, Liu, Bordes, Grandvalet (2015)는 단순 생의학 이벤트 추출에 있어서 트리거와 인자를 올바르게 추출할 수 있도록, 각 개체에 대한 유형 등에 대한 정보, 단어에서 나타난 기호 및 숫자 정보 등 다양한 자질을 복합적으로 활용하였으며, 단순 이벤트를 다시 인자(Argument)로 받는 복합적인 생의학 이벤트 추출을 위해 트리거와 인자를 짝으로 묶어서 예측을 수행하는 방법을 고안하여 기존 연구 성능의 3.7% 이상의 성능(F-점수) 향상을 보였다.

Huang과 Lu(2016)는 현재까지 2002년부터 2014년까지 진행 되었던 생의학 텍스트 마이닝 분야에 대한 주요 과업을 정리하였다. 그

들이 정리한 다양한 챌린지 중에서도 생의학 이벤트 추출과 관련하여 The BioNLP Shared Task(BioNLP-ST)는 생체 의학 범위 내의 세밀한 정보 추출을 위한 과제로 2009년을 시작으로 2011년, 2013년, 2016년에 생의학 이벤트를 추출하는 경진대회를 진행하기도 하였다. Kim et al.(2011)의 경우, BioNLP-ST 2011에 대한 다양한 기계학습 모델 성능을 비교하고 있다. 최고 성능은 SVM 분류를 사용한 FAUST 모델로, 생의학 이벤트 추출 부분에 대해 57.46%의 F-점수를 보였다. Kim, Wang, Yasunori(2013)의 경우, BioNLP-ST 2013에 대한 다양한 기계학습 모델의 성능을 비교했으며, 생의학 이벤트 추출 부분에 대한 최고 성능은 SVM 분류를 사용한 EVEX 모델로, 전체 분야의 생의학 이벤트 추출에 대해 50.97%의 F-점수를 보였다. BioNLP-ST 2011보다 최고 성능이 낮은 이유를 BioNLP-ST 2013부터는 연구 논문의 초록 외에도 본문까지 담고 있어 데이터 자체의 복잡성이 늘어났고, 컬렉션의 최신성이 높은 논문을 추가하여 최신 용어가 다수 포함되었기 때문으로 분석하였다(Kim, Wang and Yasunori 2013).

최근에는 자연어 처리의 성능 향상에 높은 가능성을 보이는 심층 학습(Deep Learning) 기술이 대두되면서, 이를 생의학 분야의 정보 추출에 적용하는 연구가 활발히 진행되고 있다. 예를 들어, Choi(2016)는 생의학 관계를 추출 과정에 대해 문장 구조 분석에 우수한 성능을 보이고 있는 심층 학습(Deep Learning) 모델인 컨볼루션 네트워크(Convolutional Neural Networks, CNN)를 사용하여 우수한 성능을 보였다.

그러나 생의학 이벤트 추출에 대하여 Li et al.(2015)와 Liu, Bordes, Grandvalet(2015) 같이 다양한 연구 방법론이 연구되었음에도, 이를 심층 학습 기반으로 어떻게 적용할 수 있는지에 대한 연구가 부족한 실정이다. 본 연구는 Kim et al.(2011)과 Kim, Wang, Yasunori(2013)에서 다룬 생의학 이벤트 추출의 주요 과업인 BioNLP-ST를 사용하였다. 그 중에서도 최신 컬렉션인 BioNLP-ST 2016을 사용하여, 심층 학습 기법 중 하나인 컨볼루션 네트워크를 적용한 생의학 이벤트 추출 실험을 수행한다. 특히 Li et al.(2015)와 Liu, Bordes, Grandvalet(2015)에서 공통적으로 사용한 개체 유형을 주요 자질로 보고 이를 어떠한 구조로 사용하는지에 대하여 연구하기 위해, 다양한 토폴로지의 컨볼루션 네트워크 모델을 이벤트 추출에 적용하여 비교 실험하였다.

### 3. 생의학 이벤트 추출을 위한 컨볼루션 네트워크 모델

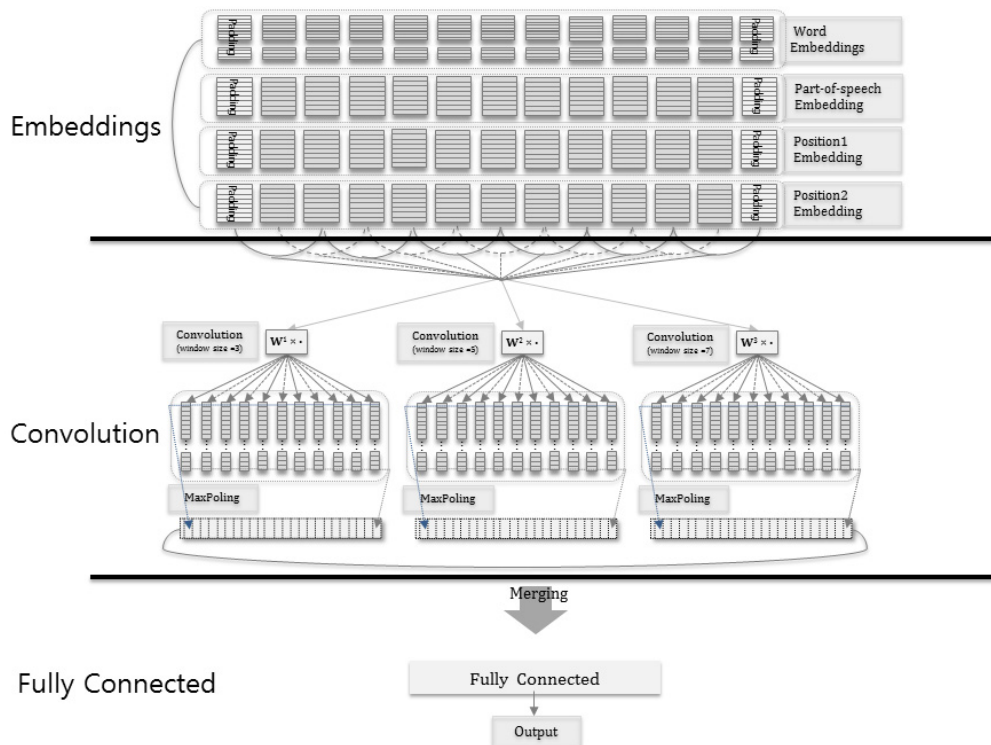
본 연구에서는 심층 학습 기법 중 컨볼루션 네트워크를 기반으로 하여 생의학 이벤트를 추출하였다. 특히, 유관 연구를 참고하여 개체 유형 정보를 핵심적인 자질로 판단하여, 이를 적용하는 위치 및 방법에 따라 모델 구조를 나누어 각 성능을 비교 분석하였다. 먼저 컨볼루션 네트워크 기반의 생의학 이벤트 추출 모델의 기본적인 구조에 대해 설명한 후, 개체 유형 정보의 적용 위치 및 방법에 따른 모델 구조들을 다룬다.

### 3.1 컨볼루션 네트워크 기반 생의학 이벤트 추출 기본 모델

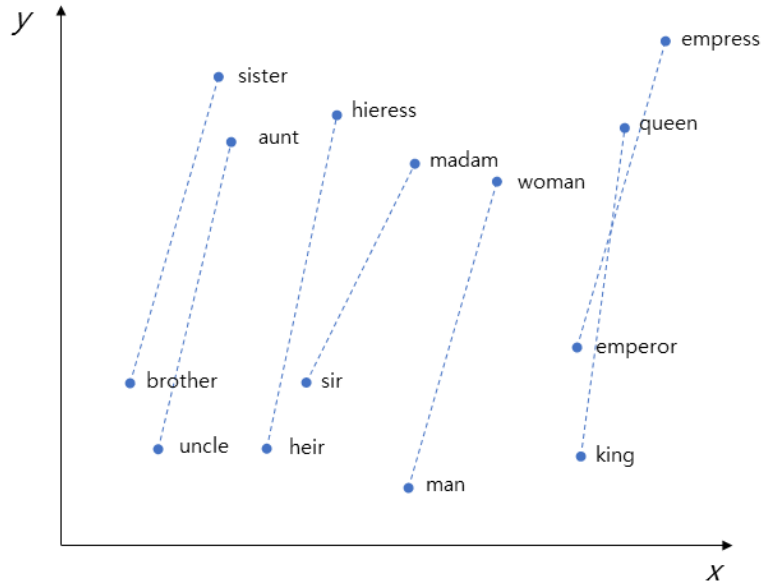
본 연구에서 제안하는 모델은 학술 문헌 내의 개체 쌍을 기반으로 문장 정보를 입력하면 개체 간에 유의미한 이벤트가 있는지 식별하고, 있다면 어떤 유형의 이벤트인지를 분류하는 것을 목적으로 한다. 이에 대한 성능 향상을 위해 본 연구에서는 컨볼루션 네트워크를 기반으로 한 심층 학습 모델을 구성한다. 기본적인 모델의 전반적인 구성은 <그림 3>과 같이 각 정보의 임베딩, 컨볼루션 네트워크, 완전연결 층으로 나누어져 구성된다.

#### 3.1.1 각 자질 정보의 임베딩 층

본 연구에서 구성한 모델은 주요 자질 정보로서 문장에 대한 각 단어 정보, 각 단어에 대한 품사 정보, 첫 번째 개체에 대한 상대적 위치 정보, 두 번째 개체에 대한 상대적 위치 정보를 사용한다. 각 주요 자질 정보는 학습 집합을 기반으로 구성된 사전을 통하여 각 정보를 가리키는 인덱스로 치환되며, 미리 설정된 문장 최대 길이에 맞추어 인덱스 외의 정보는 0으로 채워주는 전처리를 거친다. 각 주요 자질 정보가 토큰 단위의 인덱스 리스트로 입력되면, 각 인덱스 리스트 정보를 기반으로 주요 자질에 대한 벡터를 구성하기 위하여 임베딩(Embedding) 매트릭스에서 해당 인덱스에 해당하는 벡터를 추출



<그림 3> 컨볼루션 네트워크 기반 생의학 이벤트 추출 기본 모델의 구조



〈그림 4〉 학습된 단어 임베딩 벡터 2차원 그래프의 예시

한다. 〈그림 4〉는 학습된 임베딩 벡터의 각 정보를 2차원 그래프로 표현한 예시이다.

〈그림 4〉를 보면, 각 단어에 여성적인 의미가 포함된 단어는 위쪽에, 남성적인 의미가 포함된 단어는 아래쪽에 위치한다. 또한, “sister”, “aunt”, “brother”, “uncle”과 같이 가족적인 의미를 가진 용어는 좌측에 모여서 위치하는 등, 의미적으로 유사한 개념끼리 공간적으로 가까운 거리를 갖는 벡터가 구성된다. 이러한 임베딩 결과는 각 정보에 대한 대표성을 띄며, 이는 해당 정보에 대한 자질 정보로 활용할 수 있다.

단어 정보의 경우, 선학습 임베딩 매트릭스를 정적으로 사용하였고, 그 외의 정보의 경우에는 임의로 구성하여 학습하는 동적 임베딩 매트릭스를 구성하였다. 각 임베딩 매트릭스는  $n \times d$ 에 해당하는 크기를 갖는다.  $n$ 는 선학습 임베딩 매트릭스의 경우, 미리 설정된 사전의 크기를 의미한다. 반면 학습 임베딩 매트릭스

의 경우, 학습 집합 내에서 자체 구성한 각 정보에 대한 사전의 크기를 의미한다.  $d$ 는 각 임베딩 매트릭스를 거쳐 나오는 차원의 수를 의미한다. 선학습 임베딩 매트릭스는 미리 설정된 차원 수를 가지고, 학습 임베딩 매트릭스는 학습 과정에서 설정된 임의의 차원의 수를 가진다. 각 임베딩 매트릭스를 통해, 각 정보 중 입력되는 인덱스 값  $x_i$ 에 대해 임베딩 벡터를 구성한 결과는 수식 (1)과 같다.

$$E = [e_s, e_1, e_2, \dots, e_T, e_e] \quad (1)$$

$e_i$ 는 각 인덱스에 대한 임베딩 매트릭스 내의 벡터를 찾아 추출한 값이며,  $e_s, e_e$ 는 각각 각 인덱스 정보에 대한 문장의 시작과 끝을 나타내는 패딩 값이다. 각 정보의 임베딩 벡터 ( $W, P^{pos}, P^1, P^2$ )는 연결(Concatenate)되어 컨볼루션 네트워크의 입력 값인  $X = [x_s, x_1, x_2, \dots, x_T, x_e]$ 가 구성된

다. 이 중  $x_i$ 는 문장 내의 한 단어에 대한 값을 의미한다. 결과적으로  $x_i$ 는 단어에 대한 임베딩 벡터 값, 단어의 품사에 대한 임베딩 벡터 값, 첫 번째 개체에 대해 해당 단어가 가지는 상대적 위치에 대한 임베딩 벡터 값, 두 번째 개체에 대해 해당 단어가 가지는 상대적 위치에 대한 임베딩 벡터 값이 이어진 형태로 구성된다.

### 3.1.2 컨볼루션 층(Convolution Layer)

컨볼루션 층은 입력된 정보를 일정 부분씩 나누어서 연달아 연산한다. 주로 이미지 정보를 다룰 때 사용하는 방법이지만, 최근 자연어 처리의 방법으로도 사용되기 시작하였다. 문장 단위로 입력된 정보에 대하여, 설정된 크기만큼의 단어들을 대상으로 연산을 수행하여 벡터를 구성하고, 다시 설정된 크기만큼의 다음 단어들을 대상으로 연산을 수행하여 벡터를 구성한다. 이러한 컨볼루션 층은 입력된 문장 정보에 대한 구조적인 연산을 통해, 핵심적인 자질을 추출하는데 효과적이다(Stanford CS231n 2017).

3.1.1을 통하여 구성된 연결된 임베딩 벡터  $X = [x_s, x_1, x_2, \dots, x_T, x_e]$ 를 컨볼루션 층(Convolutional Layers)에 입력한다. 본 모델의 구조는 다중 컨볼루션 필터를 통한 다중 컨볼루션 층을 사용한다. 각 컨볼루션 층은 컨볼루션 연산을 수행하기 위한 3차원의 가중치 필터를 가지는데, 필터의 길이는  $x_i$ 가 가지는 벡터의 길이와 동일하게 구성되며, 너비는 컨볼루션 층마다 지정된 필터 크기  $f_d$  만큼의 값을 가진다. 높이는 컨볼루션 층의 출력 값의 차원 수이며, 변수로서 지정된다. 컨볼루션 층은  $X = [x_s, x_1, x_2, \dots, x_T, x_e]$  내에서 시작 단어인  $x_s$ 로부터  $f_d$ 만큼의 단어 값 벡터를 가중치 필터의 각 값에 곱하고, 그

값을 합산하는 형태로 연산을 수행한다. 연산이 끝나면, 시작 단어의 다음 단어 벡터인  $x_1$ 부터  $f_d$ 만큼의 단어 값 벡터를 통해 같은 연산을 반복하여, 이에 대한 결과 값을 앞서 계산한 값에 이어 붙인다.  $x_e$ 까지의 연산을 마치면, 연결된 연산 결과 값이 가중치 필터의 높이만큼 쌓여 2차원 벡터를 구성한다. 이 2차원 벡터 내에서 효율적인 학습을 위해, 중요 자질 정보를 포함하고 있는 벡터만 사용하기 위하여 최댓값 풀링 층(Max Pooling Layer)의 연산을 적용한다. 최댓값 풀링은 일정 풀링 사이즈로 높이와 너비가 설정된 풀링 필터 내에서 최댓값에 해당하는 값만을 추출하여 계산할 벡터의 크기를 줄이는 방법이다.

다중 컨볼루션 층에서는 컨볼루션 층마다  $f_d$ 의 값을 달리 가진다. 때문에 위 설명한 연산을 각 컨볼루션 층에서 가지는  $f_d$ 에 맞춰 연산을 동시에 수행하며, 이를 통해 각각 다른 값을 얻는다. 각 컨볼루션 층의 결과 값을 연결(concatenate)하여 문장 구조적 자질 벡터  $S = (S_{f_1}, \dots, S_{f_d})$ 를 얻는다.

### 3.1.3 완전연결 층(Fully Connected Layer)

마지막으로, 3.1.2의 컨볼루션 네트워크 과정을 통하여 출력된 벡터  $S = (S_{f_1}, \dots, S_{f_d})$ 를 완전연결 층(Fully Connected Layer)에 입력하여 주요 자질에 따른 예측 가능 유형들에 대한 예측 값들을 제시한다. 미리 지정된 차원  $h$ 만큼의 차원을 가진 가중치(weight) 벡터  $W_{fc}$ 과 바이어스(bias) 벡터  $b_{fc}$ 로 수식 (2)와 같은 연산을 수행한다.

$$fc = g((W^{fc} \cdot x) + b^{fc}) \quad (2)$$

해당 가중치 벡터  $w^{fc}$ 와 바이어스 벡터  $b^{fc}$ 는 랜덤으로 초기화되어 학습 과정에서 최적화 과정을 통해 학습된다. 특히,  $g$ 는 활성화함수로 모델 구성 과정에서의 하이퍼파라미터로 결정되는 비선형 함수이다. 이  $g$ 의 값에 따라 해당 개체 쌍이 유의미한 이벤트를 가지는지를 판별하는 이벤트 식별(Event Identification) 모델과 어떠한 이벤트 유형인지를 분별하는 이벤트 분류(Event Classification) 모델로 나누어 <그림 5>와 같이 구성할 수 있다. 이벤트 식별 모델의 경우에는  $g$ 에 대하여 시그모이드(Sigmoid) 활성화함수(Activation)를 적용하여 0~1 사이의 확률 값을 제시하여 준다. 반면, 이벤트 분류 모델의 경우  $g$ 에 대해, 소프트맥스(Softmax) 활성화함수를 적용하여 전체 이벤트 유형에 대하여 예측 결과 값의 합이 1이도록 값을 출력하여, 각 유형에 따른 예측 가능성을 제시한다.

### 3.2 개체 유형 정보를 활용한 변형된 컨볼루션 네트워크 모델 토폴로지

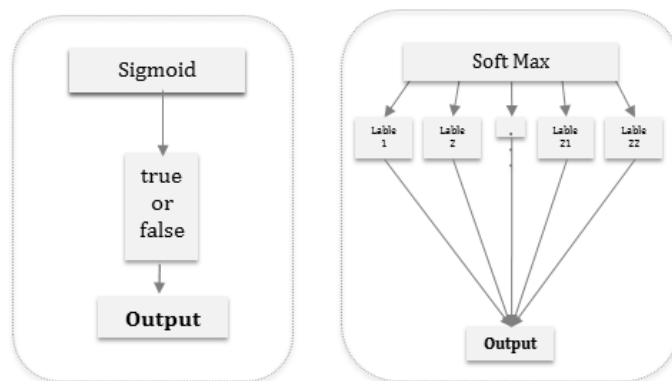
BioNLP-ST 2016 컬렉션에서 개체 유형은

매우 중요한 자질이다. 두 개체가 각각 어떤 유형을 가지고 있는지에 따라 분류 가능한 생의학 이벤트 유형이 한정될 수 있다. 이에 따라, 개체 유형을 어디에서 어떤 방식으로 모델에 적용할지에 대해 개체 유형 완전연결 모델, 개체 유형 임베딩 모델, 개체 유형 선입 모델, 개체 유형 후입 모델로 나누어 구성하였다.

#### 3.2.1 개체 유형 완전연결 모델

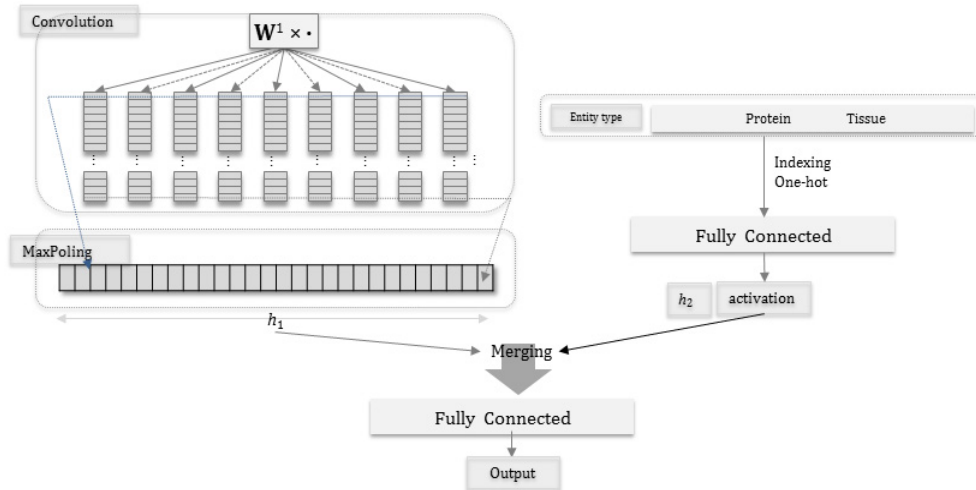
개체 유형 정보를 다른 주요 자질과는 별개로 인덱스에 맞게 One-hot 벡터 형태로 변환하여 완전연결 층으로 입력한다. 여기서, One-hot 벡터는 0과 1로 필요한 정보를 표현하는 벡터로, 필요한 정보를 강조할 때 1과 0으로 그 차이를 극명하게 표현할 수 있어 효과적인 방법이다. 해당 모델은 완전연결 층의 결과 벡터를 다른 자질들의 풀링 층의 결과 벡터와 병합하여 다룬다.

해당 모델의 장점은 개체 유형 정보를 One-hot 벡터로 표현하여, 학습에 있어 개체 유형 정보의 중요성을 높일 수 있다는 점이다. 반면, 완전연결 층을 별개로 추가하여 학습하기 때문에 이에 대한 변수가 필요하다는 단점이 있다. 결과



<그림 5> 시그모이드와 소프트맥스 활성화함수의 비교





〈그림 6〉 개체 유형 완전연결 모델의 벡터 병합

적으로 병합한 값이 최종 완전연결을 거친 후의 활성화함수에 따라 식별 모델과 분류 모델이 나뉜다. 〈그림 6〉은 개체 유형을 One-hot 벡터로 구성하여 완전 연결 형태로 구성하는 부분을 표현한 그림이다.

### 3.2.2 개체 유형 임베딩 모델

개체 유형 정보를 다른 주요 자질과 같이 인덱스 리스트로 구성하여 임베딩 벡터를 적용하는 형태의 모델이다. 임베딩 벡터로 구성된 개체 유형 정보는 다른 주요 자질과 함께 컨볼루션 층의 입력되기 위해 병합되며, 이후에는 기본 모델의 구조를 따른다. 〈그림 7〉은 개체 유형 정보를 다른 자질 정보와 같이 임베딩 벡터로 구성하여 적용하는 부분에 대한 약도이다.

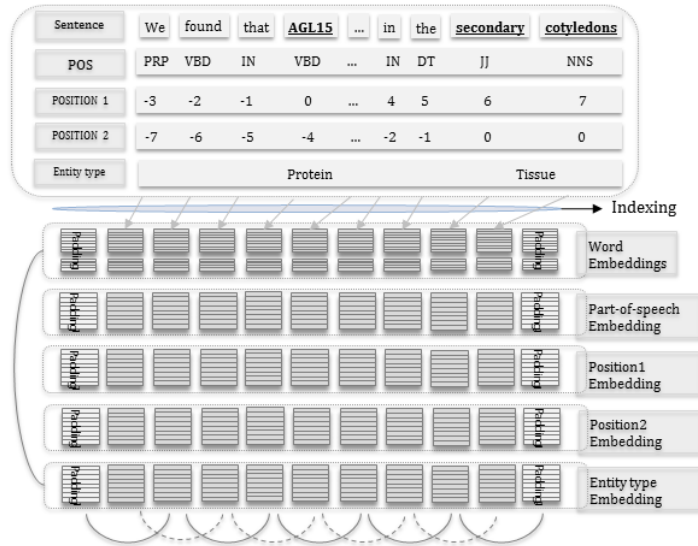
해당 구조는 임베딩을 통하여 개체 유형 정보에 대한 가중치를 학습하고, 다른 모델과 함께 컨볼루션 층에 입력되어 연산되기 때문에, 구조적으로 단순하며 개체 유형 정보에 대한 임베딩 차원 외의 변수 추가가 없다. 하지만, 개체 유형

정보는 2개씩만 입력되기 때문에, 임베딩을 적용할 때 학습의 효율이 비교적 적을 수 있다.

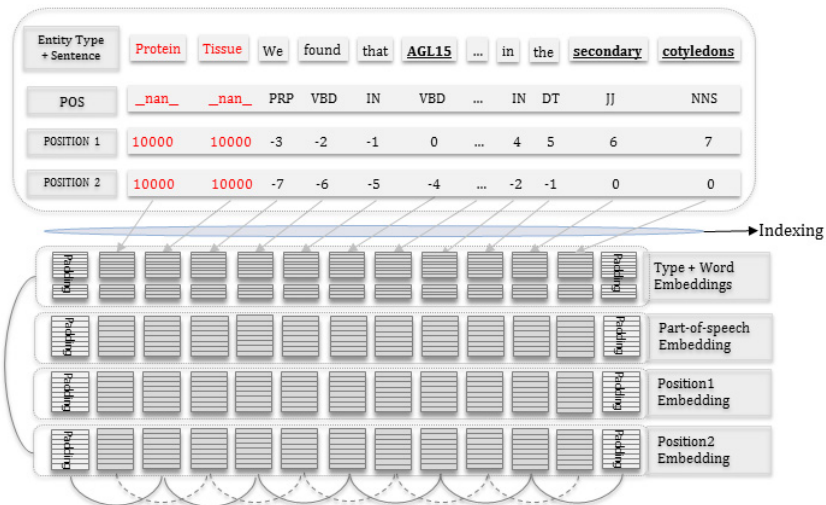
### 3.2.3 개체 유형 선입 모델

개체 유형 정보를 다른 주요 자질의 인덱스 리스트의 가장 앞쪽에 삽입하여 다른 자질들의 정보와 같이 임베딩을 거치도록 하는 형태의 모델이다. 다른 자질들과 같이 임베딩 되어 나온 결과 값이 컨볼루션 층에 입력되며, 이후에는 기본 모델의 구조를 따른다. 〈그림 8〉은 다른 주요 자질의 인덱스 리스트 앞쪽에 삽입하여 임베딩을 같이 진행하도록 구성한 것에 대한 예시이다.

해당 모델은 개체 정보의 유실 없이, 다른 자질 정보 같이 입력된 상태로 임베딩과 컨볼루션의 각 연산을 수행한다. 때문에 모델 구조적인 변화 및 변수 추가가 전혀 없다는 장점이 있다. 반면, 각 자질 정보 앞쪽에 삽입한 개체 유형 정보가 각 임베딩 및 컨볼루션 층의 연산에 대한 성능을 저해할 수 있는 문제점이 있다.



〈그림 7〉 개체 유형 임베딩 모델의 임베딩 적용



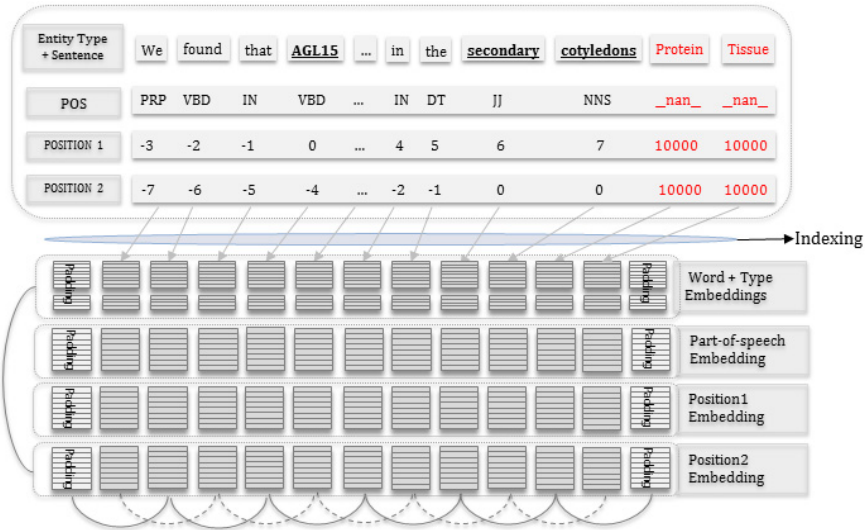
〈그림 8〉 개체 유형 선입 모델의 임베딩 구조

### 3.2.4 개체 유형 후입 모델

개체 유형 정보를 다른 주요 자질의 인덱스 리스트의 가장 뒤쪽에 삽입하여 다른 자질들의 정보와 같이 임베딩을 거치도록 하는 형태의 모델이다. 다른 자질들과 같이 임베딩 되어 나온 결과

값이 컨볼루션 층에 입력되며, 이후에는 기본 모델의 구조를 따른다. 〈그림 9〉는 다른 주요 자질의 인덱스 리스트 뒤쪽에 삽입하여 임베딩을 같이 진행하도록 구성한 것에 대한 예시이다.

해당 모델은 개체 유형 정보를 다른 자질 정



〈그림 9〉 개체 유형 후입 모델의 임베딩 구조

보에 추가하여 학습하는 형태이기 때문에, 모델 구조의 변화 및 변수 추가 등이 없다는 장점이 존재한다. 하지만, 각 자질 정보의 뒤쪽에 추가한 개체 유형 정보가 각 임베딩 및 컨볼루션 층의 연산에 대한 성능을 저해할 수도 있다는 단점이 있다. 또한, 각 자질 정보의 뒤쪽에 추가하기 때문에, 설정된 문장의 최대 길이보다 입력된 문장 길이가 길 경우, 개체 유형 정보를 유실할

수도 있다는 문제점이 존재한다.

본 연구에서 제안하는 네 가지의 변형된 컨볼루션 네트워크 모델 토폴로지의 특징과 장단점을 정리하면 <표 1>과 같다. 이를 기준으로 최종 활성화함수를 시그모이드와 소프트맥스로 나누어 구성하였다. 이에 따라 식별 모델과 분류 모델이 각각 생성되어, 총 8종의 이벤트 추출 모델을 구성하였다.

〈표 1〉 각 변형된 컨볼루션 네트워크 모델 토폴로지의 특징 및 장단점

	완전연결 모델	임베딩 모델	선입 모델	후입 모델
특징	• 개체 정보를 One-hot 벡터로 표현하여 완전 연결하여, 컨볼루션 층의 결과 값에 연결	• 개체 정보를 임베딩하여 다른 자질 정보와 같이 컨볼루션 층에 같이 입력	• 개체 정보를 다른 자질 정보의 앞부분에 입력하여 임베딩을 수행	• 개체 정보를 다른 자질 정보의 뒷부분에 입력하여 임베딩을 수행
장점	• 개체 정보의 중요도를 높임	• 모델 구조 변화 없음	• 개체 정보 유실 없음 • 모델 구조 변화 및 변수 추가 없음	• 모델 구조 변화 및 변수 추가 없음
단점	• 추가 완전연결 층을 위한 별개의 파라미터가 필요	• 관계 마다 2개씩만 입력되는 개체 유형의 임베딩 효과 미비	• 다른 자질 정보에 대한 임베딩 및 컨볼루션 결과를 저해할 수 있음	• 다른 자질 정보에 대한 컨볼루션 결과를 저해할 수 있음 • 문장 길이가 최대 길이 보다 길면 개체 유형이 유실될 수 있음

## 4. 실험 방법 및 환경

본 장에서는 본 연구에서 사용한 컬렉션에 대한 소개와 본 연구에서 제안하는 토폴로지의 성능을 최적화하기 위한 성능 최적화 실험 방법, 실제 토폴로지별 비교 실험 방법, 이벤트 분류 실험 방법을 설명한다.

### 4.1 BioNLP-ST 2016 컬렉션

BioNLP-ST 2016 컬렉션은 기존의 BioNLP-ST의 컬렉션들과 다르게 생의학 이벤트에 대한 트리거가 제거되었다. 또한, 기반 데이터가 생의학 분야의 전문과 초록으로 구성되어 있으며, 학습, 검증, 실험 집합으로 분할이 되어 있다. 학습 집합과 검증 집합의 경우, 해당 레이블이 공개되어 있으며, 실험 집합의 경우 레이블이 공개되어 있지 않아, BioNLP-ST 2016의 공식 홈페이지(<http://2016.bionlp-st.org/>)를 통하여 성능을 확인할 수 있다. 학습 집합의 유의미한 생의학 이벤트가 1,638건으로 BioNLP-ST 2013의 학습 집합 규모(3,152건)에 비하여 그 규모가 축소되었다. 반면, 예측 결과로 나타날 이벤트 유형은 BioNLP-ST 2013이 11종이었던 것에 비하여 13종이 추가되어 24종으로 증가하였다. 이에 따라, 이벤트 분류의 복잡성이 높아졌으며, 이와 함께 규모 축소로 인한 데이터 부족 현상도 나타날 위험성이 있다.

한편, BioNLP-ST 2016은 트리거가 제거되면서, 생의학 이벤트의 요소가 될 수 있는 개체의 유형이 보다 세분화되고 중요성이 높아졌다. 세분화된 개체 유형에 맞게, 생의학 이벤트의 유형도 특정 개체 유형만을 요소로 가지게 되었다.

이는 개체 유형을 주요 자질로서 활용하기 위한 주요 단서로서 활용할 수 있음을 뜻한다. 개체 정보에 대한 중요도가 높아지고 트리거가 제거되면서, 트리거 추출을 중심으로 이루어졌던 기존의 방법론과 다르게 개체 정보를 기반으로 한 생의학 이벤트 추출이 가능하다. 본 연구에서는 개체 쌍 기반의 생의학 이벤트 추출을 위해 문장의 의미 정보, 품사 정보, 첫 개체의 위치 정보, 두 번째 개체의 위치 정보, 두 개체의 유형 정보를 주요 자질로서 추출하였다.

추출하는 과정에서, 개체의 위치 정보가 다중으로 분할되어 있는 경우, 3개 이상의 분할을 가진 데이터와 위치 정보에 대해 오류가 있는 데이터 28건을 제외하였고, 그 결과 학습 집합에서 <표 2>와 같이 학습 집합에서 1,610건이, 검증 집합에서 809건이 추출되어, 총 2,419건이 추출되었다.

### 4.2 성능 최적화 실험

심층 학습 기술의 특징 상, 모델에 적용되는 변수를 변경함에 따라 모델 성능의 변화가 유동적으로 나타난다. 이에 대해서 최적의 성능을 발휘할 수 있는 최적의 변수를 찾는 실험이 필요하며, 이를 성능 최적화 실험이라 한다. 이에 학습과 검증에 대한 실험은 BioNLP-ST 2016 컬렉션에서 자체적으로 분류되어 있는 학습 집합과 검증 집합을 그대로 사용하였지만, BioNLP-ST 2016 컬렉션의 특징 상, 학습 집합과 검증 집합은 유의미한 이벤트에 대한 개체 쌍만이 기재되어 있어, 이에 대하여 임의로 개체 간의 토큰 거리가 10 이하인 모든 개체 쌍을 짝지어 식별 및 분류 모델 실험에 사용하였다. 그 외에 실험에 사용한 변수들은 <표 3>과 같다.

〈표 2〉 BioNLP-ST 2016의 학습 집합 및 검증 집합 통계

이벤트 유형	학습 집합	검증 집합
Is_Linked_To	44	23
Occurs_During	8	10
Transcribes_Or_Translates_To	21	12
Composes_Protein_Complex	16	-
Regulates_Expression	201	111
Has_Sequence_Identical_To	61	20
Is_Functionally_Equivalent_To	58	39
Is_Localized_In	107	47
Regulates_Development_Phase	104	59
Regulates_Molecule_Activity	16	-
Is_Involved_In_Process	23	20
Binds_To	60	24
Is_Protein_Domain_Of	46	28
Regulates_Process	430	177
Regulates_Tissue_Development	9	9
Is_Member_Of_Family	89	52
Composes_Primary_Structure	20	15
Regulates_Accumulation	36	29
Exists_At_Stage	13	8
Exists_In_Genotype	168	79
Interacts_With	62	31
Occurs_In_Genotype	18	16
파일별 총합	1,610	809
전체 파일 총합	2,419	

〈표 3〉 최적화 실험을 위한 변수의 종류와 범위

Parameter	Scope
Window size(Filter size)	3, 5, 7, 9
Drop out	-0.3, -0.5, -0.7
Word vector dictionary	collobert, PubMed (50, 100 dimentions)
Epoch	50, 100, 300, 500
Batch size	32, 128
Optimizer	Rmsprop, Adam, Adadelata, SGD
Learning rate	0.0001, 0.0002, 0.0003, 0.0004, 0.0005, 0.0006, 0.0007, 0.0008, 0.0009, 0.001
Activation	Relu, Sigmoid, Tanh
Filter number	200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800
Hidden dims	100, 300, 500, 700, 900
Early stopping	None, 2, 3, 5, 10

윈도우 사이즈 변수는 컨볼루션에서 적용되는 필터의 크기를 의미한다. 기본적으로 단일 값으로 들어가지만, 그래프 모델을 사용하여 동시에 컨볼루션을 적용할 때는 리스트화 하여 한 번에 사용된다. 실험 진행 과정에서 '9'는 크게 효용성이 없는 것으로 나타나 실험 도중 제외되었다. 드롭아웃 변수는 각 층 사이에서 드롭아웃을 사용할 때의 제외시키는 비율을 의미한다. "Word vector dictionary" 변수는 Collobert et al.(2011)가 구성한 Collobert 사전과 본 연구에서 자체적으로 구성한 PubMed 데이터베이스로 구성된 각 50 차원, 100 차원의 단어 벡터 사전을 사용했다. "Epoch" 변수는 학습 반복의 양을 의미하며, "Batch size" 변수는 데이터를 다루는 배치의 크기를 뜻한다.

"Optimizer" 변수는 1 Epoch마다 모델의 오류율을 줄이는 최적화 방식을 뜻하며, 실험 도중 'Adadelta'와 'SGD'가 성능불량으로 제외되었다. "Learning rate" 변수는 학습 비율을 의미하며, 오류율을 줄이기 위해 어느 정도씩 움직일지를 결정하는 요소이다. "Activation" 변수는 각 층마다의 활성화함수를 의미하며, 실험 도중 'Sigmoid'가 성능이 좋지 않아 제외되었다. "Filter number" 변수는 컨볼루션 모델에서 필터의 개수를 의미한다. "Hidden dims" 변수는 개체유형이 특수 자질로서 사용되는 모델, 즉 개체유형 One-hot 식별 및 분류 모델에서 사용하며, 개체유형 자질을 Dense 층으로 다룰 때 사용한다. 마지막으로, "Early stopping" 변수는 주어진 Epoch 변수 값 내에서 학습 오류율(Loss)이나 검증 오류율(val\_loss)이 더 이상 떨어지지 않을 때 주어진 Epoch를 다 수행하지 않아도 학습을 종료하도록 하여 실험을 좀

더 신속하게 이뤄지도록 하는 변수이다. 각 숫자는 각 오류율이 몇 번까지 떨어지지 않는 지를 결정하는 사항이며, 'None'은 Early stopping을 수행하지 않고 주어진 Epoch를 완수하도록 한다.

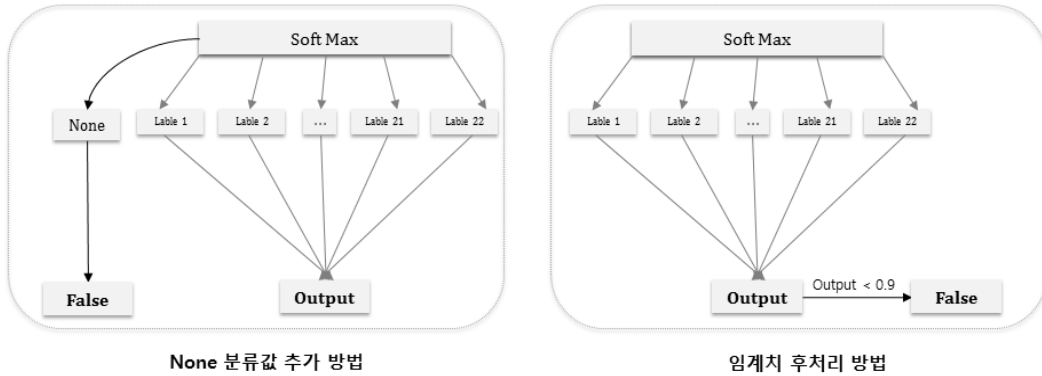
#### 4.3 토폴로지별 이벤트 추출 실험

이벤트 추출에 있어 기본적인 구조는 이벤트 식별을 거친 후, 유의미한 이벤트로 판단되는 개체 쌍에 대해 이벤트 분류를 거쳐 최종적인 이벤트를 추출하는 것이다. 이에 따라, 선정한 식별 모델 2종과 분류 모델 1종을 조합하여 2개의 식별 및 분류 구조를 구성할 수 있다. 또한, "None" 등의 이벤트 유형이 없음을 나타내는 유형을 추가하거나, 각 이벤트 유형의 예측 가능성이 일정한 임계치에 도달하지 못할 경우 이벤트가 없음으로 분류하는 등의 후처리를 수행하면 이벤트 분류 모델을 단독으로 학습하는 형태로도 이벤트 추출을 수행할 수 있다. <그림 10>은 None과 임계치를 활용한 방법을 간략하게 표현한 그림이다.

#### 4.4 이벤트 분류 실험

본 연구에서는 이벤트 분류 모델의 실제적인 이벤트 분류 성능을 파악하기 위해, 유의미한 이벤트가 있는 개체 쌍에 대한 이벤트 분류 실험을 수행한다. 이벤트 분류 성능은 이벤트 추출 성능과 달리 유의미한 이벤트를 추출하는 식별 과정이 제외되고, 유의미한 이벤트가 이벤트 유형 중 어떤 이벤트에 해당하는지를 분류하는 성능이다.

실험을 위하여, 성능 최적화 실험을 통해 가



〈그림 10〉 None 분류값 추가 방법과 임계치 후처리 방법

장 높은 성능을 보인 이벤트 분류 모델을 사용하여 유의미한 이벤트를 가지고 있는 개체 쌍만을 담고 있는 검증 집합을 분류한다.

## 5. 실험 및 결과 분석

본 논문에서 제시한, 먼저 변수 최적화 실험을 통해 각 모델의 기초적인 성능 비교를 수행하였다. 변수 최적화 실험 과정에서 우수한 성능을 보인 식별 모델 2종과 분류 모델 1종을 선택하고, 이에 대한 이벤트 추출 구조를 설정하여 성능 비교 실험을 수행하였다. 추가적으로 이벤트 분류 단독 실험 성능을 파악하기 위한 실험을 별도로 진행하였다. 자세한 실험 내용은 다음과 같다.

### 5.1 성능 최적화 실험 결과 분석

변수마다의 모델로, 총 15,552,000번의 실험이 이뤄질 예정이었으나, 중도에 빠진 변수들을 고려하였을 때, 약 777,600의 모델을 비교하

여 결과를 얻었다.

실험 결과, 학습이 전혀 이루어지지 않거나 저조한 성능을 보인 모델을 제거하고, 비교적 우수한 성능을 보인 모델로 식별 모델 2종(개체 유형 완전 연결, 개체 유형 선입), 분류 모델 1종(개체 유형 완전 연결)을 선정하였다. 〈표 4〉는 선정된 모델과 최적의 성능을 보인 변수의 값이다. 표에서의 성능은 검증 집합에 대한 성능으로, 개체 간의 토큰 거리가 10 이하인 개체 쌍을 임의로 묶어 성능을 평가한 것이다. 때문에 이벤트 분류 모델의 경우, 이벤트 분류만을 단독으로 실험한 것이 아니라 무의미한 개체 쌍에 대해서 성능 평가를 실시한 값이다.

### 5.2 토폴로지별 이벤트 추출 실험 결과 분석

변수 최적화 실험 결과로 선정된 3종의 모델에 대하여, 4.3의 방법론을 기준으로 〈표 5〉와 같이 네 가지 유형으로 구조를 나누어 이벤트 추출 실험을 수행하였으며, 이에 대하여 각각의 성능을 비교하였다.

〈표 4〉 선정된 파라미터 및 검증집합에 대한 성능

모델	식별		분류
	개체 유형 완전연결	개체 유형 선입	개체 유형 완전연결
Number of Filter	600	200	200
hidden dims	900	-	900
optimizer	adam	rmsprop	adam
activation	relu	tanh	relu
Common	Word vector: Collobert 50 Epoch: 300 Early Stopping: 5 (loss) Batch size: 128 Drop out: 0.5 Filter sizes: [3, 5, 7] Learning rate: 0.0001		
재현율	0.6556	0.9438	0.4302
정확도	0.7447	0.9801	0.4008
F-점수	0.6956	<b>0.9619</b>	<b>0.4177</b>

〈표 5〉 실험을 수행한 토폴로지의 명칭과 설명

번호	토폴로지 명칭	설명
1	FCI-FCC (Fully Connected Identification -Fully Connected Classification)	개체 유형 완전연결 이벤트 식별 모델 + 개체 유형 완전연결 이벤트 분류 모델
2	Fisrt in I-FCC (First in Idendification -Fully Connected Classification)	개체 유형 선입 이벤트 식별 모델 + 개체 유형 완전연결 이벤트 분류 모델
3	Only FCC(None)	개체 유형 완전연결 이벤트 분류 모델 단독 None 분류 값을 추가하여 식별 수행
4	Only FCC(Threshold)	개체 유형 완전연결 이벤트 분류 모델 단독 임계치 후처리 기법을 추가하여 식별 수행

이러한 네 가지의 방법에 대한 실험 집합의 성능은 〈표 6〉과 같다.

〈표 6〉과 같이 성능이 가장 좋은 구조는 Only FCC(Threshold) 토폴로지이다. 실험 결과, 임계치 수치에 따라서도 다양한 성능이 나왔는데, 최고 성능은 임계치를 0.9 이상으로 두는 것이었다. 전반적으로 분류 모델을 단독 사용한 Only FCC(None)과 Only FCC(Threshold)의 성능

이 식별 모델과 분류 모델을 적용하여 수행한 모델에 비하여 높은 성능을 낸 것에 비하여 높은 편이다. 이는 식별 모델 자체의 성능이 저조하거나, 두 모델을 함께 사용하면서 각각의 오류율이 곱으로 반영되어 더욱 낮은 성능을 보인 것으로 추론된다. 또한, 전반 구조에서 재현율(Recall)이 정확률(Precision)에 비해 월등히 높은 것을 볼 수 있는데, 이는 그만큼 많은



〈표 6〉 토폴로지별 이벤트 추출 성능

	Predictions	False Negative	False Positive	재현율	정확도	F-점수
FCI-FCC	2180	812	1864	0.2801	0.1450	0.1911
First in I-FCC	4798	591	4261	0.4761	0.1119	0.1812
Only FCC(None)	2908	710	2490	0.3706	0.1437	0.2071
Only FCC(Threshold)	2238	761	1871	0.3254	<b>0.1640</b>	<b>0.2181</b>

개체 쌍을 식별하였다는 것으로 추리되며, 이에 따라 식별 모델 자체의 성능이 저조한 것으로 보인다. 특히, 변수 최적화 실험에서 가장 높은 성능을 보였던 개체 유형 선입 식별 모델의 경우, 오히려 적용 시에 First in I-FCC 토폴로지와 같이 성능이 가장 저조한 것을 볼 수 있다. 이러한 현상은 이벤트 분류 모델에 비해 이벤트 식별 모델이 다루는 학습 집합의 불균형 정도가 심하여 나타난 과적합 현상 때문으로 추론된다. 학습 집합에 대한 과적합 현상은 성능 최적화 실험의 방해요소로서 작용하여, 적절한 학습 변수를 추출하지 못한 것으로 보인다.

결과적으로 변형된 컨볼루션 네트워크 모델 구조 중에서 가장 좋은 성능을 보인 것은 개체 유형 완전연결 모델이다. 이는 개체 유형을 One-hot 벡터로 표현하여 개체 유형 정보의 중요도를 높여 학습하였고, 이러한 방법이 다른 임베딩 및 선입, 후입 방법보다 효과적이기 때문으로 추론된다.

또한, 위 실험 집합에 대한 이벤트 추출 모델의 성능은 대체로 저조한 편이다. BioNLP-ST 2016이 개최 당시 참가했던 팀들의 성능도 대체로 낮은 점수를 기록하였지만, 그 중 최고 성능을 보인 Litway 모델(Li, Rao and Zhang 2016)의 0.432나, 두 번째로 높은 성능을 보인 UniMelb 모델(Panyam et al. 2016)의 0.364의 F-점수 성과와 비교하였을 때에도 현저히

낮은 성능이다. 이는 두 모델의 경우, SVM과 다양한 자질을 기반으로 학습한 모델이지만, 본 연구에서의 모델은 심층 학습 모델로서 방대한 양의 데이터를 필요로 하는 모델이기 때문에 BioNLP-ST 2016의 데이터 규모가 학습 집합으로서 부족하였기 때문으로 보인다. 또한, 기존 연구들에서 적용한 방대한 양의 언어 자질을 추출하여 사용하는 방법을 사용하지 않고, 단순히 컨볼루션 네트워크 모델의 변형만을 이용하여 접근하여 이벤트 추출 과정에 대한 충분한 언어적 자질을 학습하지 못한 결과로 추론된다.

### 5.3 이벤트 분류 성능 실험 결과 분석

위 실험 이후, 이벤트 추출 구조에 있어 이벤트 분류 모델을 단독으로 사용하였을 때 오히려 성능이 좋게 나오는 점과 재현율이 정확률에 비해 성능이 높게 나오는 점 등을 통해, 이벤트 식별 모델의 성능이 저조하다는 것을 추론하였다. 이에 개체 유형 완전연결 분류 모델 자체의 이벤트 분류 성능 평가를 수행하기 위한 실험을 수행하였다. 실험 집합이 공개되어 있지 않아, 유의미한 이벤트를 가진 개체 쌍만을 출력하기 어렵기 때문에, 본 실험은 검증 집합을 통해 이루어졌다. 기존 이벤트 추출 실험은 이벤트 식별 과정을 추가하기 위해, 토큰 거리가 10 이하인 개체 쌍을 임의로 엮어 실험하였으나, 본 실험에

서는 이벤트 분류 과정 자체에 대한 성능 파악을 위해 유의미한 개체 쌍만을 가지고 실험하였다. 실험 결과는 <표 7>과 같다.

실험 결과, <표 7>과 같이 0.7209라는 높은 분류 성능을 보였다. 이는 개체 유형 정보가 이벤트 유형을 분류함에 있어서 중요한 단서가 되기 때문에, 이를 별도로 학습에 사용할 때, 그 의미 정보를 강조할 수 있는 One-hot 벡터로 변환하여 완전연결 층을 통해 학습한 것이 성능의 주요 원인으로 추론된다.

이러한 개체 유형 완전연결 모델의 성능은 식

별 모델과 함께 사용하였을 때보다 단독 사용 성능이 높게 나타난 하나의 이유로 이해할 수 있다. 분류 모델의 성능이 비교적 높기 때문에, 분류 모델을 통해 식별 과정을 함께 수행하는 것이 저조한 성능을 보이는 식별 모델을 통해 식별 과정을 수행하는 것보다 높은 이벤트 추출 성능을 보이는 것으로 추론된다. 또한, 두 가지의 모델을 통해 이벤트 추출을 수행하게 되면, 이벤트 식별 모델의 오류율과 이벤트 분류 모델의 오류율이 곱하여져서 이벤트 추출의 오류율이 나타나기 때문에, 더더욱 저조한 성능이 나

<표 7> 개체 유형 완전연결 모델의 이벤트 분류 성능

이벤트 유형	Predictions	False Negatives	False Positives	재현율	정확도	F-점수
Binds To	18	18	12	0.2500	0.3333	0.2875
Composes Primary Structure	0	15	0	-	-	-
Composes Protein Complex	0	0	0	-	-	-
Exists At Stage	0	8	0	-	-	-
Exists In Genotype	74	7	0	0.9136	1	<b>0.9548</b>
Interacts With	25	20	13	0.3759	0.4800	0.4211
Is Involved In Process	0	20	0	-	-	-
Is Localized In	39	8	0	0.8298	1	0.9070
Is Member Of Family	67	14	26	0.7455	0.6119	0.6721
Is Protein Domain Of	32	3	6	0.8966	0.8125	0.8525
Occurs In Genotype	0	16	0	-	-	-
Occurs During	0	10	0	-	-	-
Regulates Accumulation	2	27	0	0.06897	1	0.1290
Regulates Development Phase	58	9	8	0.8475	0.8621	0.8547
Regulates Expression	83	43	15	0.6126	0.8193	0.7010
Regulates Molecule Activity	0	0	0	-	-	-
Regulates Process	201	2	24	0.9888	0.8806	0.9316
Regulates Tissue Development	0	9	0	-	-	-
Transcribes Or Translates To	2	11	0	0.1538	1	0.2667
Is Linked To	6	19	2	0.1739	0.6667	0.2759
Is Functionally Equivalent To	22	21	3	0.4750	0.8636	0.6129
Has Sequence Identical To	25	12	7	0.6000	0.7200	0.6545
전체	643	292	116	0.6435	0.8196	<b>0.7209</b>

타날 수도 있다. 결과적으로, 성능이 높은 이벤트 분류 모델을 통하여 이벤트 추출을 통합적으로 수행하면, 이벤트 추출의 성능이 향상될 것으로 전망된다.

## 6. 결론 및 향후 연구

본 연구에서는 최근 텍스트 마이닝 및 정보 추출 분야에서 높은 성능을 보이는 심층 학습 기법 중 컨볼루션 네트워크 기반의 모델을 구성하여 BioNLP-ST 2016 컬렉션을 중심으로 생의학 이벤트 추출 모델을 구성하였다. 특히, 기존 연구의 생의학 이벤트 추출 관련 방법론들을 참고하여 개체 유형을 중요 자질로 판단하고, 이를 컨볼루션 네트워크 기반의 심층 학습 모델에 구조적으로 어떻게 적용할 것인지에 대하여 중심으로 연구하였다. 결과적으로, 개체 유형 적용 구조에 따른 모델 4종을 고안하였고, 각 구조의 모델 마다 이벤트 식별 및 분류 모델을 따로 구성하여 총 8종의 이벤트 추출 모델을 구축하여 비교분석하였다.

실험 결과, 개체 유형 완전연결 모델을 단독으로 사용하며 임계치를 0.9 이상으로 둔 구조가 가장 높은 성능을 보였다. 그러나 이벤트 추출 전반에 대한 최고 성능은 F-점수 0.2181로 세계 최고 수준에 비하여서는 다소 저조한 성능이다. 이는 대용량의 데이터를 기반으로 하는 심층 학습 모델인 만큼, 규모가 작은 BioNLP-ST 2016

컬렉션에 대해 다른 자질 및 룰 기반의 기계학습 모델에 비해 저조한 성능을 보이는 것으로 추론된다. 이후, 개체 유형 완전연결 분류 모델의 이벤트 분류 자체에 대한 성능 파악을 위한 실험을 진행한 결과, 분류 자체에 대한 검증 집합 성능은 F-점수 0.7209로, 상당히 준수한 성능을 보였다.

결과적으로 이벤트 분류 성능은 비교적 준수한 성능을 보였지만, 생의학 이벤트 추출 전반에 대해서는 아쉬운 점이 많았다. 그러나 본 연구에서는 이벤트 추출을 위한 다양한 컨볼루션 네트워크 변형 토폴로지를 제시하여 비교 실험을 하였다. 이러한 실험 결과는 향후 해당 연구를 진행할 유관 연구자들에게 가이드 역할을 수행할 수 있다.

추후에는 BioNLP-ST 2016 컬렉션의 복잡성과 데이터 부족 문제를 방지할 수 있는 적합한 규모의 컬렉션을 찾아 생의학 이벤트 추출 실험을 진행할 것이다. 또한, 식별 모델의 성능이 저조한 것을 고려하여, 식별 모델의 성능을 높일 수 있는 방안이나 생의학 이벤트의 식별과 분류를 동시에 진행할 수 있는 새로운 형태의 생의학 이벤트 추출 모델에 대한 구조를 연구할 필요가 있다. 마지막으로, 문장 구조 분석 기반의 컨볼루션 네트워크 모델 외에도 현재 심층 학습 기술에서 연속 정보 및 문맥 정보 파악에 탁월한 성능을 보이고 있는 RNNs 기반의 생의학 이벤트 추출 모델을 연구할 필요가 있다.

## 참 고 문 헌

- [1] 김정균, 조혜진, 이현주. 2015. 생의학 문헌에서 질병 관련 정보를 추출하기 위한 텍스트 마이닝 기법. 『정보과학회지』, 33(4): 13-19.
- [2] 김태현 외. 2002. 정보 추출을 위한 이벤트 문장 추출. 『한국정보과학회 언어공학연구회 학술발표 논문집』, 2002년 10월 11-12일, 청주: 충북대학교: 325-331.
- [3] 이기현, 허고은, 송민. 2015. 생의학 텍스트 마이닝: 새로운 생의학 지식 발견 방법 연구 동향. 『정보과학회지』, 33(4): 30-38.
- [4] Choi, S. 2016. "Extraction of Protein-Protein Interactions(PPIs) from the Literature by Deep Convolutional Neural Networks with Various Feature Embeddings." *Sage Journal*.
- [5] Huang, C. and Lu, Z. 2016. "Community Challenges in Biomedical Text Mining over 10 Years: Success, Failure and the Future." *Briefings in Bioinformatics*, 17(1): 132-144.
- [6] Kim, J. D., Wang, Y. and Yasunori, Y. 2013. "The Genia Event Extraction Shared Task, 2013 Edition-Overview." In *Proceedings of the BioNLP Shared Task 2013 Workshop*, 8-15.
- [7] Kim, J. D. et al. 2011. "Overview of Genia Event Task in BioNLP Shared Task 2011." In *Proceedings of the BioNLP Shared Task 2011 Workshop*, 7-15.
- [8] Li, C., Rao, Z. and Zhang, X. 2016. "LitWay, Discriminative Extraction for Different Bio-Events." In *Proceedings of the 4<sup>th</sup> BioNLP Shared Task 2015 Workshop*, 32-41.
- [9] Li, C. et al. 2015. "Using Word Embedding for Bio-event Extraction." *ACL-IJCNLP*, 2015: 121-126.
- [10] Li, F. et al. 2015. "Mapping Publication Trends and Identifying Hot Spots of Research on Internet Health Information Seeking Behavior: A Quantitative and Co-Word Biclustering Analysis." *J Med Internet Res*, 17(3).
- [11] Liu, X., Bordes, A., and Grandvalet, Y. 2015. "Extracting Biomedical Events from Pairs of Text Entities." *BMC Bioinformatics*, 16(10): 45-49.
- [12] Panyam N. C. et al. 2016. "SeeDev Binary Event Extraction using SVMs and a Rich Feature Set." In *Proceedings of the 4<sup>th</sup> BioNLP Shared Task 2015 Workshop*, 82-87.
- [13] Collobert, R. et al. 2011. "Natural Language Processing (Almost) from Scratch." *Journal of Machine Learning Research*, 12: 2493-2537.
- [14] Xia, J., Fang, A. C. and Zhang, X. 2014. "A Novel Feature Selection Strategy for Enhanced Biomedical Event Extraction Using the Turku System." *BioMed Research International*, 2014.

- [15] Stanford CS231n. 2017. Convolutional Neural Networks for Visual Recognition. *Stanford University CS Class*. [online] [cited 2017. 10. 9.]  
〈<http://cs231n.github.io/convolutional-networks/>〉
- [16] GENIA. 2017. The 4<sup>th</sup> BioNLP Shared Task 2016. *The BioNLP Shared Task*. [online] [cited 2017. 10. 9.] 〈<http://2016.bionlp-st.org/>〉

• 국문 참고자료의 영어 표기

(English translation / romanization of references originally written in Korean)

- [1] Kim, Jeong kyun, Jo, Hye Jin and Lee, Hyeon Ju. 2015. "Text Mining Approaches to Extract Disease Information from Biomedical Articles." *Communications of the Korean Institute of Information Scientists and Engineers*, 33(4): 13-19.
- [2] Kim, Tae-Hyun, et al. 2002. Event Sentence Extraction for Information Extraction. In *Proceedings of the 14th Annual Conference on Human and Cognitive Language Technology*, October 11-12, 2002, Cheongju: Chungbuk National University: 325-331.
- [3] Lee, Keeheon, Heo, Go Eun and Song, Min. 2015. Biomedical Text Mining: A Trend in Biomedical Knowledge Discovery Based on Text Mining. *Communications of the Korean Institute of Information Scientists and Engineers*, 33(4): 30-38.

