

# 코로나 백신 지식그래프 구축: 백신 접종 후 질환 및 증상을 중심으로



삼민이(강민지, 이수민, 김민주)

# 목 차

1. 프로젝트 목표 및 의의
2. 프로젝트 진행 결과
  - 2.1. Data Collection
  - 2.2. Named Entity Recognition
  - 2.3. Relation Extraction
  - 2.4. Knowledge Graph
3. 한계 및 보완점

# 1. 프로젝트 목표 및 의의

## 목표

- 1) ‘코로나 백신 부작용’과 관련된 한국어 데이터셋을 구축한다.
- 2) 개체명 인식(NER), 관계 추출(RE) 기술을 사용하여, 코로나 백신과 백신으로 인해 발생할 수 있는 질환/증상 간의 관계를 지식 그래프로 구축한다.



## 의의

- 1) 한국어 데이터를 사용하여 구축된 ‘첫 코로나 백신 부작용 관련 지식그래프’이다.
- 2) 지금까지 COVID-KG가 임상적/연구적 목적을 위해 개발되어 왔다면, 본 프로젝트의 지식그래프는 전문적인 의학 지식이 없는 정부관계자/일반인에게 유용한 정보를 제공할 수 있다.
- 3) 서로 다른 백신으로 인해 발생하는 부작용의 정보를 체계적/효율적으로 정리, 관리할 수 있다.

## 2. 프로젝트 진행 결과

### 2.1. Data Collection

## 2. 프로젝트 진행 결과: Data Collection



### 코로나19 예방접종 후 이상반응 관련 혈소판감소성 혈전증에 대한 이해

- 혈소판감소성 혈전증이란?**
  - 코로나19 백신 접종 후 발생하는 부작용으로, 혈소판 감소로 인한 출혈과 혈전증이 동시에 생기는 것이 특징입니다.
  - 매우 드물지만 코로나19 백신 접종 후 4일에서 6주 사이에 발생할 수 있습니다.
  - 우리나라에서는 현재까지(9.29일 기준) 3건이 확인되었습니다.
- 코로나19 예방접종 후 혈소판감소성 혈전증 의심증상으로 의사의 진료가 필요한 경우**
  - 코로나19 백신접종 후 4일에서 6주 사이에 아래와 같은 증상이 있는 경우



#### 15일부터 '학교로 찾아가는 백신 접종'... 학교에 구급차 배치

앞서 교육부는 등교 전 코로나19 의심 증상 발생 여부 등을 기록하는 건강상태 자가진단 애플리케이션(앱)을... 백신 부작용에 대한 학부모, 학생의 불안 심리가 반...

| '방역패스' 반발에도 학교 찾아가는 백신접... 한국경제TV 1일 전 네이버뉴스  
청소년 '찾아가는 백신접종' 13일 시작...8만여명... 한겨레 1일 전 네이버뉴스





#### 野 "무책임한 '위드코로나'...청소년 백신, 자율권 존중돼야"

국민의힘이 12일 신종 코로나바이러스 감염증(코로나19) 확진자 급증과 관련해 "정부의 준비없는 무책임한 위드코로나로 우리 국민이 중대한 위험에 빠져들었...

| 원희룡 "청소년-아동 백신 강제 접종 말... 매일신문 PICK 1일 전 네이버뉴스  
국민의힘 "아동-청소년 백신 우려 有..... 이데일리 PICK 1일 전 네이버뉴스





#### "문정부 위드코로나 사과하라"...K방역 실패론 띄우는 국민의힘

국민의힘 대선 중앙선대위는 출범식 다음 날인 지난 7일부터 코로나 백신 부작용 대책, 코로나 중증환자 병상 확보, 코로나 피해 손실보상 등의 공약을 발표하며 ...



Data 1: 질병관리청에서 제공하는 코로나19 예방접종 후 이상반응 관련 정부 공식 자료

Data 2: 올해 6/1~11/30 사이에 발행된 “코로나 백신 부작용” 관련 네이버 뉴스 기사 197건

## 2. 프로젝트 진행 결과: Data Collection

- ❖ 1) 정부 공식 자료와 2) 뉴스 기사 본문을 합쳐 ‘코로나 백신 부작용 데이터셋’을 구축
- ❖ 한국어 문장 분리기(kss)를 사용하여 데이터셋 → 문장 단위로 분리
- ❖ 최종 문장 개수: 4,044개

	press	title	url	content
0	MBN	전두환 백신 부작용 의심'에 이재갑 "가당치도 않은 주장"	http://mbn.mk.co.kr/pages/news/newsView.php?ca...	전두환 전 대통령이 코로나19 화이자 백신을 접종한 뒤에 건강이 급격히 악화됐다는 ...
1	세계일보	'전두환, 코로나 백신 부작용으로 사망' 주장에 정부 "절차 따라 조사할 것"	http://www.segye.com/content/html/2021/11/24/2...	민정기 전 청와대 공보비서관이 지난 23일 사망한 전두환 전 대통령이 화이자사(社)...
2	조선일보	정은경 부스터샷 맞은 날...백신 피해 유족 "내 딸 살려내" 길 막고 항의	https://www.chosun.com/national/national_gener...	코로나 백신 부작용으로 가족이 사망했다고 주장하는 유족들이 19일 정은경 질병관...
3	강원일보	당국 "백신 접종 후 첫 사망신고 고3학생, 접종과 인과성 없다"	http://www.kwnews.co.kr/nview.asp?aid=22111170...	코로나19 백신 접종 후 사망한 고3 학생 사례와 관련해 정부는 백신과의 인과성이 ...
4	세이프타임즈	코로나 백신 부작용 누구 책임 ? ... '부스터샷' 꼭 맞아야 하나	http://www.safetimes.co.kr/news/articleView.ht...	위드 코로나가 시작되면서 위중증 환자가 500명 안팎을 넘나들며 백신 추가접종(부...

scraping한 네이버 뉴스 기사(197건)의 모습

1 전두환 전 대통령이 코로나19 화이자 백신을 접종한 뒤에 건강이 급격히 악화됐다는 전 씨 측 주장이 제기된 가운데 이에 대해 이재갑 한림대 감염내과 교수는 "가당치도 않은 주장"이라고 지적하고 나  
2 이재갑 한림대 감염내과 교수는 25일 CBS 라디오 '김현정의 뉴스쇼'와의 인터뷰에서 전 씨 사망 원인으로 코로나 백신 부작용이 언급된 것에 대해 "사실 가당치도 않은 주장"이라며 "최근에 혈액종양  
3 이 교수는 "전 세계적으로 관련이 증명된 사례는 전혀 없는 상황이고 다른 백신에서도 사례가 전혀 없다"며 "지금까지는 인과관계가 없다"고 강조했습니다.  
4 특히 "전 전 대통령이 걸린 만성골수성백혈병 같은 경우는 원래 꽤 오래 전부터 시작되면서 서서히 시작되는 백혈병 중에 하나이기 때문에 그럴 가능성은 훨씬 더 떨어진다고 볼 수 있다"고 전했습니다.  
5 한편, 전 씨는 지난 23일 오전 8시 40분쯤 연희동 자택에서 쓰러져 숨졌습니다.

문장 단위로 분리된 데이터셋

## 2. 프로젝트 진행 결과

### 2.2. Named Entity Recognition

## 2. 프로젝트 진행 결과: Named Entity Recognition

### 개체명 태그 및 정의

태그	정의/예시
VACC	코로나 백신 (예: 화이자, 모더나, AZ)
DIS	백신 접종 후 발생가능한 질환명 (예: 길랑-바레증후군)
SYMP	백신 접종 후 발생가능한 증상명 (예: 두통, 근육통, 오한)

BIO 형식을 따른 최종 개체명 태그(7개):

O, VACC-B, VACC-I,  
DIS-B, DIS-I, SYMP-B, SYMP-I



### 개체명 태깅

#### 1) 자동 태깅

- ❖ {개체명:태그}(예: '화이자': 'VACC-B')로 이루어진 dictionary 자료형을 사전에 구축하고, 어절(띄어쓰기) 단위로 분리된 문장에 대해 자동 태깅 수행

#### 2) 수동 태깅

- ❖ 자동 태깅 과정에서 태깅이 제대로 이루어지지 않은 어절 token에 대해 수동 태깅 진행



## 2. 프로젝트 진행 결과: Named Entity Recognition

### NER 태깅 예시

영국이 **아스트라제네카(AZ)**의 신종 코로나바이러스 감염증(코로나19) 백신 부작용에 **길랭-바레 증후군**을 추가했다.

○ **VACC-B** ○ ○ ○ ○ ○ ○ ○ **DIS-B DIS-I** ○ ○

**혈전증**이 발생하면 **호흡곤란**과 **흉통**이 나타날 수 있고 심하면 **사망**에 이르게까지 하는 부작용이다.

**DIS-B** ○ **SYMP-B SYMP-B** ○ ○ ○ ○ **DIS-B** ○ ○ ○ ○



KoBERT-NER을 통해 학습, 성능 평가

## 2. 프로젝트 진행 결과: Named Entity Recognition

### Train\_Test\_Split

- ❖ NER 태깅 시 ‘코로나 백신 부작용’과 관련이 없는 문장은 삭제  
(4,044 문장 → 1,758 문장)
- ❖ 〈NE 태그가 O로만 이루어져 있는 문장〉 vs. 〈VACC, DIS, SYMP가 포함되어 있는 문장〉 간의 비율을 고려하기 위해 stratify parameter 사용
- ❖ test\_size = 0.1
- ❖ train: 1,582 test: 176



### 학습 후 성능 평가

	precision	recall	f1-score	support
DIS	0.86	0.78	0.82	41
SYMP	0.75	0.82	0.78	50
VACC	1.00	1.00	1.00	36
micro avg	0.85	0.86	0.85	127
macro avg	0.87	0.87	0.87	127
weighted avg	0.86	0.86	0.86	127

최종 Weighted Average F1-score:

0.86

## 2. 프로젝트 진행 결과

### 2.3. Relation Extraction

## 2. 프로젝트 진행 결과: Relation Extraction

### 관계 타입, 구성 엔티티 타입 및 정의

관계 타입	X	Y	정의/예시
Triggers	VACC	DIS, SYMP	백신에 의해 질환 혹은 증상이 발생하다 (예: 화이자-심낭염)
IsTriggeredBy	DIS, SYMP	VACC	질환 혹은 증상이 백신에 의해 발생하다 (예: 심낭염-화이자)
Accompanies	DIS	SYMP	질환이 특정 증상을 동반하다 (예: 아나필락시스 쇼크-호흡곤란)
IsAccompaniedBy	SYMP	DIS	특정 증상이 질환에 의해 동반되다 (예: 호흡곤란-아나필락시스 쇼크)
Other	VACC	same as X	두 엔티티가 관계를 가지지 않음 (예: 화이자-모더나)

## 2. 프로젝트 진행 결과: Relation Extraction

### RE 자동 태깅

```
re_tag_match = {  
    ('VACC', 'DIS'): 'Triggers',  
    ('VACC', 'SYMP'): 'Triggers',  
    ('SYMP', 'VACC'): 'IsTriggeredBy',  
    ('DIS', 'VACC'): 'IsTriggeredBy',  
    ('DIS', 'SYMP'): 'Accompanies',  
    ('SYMP', 'DIS'): 'IsAccompaniedBy',  
}
```

사전에 ('개체명 태그', '개체명 태그')의 조합으로 만들어질 수 있는 관계 타입을 dictionary 형태로 정의하여, 문장 각각에 대해 RE 자동 태깅

최종 RE 태깅된 문장 개수: 2,392개

### RE 태깅 결과, 예시

Other	1533
Triggers	481
Accompanies	170
IsAccompaniedBy	144
IsTriggeredBy	64
Name: re_tag, dtype: int64	

#### Accompanies (DIS-SYMP)

<e1>아나필락시스 쇼크는 </e1>약제 혹은 꽃가루 등 외부 자극으로 인해 <e2>가려움증, </e2>두드러기, 호흡곤란 등 증상이 나타나는 중증 알레르기 반응이다.

## 2. 프로젝트 진행 결과: Relation Extraction

RE 학습 모델

R-BERT를 사용해 학습, 성능 평가

```
# def load_tokenizer(args):  
#     tokenizer = BertTokenizer.from_pretrained(args.model_name_or_path)  
#     tokenizer.add_special_tokens({"additional_special_tokens": ADDITIONAL_SPECIAL_TOKENS})  
#     return tokenizer  
  
def load_tokenizer(args):  
    tokenizer = KoBertTokenizer.from_pretrained('monologg/kobert')  
    tokenizer.add_special_tokens({"additional_special_tokens": ADDITIONAL_SPECIAL_TOKENS})  
    return tokenizer
```

기존의 R-BERT는 영어 데이터 학습/예측에 사용되었던 모델이었기에,  
한국어 데이터를 학습시키기 위해 tokenizer의 종류를 KoBertTokenizer로 변경 후 학습

## 2. 프로젝트 진행 결과: Relation Extraction

### 학습 후 성능 평가

```
python main.py --do_train --do_eval --train_batch_size 4 --eval_batch_size 4  
--model_name_or_path 'monologg/kobert'
```

	precision	recall	f1-score	support
Others	0.96	0.99	0.98	186
Triggers	1.00	0.95	0.97	20
IsTriggeredBy	1.00	0.50	0.67	2
Accompanies	0.94	0.94	0.94	17
IsAccompaniedBy	0.92	0.73	0.81	15
micro avg	0.96	0.96	0.96	240
macro avg	0.80	0.69	0.73	240
weighted avg	0.96	0.96	0.96	240

최종 Weighted Average F1-score: 0.96

## 2. 프로젝트 진행 결과

### 2.4. Knowledge Graph



## 2. 프로젝트 진행 결과: Knowledge Graph

RE 결과 → KG input 과정에서의 전처리

- ❖ RE 태깅된 문장 중 Other가 아닌 문장만 추출 → 859문장
- ❖ NER, RE 학습시에 '-이/가/을/를' 등 조사가 포함된 '어절' 단위로 학습하였기 때문에, 지식그래프를 구축하기 전, 조사 등의 불용어를 정의하고 엔티티를 '명사형'으로 변환

e1_tag	e2_tag	e1	e2	new_e1	new_e2
VACC	DIS	화이자	골수성 백혈병에	화이자	골수성 백혈병
VACC	SYMP	화이자	숨이	화이자	숨
VACC	DIS	아스트라제네카	급성 골수성 백혈병	아스트라제네카	급성 골수성 백혈병
VACC	DIS	화이자	급성 골수성 백혈병	화이자	급성 골수성 백혈병
VACC	DIS	모더나	골수성 백혈병	모더나	골수성 백혈병

조사 등의 불용어가 제거된 명사형 new\_e1, new\_e2

## 2. 프로젝트 진행 결과: Knowledge Graph

### Neo4j 지식 그래프 구축 결과

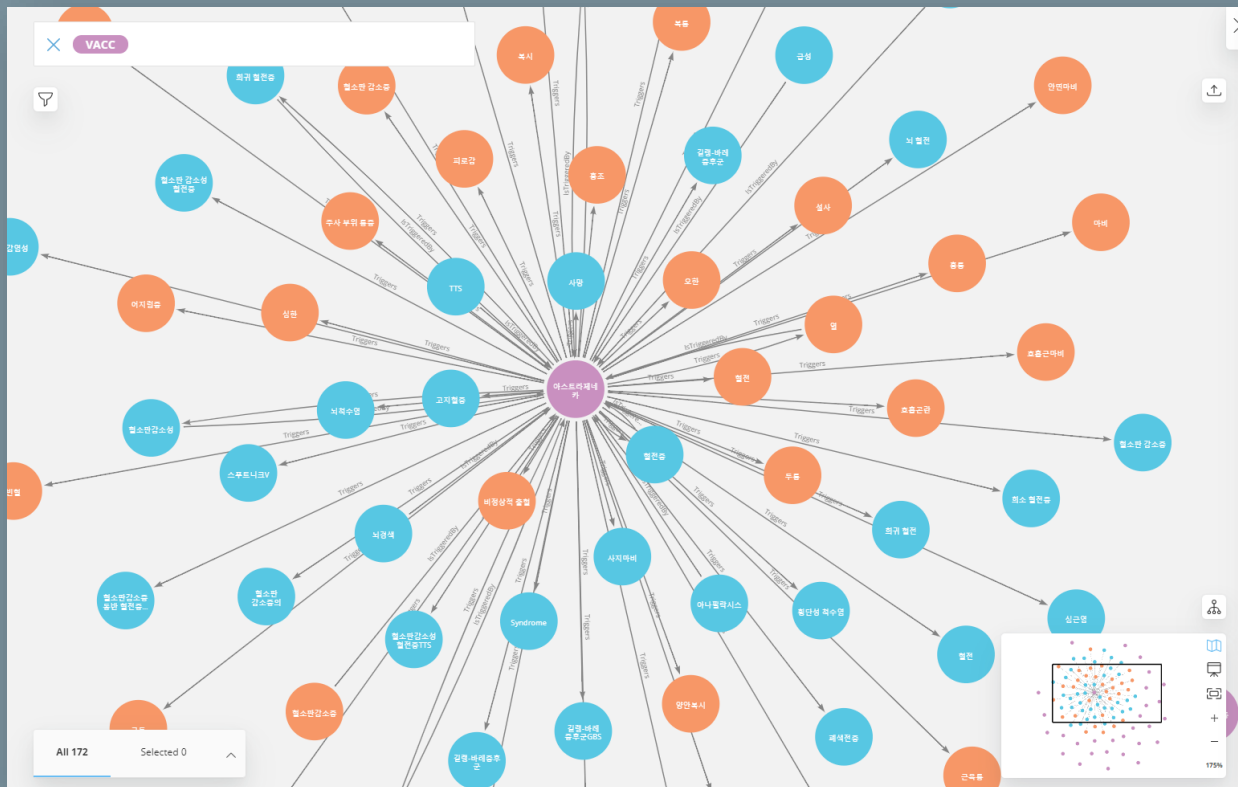
엔티티 태그	개수
VACC	30
DIS	129
SYMP	141

최종적으로 추출된 엔티티 태그 개수

관계 타입	개수
Triggers	291
IsTriggeredBy	53
Accompanies	151
IsAccompaniedBy	124

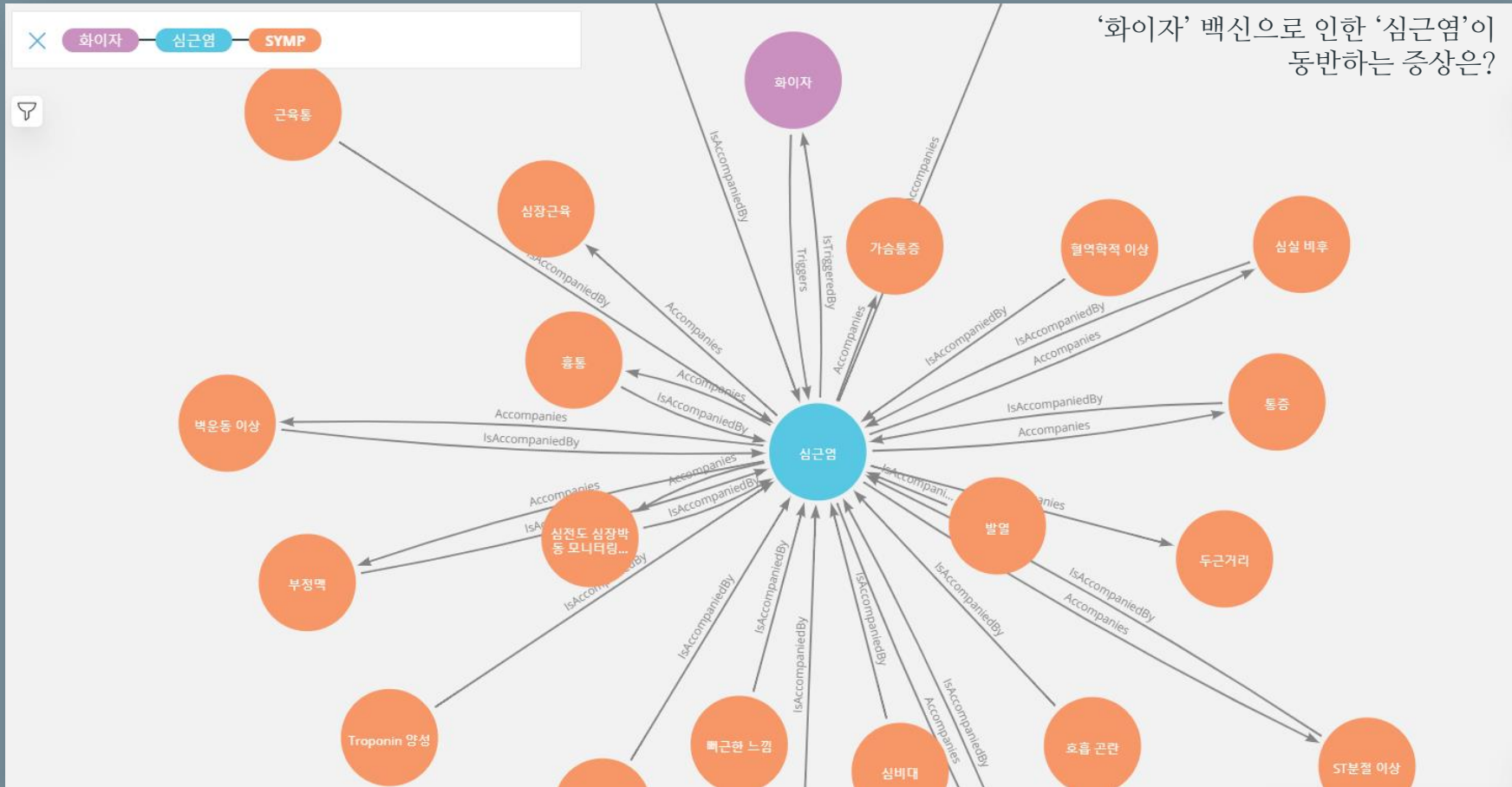
최종적으로 추출된 관계 타입 개수

## 2. 프로젝트 진행 결과: Knowledge Graph

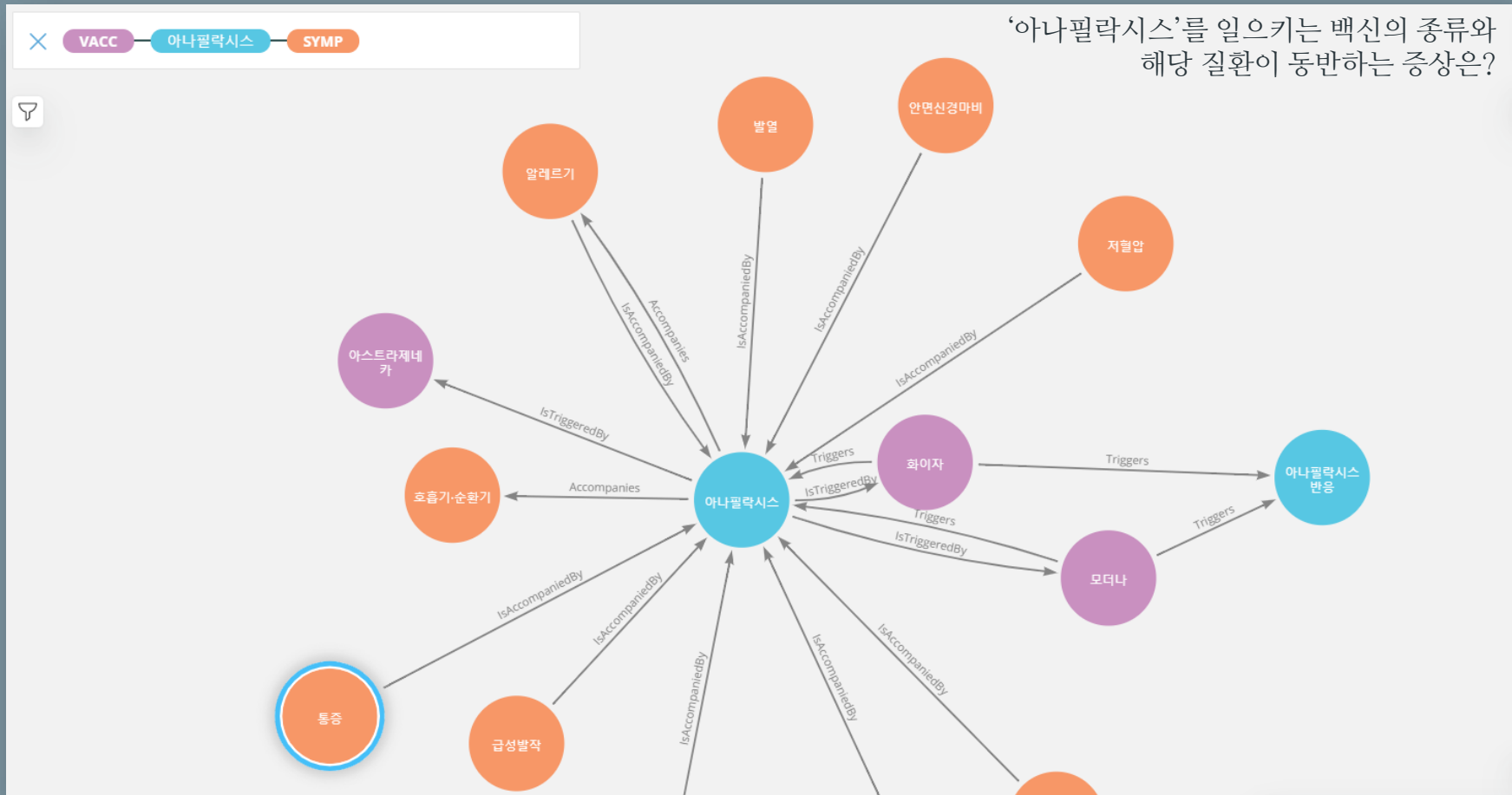


하나의 백신(예: '아스트라제네카')을 중심으로 엔티티들이 관계를 형성하고 있는 모습

## ‘코로나 백신 지식그래프’ 검색 예시 (1)



## ‘코로나 백신 지식그래프’ 검색 예시 (2)



### 3. 한계 및 보완점

코로나 백신 부작용과 관련된 ‘신뢰 가능한’ 데이터의 수가 부족함.

더불어, 각 질환 및 증상에 대한 ‘대처/치료법’이 함께 기술되어 있는 데이터가 절대적으로 부족함.

→ 양질의 데이터 확보를 통해 개체명 태그, 관계 타입의 불균형을 해결해야 함.

정부 자료 뿐만 아니라 뉴스 기사도 함께 데이터셋에 포함시켰기 때문에,

최종적으로 구축된 지식그래프에서의 엔티티 간 관계에 대해 의학적인 검증이 필요함.

Relation Extraction에서 자동 태깅을 사용하였으나, 이는 문장 내에서 <엔티티>와 <엔티티>가 어떤 관계를 이루는지에 대한 서술부(‘~는 ~의 부작용이 아니다’ 등)를 고려하지 않은 방식임.

따라서, 부정확한 RE 관계에 대한 수정/보완이 필요함.

감사합니다