효율적인 자동 주석을 위한 단어 임베딩 인공 신경 정리 증명계 구축

양원석○, 박한철, 박종철
한국과학기술원 전산학부
derrick0511@nlp.kaist.ac.kr, hcpark@nlp.kaist.ac.kr, park@nlp.kaist.ac.kr

Neural Theorem Prover with Word Embedding for Efficient Automatic Annotation

Wonsuk Yang○, Hancheol Park, Jong C. Park
School of Computing, KAIST

요 약

본 연구는 전문기관에서 생산되는 검증된 문서를 웹상의 수많은 검증되지 않은 문서에 자동 주석하여 신뢰도 향상 및 심화 정보를 자동으로 추가하는 시스템을 설계하는 것을 목표로 한다. 이를 위해 활용 가능한 시스템인 인공 신경 정리 증명계(neural theorem prover)가 대규모 말뭉치에 적용되지 않는다는 근본적인 문제를 해결하기 위해 내부 순환 모듈을 단어 임베딩 모듈로 교체하여 재구축 하였다. 학습 시간의 획기적인 감소를 입증하기 위해 국가암정보센터의 암 예방 및 실천에 대한 검증된 문서들에서 추출한 28,844개 명제를 위키피디아 암 관련 문서에서 추출한 7,844개 명제에 주석하는 사례를 통하여 기존의 시스템과 재구축한 시스템을 병렬 비교하였다. 동일한 환경에서 기존 시스템의 학습 시간이 553.8일로 추정된 것에 비해 재구축한 시스템은 93.1분 내로 학습이 완료되었다. 본 연구의 장점은 인공 신경 정리 증명계가 모듈화 가능한 비선형 시스템이기에 다른 선형 논리 및 자연언어 처리 모듈들과 병렬적으로 결합될 수 있음에도 현실 사례에 이를 적용 불가능하게 했던 학습 시간에 대한 문제를 해소했다는 점이다.

주제어: 인공 신경 정리 증명계, 단어 임베딩, 신뢰도 향상 및 심화 정보 추가, 자동 주석 시스템

1. 서론

웹상에 존재하는 수많은 전문 문헌들의 신뢰성과 그 내용의 깊이 부재는 정보 오용의 문제를 야기할 수 있다. 이러한 문제는 웹상의 모든 문헌들이 전문적으로 검증되기 어렵다는 점에서 기인한다. 우리가 흔히 사용하는 영어 위키피디아의 경우 의학 관련 도메인에서 엄밀한 기준에 부합하지 않는 오류가 90%이상 나타난다고 보고되고 있다[1]. 따라서 우리가 흔히 접하는 전문 문헌들에 대한 오류 수정 및 내용 심화를 자동적으로 수행하기 위해 전문적으로 검증된 문서의 지식을 활용하는 전문가 시스템 개발이 요구된다.

이러한 전문가 시스템 구축을 위해서는 검증된 전문 문헌으로부터 많은 양의 신뢰도 높은 지식을 추출할 필요가 있다. 단순히 핵심 어휘를 기준으로 하여 검증된 전문 문서를 범주화할 경우 검증 여부와 관련된 중요 정보가 유실될 수 있다. 때문에 논리 구조를 기준으로 범주화 및 관련 내용의 추적 과정을 퍼지로직 및 자동 정리 증명계(automatic theorem prover)를 이용하여 시도한 연구들이 진행되었다[2, 3]. 이 중 최근 인공 신경 정리 증명계(neural theorem prover)가 소개되었고, 작은 명제 집합에 대해 매우 높은 성능을 보여주었다[4]. 그러나 이 연구의 실험 규모는 약 100개 규모의 공리를 이용하여 4개 명제를 증명하는 것이었고, 기본 구조상 계산 복잡도가 명제 개수의 계승에 비례하며 확률 최대화 계산이 수반되기 때문에, 해당 시스템이 큰 규모의 데이터에 적용될 수 없다는 근본적인 문제가 있다[4].

본 연구에서는 대규모 문장 말뭉치 전체에 대해 증명계가 작동 가능하도록 단어 임베딩 벡터 모델로 기존 인공 신경 정리 증명계의 단어 순환 학습 모델을 대체하여 재구축 하였다.

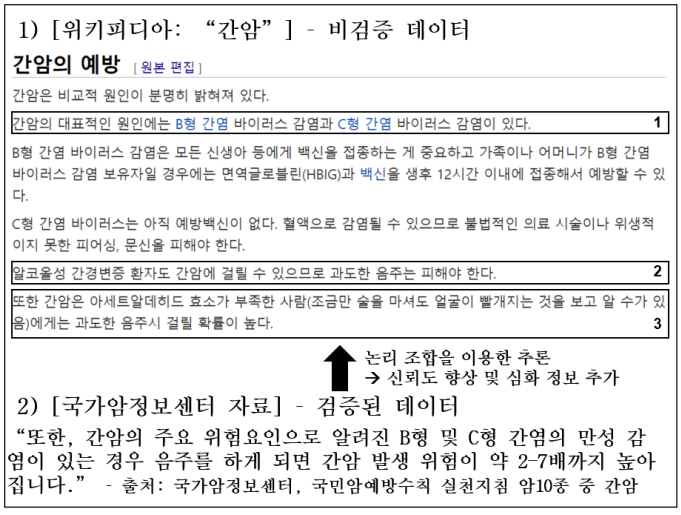

그림 1. 신뢰도 향상 및 심화 정보 추가를 위한 자동 주석 추가의 예

제안하는 방법론을 통해 구축된 전문가 시스템의 활용 예시는 그림 1과 같다. 그림 1의 1)문장은 특정 전문 기관에서 검증을 받지 않은 위키피디아 문서이며[1], 2) 국가암정보센터의 암 예방 및 실천을 위한 책자[2]에서 발췌한 내용은 본 연구에서 제안하는 방법론을 통해 주석될 내용이다. 본 연구에서 재구축된 시스템은, 형태소 분석하여 추출한 위키피디아 데이터의 7,844개 명제를 대상으로 하여 실제 국가암정보센터의 암 예방 및 실천 자료[3]의 28,844개 명제를 전처리 포함하여 약 93분 내에 자동 주석하였다. 또한 전처리 이후 계산 복잡도가 문장 수에 선형 비례한다는 점에 감안하여, 재조립된 인공 신경 정리 증명계가 실제 사례에 적용 가능함을 확인하였다.

2. 관련 연구

건강 관리 시스템 및 암 관련 전문 지식 규격화 시스템에 대한 연구가 국내외에 활발히 진행되고 있다. 이 중 문장 단위로 퍼지 논리를 기반으로 한 검색 기법과 규격 범주화에 대한 연구들이 있었고, 색인어 집합을 기반으로 성공도를 측정하거나, 색인어 간의 비선형 연결도를 단항으로 정의하는 방법을 사용하였다[5, 6].

그 동안 색인어 중심의 규격화가 가지는 자료 손실을 최소화하기 위해, 순환 신경망(recurrent neural network)를 이용한 순서 데이터의 유사도 확인 방법이 활발히 연구되었으며, 문장 일치도 확인에 있어서 매우 높은 성능을 보였다[7, 8].

또한 단어 간의 단순 유사도 측정을 위해 순환 신경망보다 계산 복잡도가 낮으면서도 정보 유실이 최소화 되도록 하는 단어 임베딩 기법이 제안되어 최근 들어 활발히 연구되고 있다[9, 10].

그리고 순서 데이터만을 이용하여 유사도를 확인하는 것을 넘어서, 순환 신경망 혹은 단어 임베딩과 같은 비선형 시스템을 이용하여 논리의 분해, 재조합, 및 전개가 가능하도록 시스템을 재구축하고자 하는 시도도 새로이 진행되고 있다[11, 12].

이런 비선형 시스템 내에 논리 모듈을 장착하는 연구 중 주목할 만한 연구로서, 단어의 벡터화를 통한 인공 신경 증명 추론계가 소개되었으며[4], 논리항을 벡터화하여 다항 논리 구조를 역추론하는 모듈도 개발되었다.

그러나 논리의 분해 및 재조합이 갖는 경우의 수가 내부 항 개수의 계승이기 때문에, 순서 데이터 기반 유사도 측정에 비해 압도적으로 많아 문장 수에 계승으로 비례하는 계산량이 선결되어야 하는 과제로 되었다[4].

본 연구는 기존 인공 신경 정리 추론계가 소개되는 시점에 주어진 명제들 내부 단어 사이의 관계를 활용하여 단어를 벡터화하기 위해 사용되는 순환 모듈 중 순환되는 확률 최대화를 위한 미분 과정에서 가장 많은 시간이 소요된다는 점에 주목하였다.

약 1,000문장 대비 주어지지 않은 약 10만 문장 내의 단어를 단어 임베딩하여 해당되는 1,000문장 내의 단어를 벡터화 하는 것이 학습 시간에 있어서 보다 효과적이라는 점에 착안하여 연구를 진행하였다. 이는 문장 수 증가에 따른 계산 시간 복잡도 증가에 있어서 순환 모듈과 단어 임베딩 모듈이 서로 현저한 차이를 보이기 때문이다. 이에 본 연구에서는 해당 순환 모듈을 외부 대규모 데이터를 활용하는 모듈로 교체하고 다른 인공 신경계 구조는 유지한 상태로 정리 증명계를 작동시켰다.

3. 단어 임베딩을 이용한 인공 신경 정리 증명계

3. 1. 형태소 분석 및 의존 문법 분석

문장 내의 논리를 규격화하여 정리 형태로 변환하기 위해서는 해당 문장을 구 구조 분석하고 의존 문법 등을 통하여 구문 분석하는 과정이 필요하다. 본 연구에서는 논리 구조를 구축하기 위한 기본 단위로 “술어(첫 번째 항, 두 번째 항)” 삼항 구조를 사용하였다. 시스템에 입력되는 모든 문장에 대해서, 세종 한국어 태그를 기준으로 술어가 VP이거나 VNP이며 동시에 두 항이 해당 술어와 직접적으로 연결되어 있을 때 해당 삼항 구조를 추출하였고, 모든 삼항 구조를 각각 하나의 정리로 다루었다.

인공 신경 정리 증명계의 장점인, 풍부한 논리 구조 및 동의어 관계가 비선형적으로 표현되어 있다는 점을 최대한 활용하기 위해 문장 내 최대한 작은 규모의 논리 구조를 하나의 명제로 간주하였다. 또한 한 문장 내에서 여러 삼항 구조가 추출된 경우 각각을 독립된 명제로 간주하였는데, 이 역시 보다 풍부한 논리 전개를 위함이었다.

3. 2. 단어 임베딩

기존의 인공 신경 정리 증명계는 주어진 명제들의 집합에 대해서 순환 모듈을 활용한다. 이는 순환모듈을 사용할 경우 다른 방법론들에 비해서 보다 정교한 시스템 조율이 가능하며, 적은 데이터에 대해서도 효과적이기 때문이다. 다만 순환 모듈의 효과적인 학습을 위해서는 약 50~100번의 순환이 필요하며 정확한 수치는 시스템 조율에 따라서 약간의 차이를 보인다.

이때 주어진 문장에 대해 50~100번의 순환을 할 때의 계산량에 비해 약 50~100배 많은 추가 관련 문장을 단어 임베딩하는 계산량이 현저히 적다. 이는 같은 회수의 유사도 및 순서 비교를 함에도 불구하고 순환 인공 신경망이 단어 임베딩보다 긴 과정의 확률 최대화 계산을 하기 때문이다. 따라서 본 시스템은 약 100배 많은 문장을 추가로 단어 임베딩하는 방법을 통해 순환 모듈을 대체하였다. 이를 통해 같은 횟수의 유사도 및 순서도 비교를 함에도 빠른 계산 시간 내에 인공 신경 정리 증명계를 작동시킬 수 있었다.

---

1) 출처: ko.wikipedia.org/wiki/간암

2) www.cancer.go.kr, 국민암예방수칙 실천지침 암10종 간암

3) 이 자료는 보건복지부 국민건강증진기금의 후원으로 구축된 보건복지부 · 국립암센터 · 국가암정보센터 (www.cancer.go.kr)에서 발췌한 것임. 또한 국가암정보센터로부터 데이터 사용 허가를 받았음을 명시함.

3. 3. 인공 신경 정리 증명계

인공 신경 정리 증명계는 크게 세 가지의 입력을 받는다. 첫 번째는 증명하고자 하는 정리들의 집합이며, 두 번째는 증명을 하는 근거가 되는 공리들의 집합이고, 세 번째는 추론 규칙에 대한 추론 틀이다. 추론 틀은 정리들의 조합을 어떤 범위 내에서 한 정리로 간주하여 논리를 전개할 것인지 결정한다. 추론 틀은 여러 구조로 정의될 수 있는데, 본 연구에서는 기존 인공 신경 정리 증명계에서 차용한 추론 틀을 그대로 사용하였다. 추론 틀은 입력되는 명제 조합과 출력 가능한 명제 사이에 논리 조합을 만들었을 때 연계되는 모순이 없는 경우 강화(reinforce)되는 규칙을 따라 학습된다.

단어 임베딩 모듈을 사전 학습시킨 이후에 각각의 단어에 대응되는 벡터들은 추후 연산 과정에서 고정된 값을 유지한다. 단어 임베딩 이후에 인공 신경 정리 증명계는 고정된 벡터들의 조합을 기준으로 입력되는 명제 조합과 출력 가능한 명제 사이에 논리 조합에 대해 가능한 추론 경로를 탐색한다 (그림 2 참고).

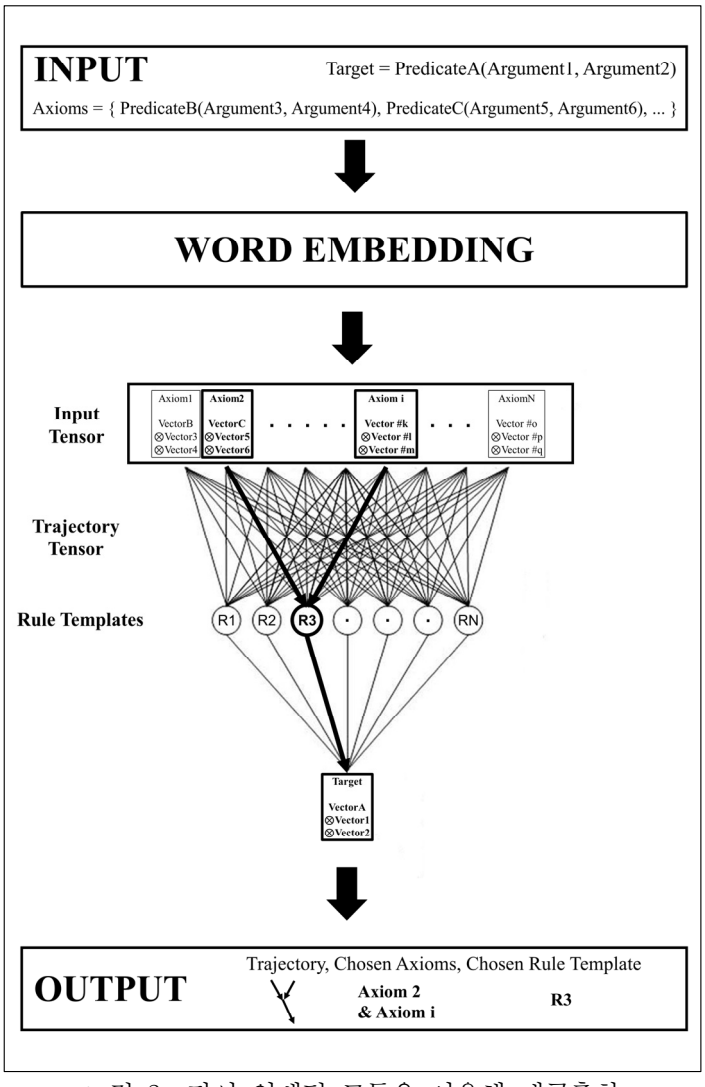

그림 2. 단어 임베딩 모듈을 이용해 재구축한 인공 신경 정리 증명계 모식도

이때 경로 내의 모든 논리 항을 구성하는 벡터들을 차례대로 나열한 희소 텐서에 같은 차원과 랭크의 경로 희소 텐서를 곱한다. 이때 경로 희소 텐서는 학습 전에 무작위로 설정하며 곱한 텐서 내에서 각각의 경로에 해당하는 벡터의 구성성분(component) 중 최솟값이 해당 경로가 선택될 확률이 되도록 정의한다. 이렇게 정의한 경로 희소 텐서를 각각의 경로에 대응하는 매개변수로 설정한다. 이 매개변수의 값에 대해 앞서 서술한 강화 규칙(reinforcement rule)을 이용하여 경로에 해당하는 희소 텐서 값을 정제학습(fine tuning)한다.

4. 실험

4. 1. 실험 세팅

개발한 시스템의 현실 사례 적용 가능성을 검증하기 위하여 사용한 데이터는 다음 표 1과 같다. 이때 위키피디아 암 관련 문서는 종양 키워드에 역링크되는 문서 중 무작위로 50개를 선정하였다.

형태소 분석 및 의존 구분 분석 도구는 NLPhub[13]를 활용하였다. 이때 명제를 추출할 때 의존 문법 분석 과정에서 오류가 발생하는 경우를 제외하고 추출된 모든 명제를 사용하였다. 결과적으로 인공 신경 정리 증명계에 입력 데이터로 활용한 데이터는 위키피디아 암 관련 문서에서 추출한 7,844개 명제와 국가암정보센터 문서로부터 추출한 28,844개 명제였다. 이 명제들에서 술어 및 항에 해당하는 형태소 수는 중복을 제외하여 총 술어 1,476개, 항 4,286개였다.

단어 임베딩을 위해서는 도메인에 맞는 데이터를 활용하기 위해 조선일보, 동아일보, 서울신문, 한겨레, 코리아타임즈의 헬스/의학/건강 섹션의 기사를 활용하였다. 각각의 문장을 형태소 단위로 띄어쓰기 하였으며, 관계언과 의존형태어미에 해당하는 단어는 대응되는 색인으로 문자열을 변경하였다. 전처리 후에 사용된 문장 수는 총 728,835개였다.

단어 임베딩을 위해서 word2vec[9]을 사용하였으며, python에서 인공 신경 정리 증명계가 동작되었다. 이 중 순환 모듈의 네트워크 연산은 (빠른 계산 속도를 위해) 별도로 C++에서 연산되었다. 또한 기존 인공 신경 정리 증명계에서 사용된 것과 같이 ADAM 확률 최적화(stochstic optimization) 알고리즘[14]을 이용하여 순환 오류역전파(backpropagation)를 계산하였다.

표 1. 입력 데이터 통계

	위키피디아 암 관련 문서	국가암정보센터 암 예방 및 실천 자료
#문서	50	33
#문장	1,474	4,927
#단어	25,976	67,888
#명제	7,844	28,844

단어 임베딩을 이용해 재구축한 인공 신경 정리 증명계와 기존의 순환 모듈을 이용한 인공 신경 정리 증명계는 모듈 스위치 온 오프 형식으로 통합된 환경 내에 구축하였다. 초기 설정 매개 변수는 ADAM에서 설정된 값과 같다. 특히 epoch은 50이었는데, 이 값은 유사한 규모의 단어 순서 데이터에 대한 순환 신경망 연구들에서 사용된 값이다 [15, 16]. 100~1000 epoch 이상을 사용하는 경우도 있지만[17], 현재 목적인 시스템의 계산속도 최적화와 부합하지 않는다고 판단하여 50 epoch을 사용하였다.

학습 시간의 비교를 위해 기존의 인공 신경 정리 증명계가 보고될 당시의 소규모 데이터(32개 공리에서부터 4개 명제를 검증함)에 대해서 역시 같은 환경 및 코드에서 구축하여 한 번 작동하는 데에 걸리는 시간을 측정하였다. 또한 기존의 순환 모듈을 탑재한 상태의 인공 신경 정리 증명계가 반복 구문 중 한 번 작동하는 데에 걸리는 시간을 통해 시스템에 의해 정의된 전체 반복 구문을 작동시키는 데에 필요한 시간을 추정하였다.

시스템의 정상 동작을 추가 확인하기 위해 기존의 인공 신경 정리 증명계에서 정의된 일치도 분포를 확인하였다. 일치도 분포가 비정상적으로 집중되어 있지 않은 지를 확인하여 재구축한 시스템이 비정상 동작을 보이지 않았음을 재검증하였다. 이때 결과 일치도 점수가 0이면 False value를 의미하며, 1이면 True value를 의미한다.

4. 2. 결과 및 토의

우선적으로 예측했던 바와 같이 학습 시간에 있어서 순환 모듈을 교체한 것은 매우 커다란 효과를 보였다. 기존에 인공 신경 정리 증명계가 제시될 당시에 소개되었던 32개 공리에서부터 4개 명제를 검증하는 "소규모 데이터 & 순환 모듈"을 사용한 시스템의 경우에는 243초의 학습 시간이 소요되었다 (그림 2 참고). 동일한 코드에 입력을 암 예방 데이터로 변경하였을 경우에는 1개 epoch 내에서 명제 1개에 대한 연산에 122초가 소요되었다. 이때 해당 순환 모듈의 전체 연산은 총 50번 epoch에 대한 반복문 각각에 대해 7,844개의 명제에 대한 연산을 요구한다. 이를 통해 총 47,848,400초(=50*7,844*122초=~553.8일)의 학습 시간을 추정하였다.

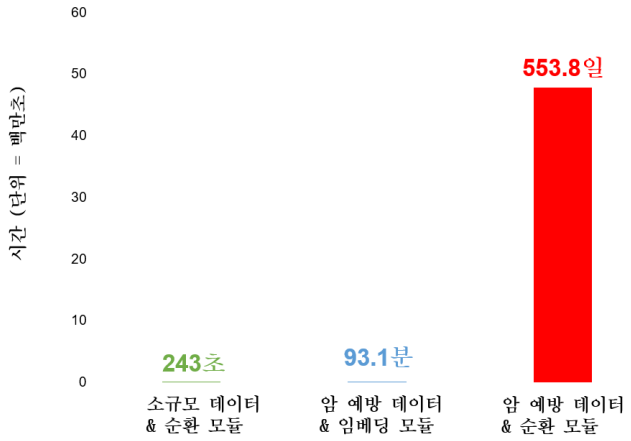

그림 3. 기존 순환 모듈과 현재 임베딩 모듈의 학습 시간 비교

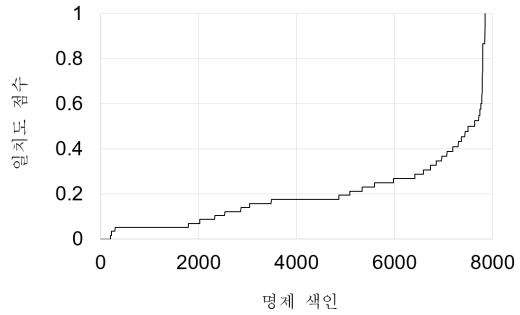

그림 4. 검증 대상 명제들의 신뢰도 분포

이에 비해 학습 시간 향상을 위해 재조립한 "암 예방 데이터 & 임베딩 모듈"을 이용해 시스템을 작동시켰을 경우 전처리과정에 해당하는 728,835개 단어 임베딩에 72.7분, 인공 신경 정리 증명 네트워크 내부에서의 경로 연산에 20.4분으로 총 93.1분이 소요되었다

또한 신뢰도 점수 분포를 확인하였을 때 과도하게 밀집된 구역이 발견되지 않음을 통해 네트워크 내부에서 심한 오버피팅과 같은 비정상적인 작동이 발견되지 않았음을 간접적으로 확인하였다.

1) [위키피디아: 대상 명제] - 비검증 데이터: 기타 아세토제닌은 항-말라리아와 항-종양 성질을 지니고 있고, 실험실 연구에서는 심지어 HIV 복제까지도 억제하는 성질이 있음이 발견되었다.
→ 삼항 구조: 발견(성질, 있)

선택된 추론 틀
: 있(기인,의하) & 발생(의하,발생)
->발견(간암,통하)

2) [국가암정보센터 자료] - 검증된 데이터1: "간염이란 간 조직에 염증이 생기는 질환으로 간염의 중요한 원인 중 하나가 바이러스에 의한 감염입니다." - 출처: 국가암정보센터 공식 사이트 암 예방과 검진 중 예방-감염
→ 삼항 구조: 받(간경변증,하)

- 검증된 데이터2: "반면 선진국에서는 성인이 된 후 이 바이러스에 감염되는 경우가 흔하며 그 결과 증상이 있는 감염성 단핵구증이 비교적 흔하게 발생합니다." - 출처: 국가암정보센터 공식 사이트 암 예방과 검진 중 예방-감염
→ 삼항 구조: 해롭(피우,피우)

그림 5. 출력 결과 예시 1 (논리 조합, 문맥 있음)

1) [위키피디아: 대상 명제] - 비검증 데이터: 바이러스나 기생충과 같은 생물의 세포를 보호하는데, 이러한 기능이 올바르게 작동하기 위해서는 개체 자신의 온전한 세포 또는 조직을 이들로부터 구별해 낼 필요가 있다
→ 삼항 구조: 위하(기능, 올바르)

선택된 추론 틀
: 작업장(시행규칙,취급)&것(대장암,발견)
->위하(예방,시작)

2) [국가암정보센터 자료] - 검증된 데이터1: "산업안전보건법 시행규칙 제 93조에 의하면, 발암성물질을 취급하는 작업장은 작업환경측정을 반드시 해야하며, 작업환경측정 횟수는 작업공정이 신규 또는 변경된 경우, 그날부터 30일내에 하고, 그 후 6개월에 1회 이상 정기적으로 실시해야하며, 발암성 물질의 측정치가 노출기준 초과시 3개월에 1회이상 실시해야한다고 규정하고 있습니다." - 출처: 국가암정보센터 공식 사이트 암 예방과 검진 중 예방-직업성 암
→ 삼항 구조: 작업장(시행규칙,취급)

- 검증된 데이터2: "간흡충은 대변검사에서 기생충란을 발견하여 진단할 수 있고 혈액검사로 진단의 도움을 받습니다." - 출처: 국가암정보센터 공식 사이트 암 예방과 검진 중 예방-감염
→ 삼항 구조: 받(혈액검사,발견)

그림 6. 출력 결과 예시 2 (논리 조합, 문맥 없음)

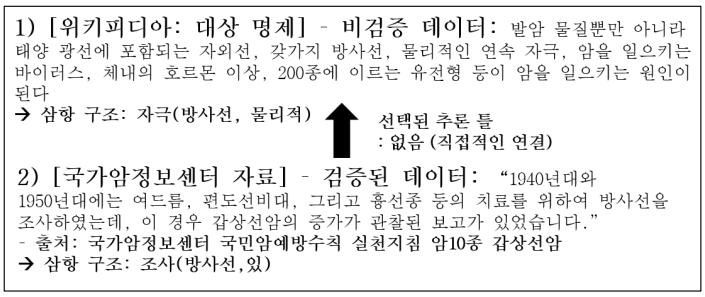

그림 7. 출력 결과 예시 3 (단항 연결, 문맥 있음)

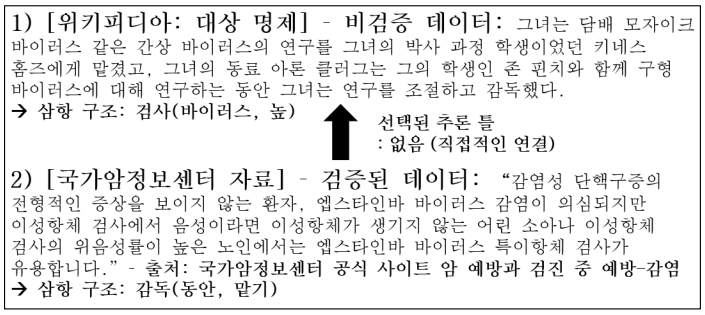

그림 8. 출력 결과 예시 4 (단항 연결, 문맥 없음)

앞서 서론에서 언급한 것과 같이 목표하는 시스템에 본 연구가 어떤 역할을 하는지를 묘사하는 목적으로 결과 예시들을 4개 나열하였다 (그림 5~8). 먼저 본 연구가 풍부한 논리 조합을 통해 최대한 넓은 범위의 연결성을 보장하고자 했음을 밝히고자 한다. 문장의 매우 작은 범위 내의 구성 요소를 기준으로 한 문장 전체를 주석한 경우가 많기 때문에, 문장 전체로 볼 경우 문맥을 전혀 고려하지 않은 결과들을 모두 포함하여 출력하였다.

본 연구가 가지는 가장 큰 강점은 기존의 인공 신경망 내부를 해석하기 힘들었던 것에 비해 내부 전개를 명확히 관찰할 수 있다는 점이다. 인공 신경 정리 증명계는 비선형적인 확률 최대화 계산과 벡터화 모듈을 이용하여 연산을 하지만 내부의 논리 네트워크는 모듈화 가능한 형태로 구성되어 있다. 따라서 시스템 내부를 포함하여 추가 모듈을 장착하는 것이 용이하며 현실 사례의 활용을 위한 경우에 인공 신경 정리 증명계 자체가 다른 선형 논리 모듈들과 병렬적으로 연결되어 하나의 모듈로서 기능할 수 있다.

문장의 작은 구성 요소를 기준으로 결과를 확인 할 때 인공 신경 정리 증명계가 성공적으로 작동했음을 알 수 있다. 다만 예시 2에서 분명하게 알 수 있듯이 문맥 정보가 전혀 고려되지 않았으며 심지어 전체 문장 흐름의 유사도가 전혀 고려되지 않았음을 확인할 수 있다. 이는 현재 제작된 시스템에 있어서 시스템의 근본적인 문제로서 고려되지는 않는다. 이는 앞 단락에서 기술한 바와 같이 인공 신경 정리 증명계의 병렬 모듈화가 가능하기 때문이다. 전역적인 비선형 시스템이 아니기 때문에, 내부 혹은 외부 모듈로 문장 유사도만을 위한 추가 모듈을 장착하는 것이 가능하다.

단어 임베딩의 결과 몇몇 단어들의 벡터 절대값이 비교적 커다란 값을 갖게 되어 단어들의 선택 빈도가 고르지 않았다는 점이 관찰되었다. 직접적인 연결의 경우 특히 예시 4번에서 이 현상을 확인할 수 있다. 즉 "동안," "맡기," "높," 이라는 단어의 빈도수 때문에 두 삼항구조 "검사(바이러스, 높)"과 "감독(동안, 맡기)" 사이에 해석이 어려운 연결이 생성되었다.

예시 1~4에서 모두 이 현상을 확인할 수 있으며, 예시 2,4번의 경우에서 보다 분명히 확인할 수 있다. 전체 출력 결과 중 본 논문에 기재하지 않은 많은 경우에 있어서 역시 주제(현재의 경우 암 및 건강)에서 벗어나는 외부 단어들에 의해 해석이 어려운 연결이 생성되었다. 특히 "있," "없," "되," "지"와 같은 형식어가 매우 빈번히 선택되었다.

본 연구에서는 사용자 서비스를 위한 필터링 모듈 없이 결과를 출력하였다. 사용자 서비스 주제가 결정되지 않은 상황에서 전처리 필터링을 할 경우에는 논리 조합 및 전개의 범위가 좁아지는 현상이 발생할 수 있다. 따라서 인공 신경 정리 증명계의 내부 구조는 그대로 두되 외부 전처리 모듈을 장착해야 하며, 이는 현재 연구의 초점인 인공 신경 정리 증명계의 내부 재구축과 별개로 설계되어야 한다. 따라서 사용자 서비스 상황에서는 주제 바깥 단어가 과도한 빈도수를 갖는 것을 방지하기 위한 전처리 필터링 과정을 인공 신경 정리 증명계의 외부 모듈로서 추가하는 것이 필요하다.

5. 결론

본 연구는 기존의 인공 신경 정리 증명계가 가지고 있던 한계점인 대규모 문장 말뭉치에 적용 불가능하다는 점을 극복하기 위해 기존 인공 신경 정리 증명계 내부의 순환 모듈을 단어 임베딩 모듈로 교체하였다. 또한 인공 신경 정리 증명계를 이용하여 현실 사례에 해당하는 1만 개 이상 문장 내 논리 구조 규격화 및 자동 주석 시스템을 작동시킬 수 있음을 입증하였다. 국가암정보센터의 검증된 문서로부터 추출된 명제들(총 28,844개)을 위키피디아의 문서에서 추출된 명제들(총 7,844개)에 자동 주석을 하는 시스템을 시험하였다. 기존의 순환 모듈을 이용한 인공 신경 정리 증명계를 이용할 경우 553.8일의 학습 시간이 추정됨에 비해 순환 모듈을 단어 임베딩으로 교체해 재구축한 현재 시스템의 경우 전처리에 72.7분 주 계산에 20.4분, 총 93.1분이 소요되었다.

본 연구에서 재구축한 인공 신경 정리 증명계는 입력으로 주어진 문장에 공리 집합 내의 논리 항과 정확히 같은 논리 항이 포함되어 있을 때 공리 집합 내의 논리 항과 입력된 문장을 정확히 연결한다. 정확히 일치하지는 않되 유사하게 일치하는 경우에 대해서도 학습된 결과를 이용하여 입력된 문장과 공리 집합 내 명제를 연결한다. 또한 추론 규칙이 주어진 경우에는 추론 규칙에 대한 논리 전개에 대해서 역시 같은 기능을 수행한다. 또한 출력된 결과를 다시 입력으로 줄 경우 긴 길이의 추론 역시 수행 가능하다. 본 연구는 그 중 특히 실제 문서에서 추출하였기에 정확히 일치하지 않는 논리 항들의 조합에 대해서 성공적으로 인공 신경 증명계가 작동함을 보였고 이를 분석하였다.

그러나 본 연구는 출력 결과의 엄밀한 정확도를 측정할 수 없다는 점을 그 한계로 가진다. 주어진 문장에 대해서 사람이 직접 주석을 달 경우에도 다양한 주석이 가능하다는 점이 자동 주석된 결과의 정확도에 대한 성능 평가 채점 기준을 정의하기 어렵게 한다. 특히 신뢰도 향상 및 심화 정보 추가를 위한 주석이 성공적인지를 평가할 수 있는 명확한 정답 및 오답 기준은 현재까지 정의되어 있지 않다. 이에 대한 가장 효과적인 해결책은 자동 주석된 결과를 전문가 집단에 의해 검증받아 정확도를 측정하는 것이다. 따라서 이를 사용자 서비스를 위한 추후의 과제로 진행하고자 한다.

이에 있어서 과거에 근본적인 문제점으로 제기되었던 인공 신경 정리 증명계의 계산시간 문제에 대한 해결안이 제안된 현재 시점에서 후속 연구로 단어 필터링을 위한 전처리 모듈과 문맥정보 확인을 위한 후처리 모듈을 추가하고자 한다. 현재 시스템은 풍부한 논리 조합을 최대한 고려하되 문장의 문맥 정보를 고려하지 않고 있는데, 인공 신경 정리 증명계의 가장 큰 장점인 모듈화 가능한 비선형 시스템이라는 점을 활용하여 필터링 모듈을 전후로 추가하는 것이 용이할 것이라 예상한다. 또한 신뢰도 향상 및 심화 정보 추가를 위한 주석에 대한 명확한 채점 기준을 마련하는 연구 역시 진행되어야 한다. 전후 필터링 모듈과 채점기준에 대한 연구가 진행된 이후에 주석의 정확도를 엄밀하게 측정하게 되면 현실 사례에서 유용한 사용자 서비스가 가능할 것으로 기대한다.

감사의 글

이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2014R1A2A1A11052310).

참고문헌

[1] H. Robert, et al., "Wikipedia vs peer-reviewed medical literature for information about the 10 most costly medical conditions", J Am Osteopath Assoc, 114.5, pp.368-373, 2014.

[2] W. Richard, et al., "Deductive Question Answering from Multiple Resources", New Directions in Question Answering, pp.253-262, 2004.

[3] 김신웅, et al., "퍼지논리를 기반으로 한 웹 문서의 효율적인 검색 기법", 한국정보과학회 2011 가을 학술발표논문집(B), 제38권, 제2호, pp.287-290, 2011.

[4] T. Rocktäschel and S. Riedel, "Learning Knowledge Base Inference with Neural Theorem Provers", NAACL Workshop on Automated Knowledge Base Construction, 2016.

[5] L. Jin, et al., "Fuzzy keyword search over encrypted data in cloud computing", IEEE INFOCOM, 2010.

[6] K. Janusz and S. Zadrożny "Queries with Fuzzy Linguistic Quantifiers for Data of Variable Quality Using Some Extended OWA Operators", Flexible Query Answering Systems 2015, Springer International Publishing, pp.295-305, 2016.

[7] S. Ilya, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks", Advances in Neural Information Processing Systems, 2014.

[8] K. Cho, et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation", Conference on Empirical Methods in Natural Language Processing, 2014.

[9] T. Mikolov and J. Dean, "Distributed representations of words and phrases and their compositionality", Advances in Neural Information Processing Systems, 2013.

[10] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation", Conference on Empirical Methods in Natural Language Processing, 2014.

[11] B. Peng, et al., "Towards neural network-based reasoning", IEEE Conference on Robotics, Automation and Mechatronics, 2015.

[12] R. Scott and N. Freitas, "Neural programmer-interpreters", International Conference on Learning Representations, 2015.

[13] 함영균, et al., "Linked Data 를 위한 한국어 자연언어처리 플랫폼", 제 24 회 한글 및 한국어 정보처리 학술발표 논문집, pp.16-20, 2012.

[14] K. Diederik and J. Ba, "Adam: A method for stochastic optimization", International Conference on Learning Representations, 2015.

[15] M. Xuezhe and E. Hovy, "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016.

[16] W. Felix, J. Bergmann, and B. Schuller, "Introducing CURRENNT-the Munich open-source CUDA RecurREnt neural network toolkit", Journal of Machine Learning Research 16.3, pp.547-551, 2015.

[17] G. Alex, N. Beringer, and J. Schmidhuber, “Rapid retraining on speech data with lstm recurrent networks”, Technical Report IDSIA-05-05, 2005.